DOCTORADO EN INFORMÁTICA

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

TESIS DOCTORAL

A Big Data and Machine Learning Model to Improve Medical Decision Support in Population Health Management

Presentado por
Fernando Enrique López Martínez

TESIS DOCTORAL

“A Big Data and Machine Learning Model to Improve Medical Decision Support in Population Health Management”

Presentado por

Fernando Enrique López Martínez

Dirigido por

Doctor D. Vicente García Díaz

Doctor D. Edward Rolando Núñez Valdez

Oviedo, 2020

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: **Modelo de Big Data y Machine Learning para mejorar el proceso de toma de decisiones en la administración de la salud de la población.**	Inglés: **A Big Data and Machine learning model to improve medical decision Support in population health management.**

2.- Autor	
Nombre: **Fernando Enrique López Martínez**	DNI/Pasaporte/NIE:
Programa de Doctorado: **Informática**	
Órgano responsable: **Centro Internacional de Postgrado**	

RESUMEN (en español)

Big data y Machine Learning son las piezas de tendencia más importantes para la innovación y el análisis predictivo en la atención médica, liderando la transformación digital del sector de la salud en todo el mundo. Varias organizaciones ya están trabajando en el desarrollo de plataformas inteligentes de análisis de big data basadas en principios de aprendizaje automático e integración de datos. Sin embargo, estas plataformas no presentan un modelo claro para interoperar, asegurar y utilizar análisis predictivos para mejorar la atención al paciente y proporcionar alertas tempranas de enfermedades. Se propone el diseño de un modelo de plataforma de salud que incluya principios de big data para manejar la complejidad de los datos clínicos, y modelos de aprendizaje automático que ayudan a mejorar la gestión de la salud poblacional, la atención basada en valores y desafíos futuros en la atención médica de hoy. Entre los beneficios tenemos, mejores resultados de atención médica, operaciones clínicas, reducción de costos de atención y generación de información médica precisa. Se implementaron tres modelos de aprendizaje automático para predecir la hipertensión y la sepsis neonatal, percibidos como necesarios en las instituciones de atención médica donde se llevaron a cabo. Los modelos desarrollados usan los conjuntos de datos, complejos y estandarizados, integrados en la plataforma para mejorar la efectividad de las intervenciones de salud, mejora del diagnóstico y apoyo a la decisión clínica. Los datos integrados en el modelo de plataforma provienen de registros electrónicos de salud (EHR), Sistemas de información hospitalaria (HIS), Sistemas de información radiológica (RIS), Sistemas de información de laboratorio (LIS), datos de

salud pública, móviles, redes sociales, y portales web clínicos. Esta cantidad de datos se integra utilizando técnicas de big data para almacenamiento, procesamiento y transformación. Esta investigación presenta el diseño del modelo de una plataforma de salud que se implementará en organizaciones de atención médica en Colombia y USA para integrar repositorios de datos operativos, clínicos y comerciales con análisis avanzados para mejorar el proceso de toma de decisiones a través de análisis descriptivos, predictivos y prescriptivos. Y, se puede adaptar fácilmente al modelo de prestación de servicios de salud para la gestión de la salud de la población.

RESUMEN (en inglés)

Big data and machine learning are the foremost trending pieces for innovation and predictive analytics in healthcare, leading the digital healthcare transformation worldwide. Several organizations are already working on developing intelligent big data analytics platforms based on machine learning, and data integration principles. However, these platforms lack of presenting a clear model for the platform to interoperate, secure and utilize predictive analytics to impact patient care and to provide early warnings of disease conditions. This work discusses how a healthcare platform model that includes big data principles can be designed to handle the complexity of healthcare data and machine learning models helping organizations to improve population health management, value-based care, and new upcoming challenges in today's healthcare settings. The benefits of using the proposed healthcare platform model for community, and population health include better healthcare outcomes, improvement of clinical operations, reducing costs of care, and generation of accurate medical information. Three machine learning models for predicting hypertension, and neonatal sepsis perceived as needed in the healthcare institutions where the research was achieved were implemented as part of this work, and they can use the large, complex and standardized data sets integrated in the platform to improve the effectiveness of public health interventions, improving diagnosis, and clinical decision support. The data integrated in the proposed platform model comes from Electronic Health Records (EHR), Hospital Information Systems (HIS), Radiology Information Systems (RIS), Laboratory Information Systems (LIS), data generated by public health platforms, mobile data, social media, and clinical web portals. This massive amount of data is integrated using big data techniques for

storage, retrieval, processing, and transformation. This research presents a conceptual model of a health platform that will be implemented in healthcare organizations in Colombia and the USA. To integrate operational, clinical, and business data repositories with advanced analytics to improve the decision-making process through descriptive, predictive, and prescriptive analytics. And, it can be easily adapted to the healthcare service delivery model for population health management.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN INFORMATICA

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

FORMULARIO RESUMEN DE TESIS POR COMPENDIO

1.- Datos personales solicitante	
Apellidos: **López Martínez**	Nombre: **Fernando Enrique**

Curso de inicio de los estudios de doctorado	**2016 / 2017**

	SI	**NO**
Acompaña acreditación por el Director de la Tesis de la aportación significativa del doctorando	X	
Acompaña memoria que incluye		
Introducción justificativa de la unidad temática y objetivos	X	
Copia completa de los trabajos *	X	
Resultados/discusión y conclusiones	X	
Informe con el factor de impacto de las publicaciones	X	
Se acompaña aceptación de todos y cada uno de los coautores a presentar el trabajo como tesis por compendio	X	
Se acompaña renuncia de todos y cada uno de los coautores a presentar el trabajo como parte de otra tesis de compendio	X	

* Ha de constar el nombre y adscripción del autor y de todos los coautores asi como la referencia completa de la revista o editorial en la que los trabajos hayan sido publicados o aceptados en cuyo caso se aportará justificante de la aceptación por parte de la revista o editorial

FOR-MAT-VOA-033

Artículos, Capítulos, Trabajos

Trabajo, Artículo 1

Titulo (o título abreviado)	**Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors**
Fecha de publicación	**15 de noviembre de 2018**
Fecha de aceptación	**3 de junio de 2018**
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	**SCI**
Factor de impacto	**JCR 4.292 5/275 Q1**

Coautor2	**X** Doctor	□ No doctor .	Indique nombre y apellidos: **Aron Schwarcz.MD**
Coautor3	**X** Doctor	□ No doctor .	Indique nombre y apellidos: **Edward Rolando Núñez Valdez**
Coautor4	**X** Doctor	□ No doctor .	Indique nombre y apellidos: **Vicente García Díaz**
Coautor5	□ Doctor	□ No doctor .	Indique nombre y apellidos
Coautor6	□ Doctor	□ No doctor .	Indique nombre y apellidos
Coautor7	□ Doctor	□ No doctor .	Indique nombre y apellidos

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Trabajo, Artículo 2

Titulo (o título abreviado)	**A neural network approach to predict early neonatal sepsis**
Fecha de publicación	**Junio de 2019**
Fecha de aceptación	**19 de abril de 2019**
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	**SCI**
Factor de impacto	**JCR 2.189 30/206 Q1**

Coautor2 **X** Doctor □ No doctor . Indique nombre y apellidos	**Edward Rolando Núñez Valdez**
Coautor3 □ Doctor **X** No doctor . Indique nombre y apellidos	**Jaime Lorduy Gomez**
Coautor4 **X** Doctor □ No doctor . Indique nombre y apellidos	**Vicente García Díaz**
Coautor5 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor6 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor7 □ Doctor □ No doctor . Indique nombre y apellidos	

Trabajo, Artículo 3

Titulo (o título abreviado)	**An artificial neural network approach for predicting hypertension using NHANES Data**
Fecha de publicación:	**June 2020**
Fecha de aceptación:	**10 June 2020**
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	**SCI**
Factor de impacto	**JCR 4.011 Q1**

Coautor2 **X** Doctor □ No doctor . Indique nombre y apellidos	**Edward Rolando Núñez Valdez**
Coautor3 **X** Doctor □ No doctor . Indique nombre y apellidos	**Rubén Gonzalez Crespo**
Coautor4 **X** Doctor □ No doctor . Indique nombre y apellidos	**Vicente García Díaz**
Coautor5 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor6 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor7 □ Doctor □ No doctor . Indique nombre y apellidos	

Trabajo, Artículo 4

Titulo (o título abreviado)	**A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management**
Fecha de publicación:	**23 de abril de 2020**
Fecha de aceptación:	**21 de abril de 2020**
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	**SCI**
Factor de impacto:	**SJR 2.20 20/51 Q3**

Coautor2 **X** Doctor □ No doctor . Indique nombre y apellidos	**Edward Rolando Núñez Valdez**
Coautor3 **X** Doctor □ No doctor . Indique nombre y apellidos	**Vicente García Díaz**
Coautor4 **X** Doctor □ No doctor . Indique nombre y apellidos	**Zoran Bursac**

Acknowledgements

During the writing of this dissertation I have received a great deal of support and assistance.

First and foremost, I would like to thank God Almighty for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

I would like to thank my parents for their support and kindly ear. You are always there for me. I wish to thank my wife, Ana, who has stood by me through all my efforts, my absences and impatience. She gave me support and help to persevere prevented several wrong turns. She also supported the family during much of my studies. My beloved daughter Maria Fernanda, I would like to express my thanks for being such a good girl always cheering me up.

I would also like to thank my tutors, Dr. Edward Rolando Nuñez Valdez and Dr. Vicente García Díaz, for their valuable guidance throughout my studies. You provided me with the tools that I needed to choose the right direction and successfully complete my dissertation. Thank you for your patient support and for all the opportunities I was given to further my research.

Abstract

Big data and machine learning are the foremost trending pieces for innovation and predictive analytics in healthcare, leading the digital healthcare transformation worldwide. Several organizations are already working on developing intelligent big data analytics platforms based on machine learning, and data integration principles. However, these platforms lack of presenting a clear model for the platform to interoperate, secure and utilize predictive analytics to impact patient care and to provide early warnings of disease conditions. This work discusses how a healthcare platform model that includes big data principles can be designed to handle the complexity of healthcare data and machine learning models helping organizations to improve population health management, value-based care, and new upcoming challenges in today's healthcare settings. The benefits of using the proposed healthcare platform model for community, and population health include better healthcare outcomes, improvement of clinical operations, reducing costs of care, and generation of accurate medical information. Three machine learning models for predicting hypertension, and neonatal sepsis perceived as needed in the healthcare institutions where the research was achieved were implemented as part of this work, and they can use the large, complex and standardized data sets integrated in the platform to improve the effectiveness of public health interventions, improving diagnosis, and clinical decision support. The data integrated in the proposed platform model comes from Electronic Health Records (EHR), Hospital Information Systems (HIS), Radiology Information Systems (RIS), Laboratory Information Systems (LIS), data generated by public health platforms, mobile data, social media, and clinical web portals. This massive amount of data is integrated using big data techniques for storage, retrieval, processing, and transformation. This research presents a conceptual model of a health platform that will be implemented in healthcare organizations in Colombia and the USA. To integrate operational, clinical, and business data repositories with advanced analytics to improve the decision-making process through descriptive, predictive, and prescriptive analytics. And, it can be easily adapted to the healthcare service delivery model for population health management.

Keywords: Decision Support Systems, Population health management, Big Data, Machine Learning, Deep Learning, Personalized Patient Care.

Resumen

Big data y Machine Learning son las piezas de tendencia más importantes para la innovación y el análisis predictivo en la atención médica, liderando la transformación digital del sector de la salud en todo el mundo. Varias organizaciones ya están trabajando en el desarrollo de plataformas inteligentes de análisis de big data basadas en principios de aprendizaje automático e integración de datos. Sin embargo, estas plataformas no presentan un modelo claro para interoperar, asegurar y utilizar análisis predictivos para mejorar la atención al paciente y proporcionar alertas tempranas de enfermedades. Se propone el diseño de un modelo de plataforma de salud que incluya principios de big data para manejar la complejidad de los datos clínicos, y modelos de aprendizaje automático que ayudan a mejorar la gestión de la salud poblacional, la atención basada en valores y desafíos futuros en la atención médica de hoy. Entre los beneficios tenemos, mejores resultados de atención médica, operaciones clínicas, reducción de costos de atención y generación de información médica precisa. Se implementaron tres modelos de aprendizaje automático para predecir la hipertensión y la sepsis neonatal, percibidos como necesarios en las instituciones de atención médica donde se llevaron a cabo. Los modelos desarrollados usan los conjuntos de datos, complejos y estandarizados, integrados en la plataforma para mejorar la efectividad de las intervenciones de salud, mejora del diagnóstico y apoyo a la decisión clínica. Los datos integrados en el modelo de plataforma provienen de registros electrónicos de salud (EHR), Sistemas de información hospitalaria (HIS), Sistemas de información radiológica (RIS), Sistemas de información de laboratorio (LIS), datos de salud pública, móviles, redes sociales, y portales web clínicos. Esta cantidad de datos se integra utilizando técnicas de big data para almacenamiento, procesamiento y transformación. Esta investigación presenta el diseño del modelo de una plataforma de salud que se implementará en organizaciones de atención médica en Colombia y USA para integrar repositorios de datos operativos, clínicos y comerciales con análisis avanzados para mejorar el proceso de toma de decisiones a través de análisis descriptivos, predictivos y prescriptivos. Y, se puede adaptar fácilmente al modelo de prestación de servicios de salud para la gestión de la salud de la población.

Palabras clave: Sistemas de apoyo a la decisión, gestión de la salud de la población, Big Data, aprendizaje automático, aprendizaje profundo, atención personalizada al paciente.

Table of Contents

1. INTRODUCTION

During the last decade, the cost of health care in numerous countries is very high and continues to grow fast. The challenge is to improve the healthcare outcome of the population while optimizing and lowering the costs of financial, clinical, and operational resources. One method is collecting and aggregating the unprecedented amount of clinical and operational data from significant data sources that reside on healthcare and financial systems. It then produces an analysis of the data to generate actionable insights through which care clinicians can get a comprehensive clinical and financial picture of the population's health to improve healthcare outcomes across the continuum of care. This research focuses on two healthcare systems: high quality and affordable healthcare system in Colombia, and the other is a costly healthcare system in the USA offering a wide variety of health care professionals, equipment, staff, and clinicians. Colombia's health system is structured by the public sector and the private sector. The general social security system has two plans, contributory and subsidized. The contributory regime covers salaried workers, pensioners, and independent workers, with the subsidized plan covering anyone who cannot pay. Enrollment coverage increased from 96.6% in 2014 to 97.6% in 2015 [1]. The National Health Authority's primary purpose in Colombia is to improve the quality of health care and strengthening supervision, surveillance, and control of the health system. The 2015 statutory health law No. 1751 places the responsibility for guaranteeing the right to health with the health system and recognizes health as a fundamental social right and makes it the state's responsibility to pursue an approach in health promotion and disease prevention [2]. The health sector in Colombia supports all initiatives for implementing new technologies to prevent chronic diseases, disabilities, and high-cost hospitalization cases [3]. There is a remarkable need to improve the prediction of the risk of conditions for the population through the integration and unification of massive amounts of data and the implementation of effective advance analytic solutions to improve the decision-making process and population health management in Colombia's population [4]. In contrast, the U.S. health system is a combination of public and private, for-profit and nonprofit insurers and health care clinicians. The federal government provides funding for the national Medicare program for adults age 65 and older and some people

with disabilities as well as for various programs for veterans and low-income people, including Medicaid and the Children's Health Insurance Program. Private insurance, the dominant form of coverage, is provided primarily by employers. It is essential to mention that public and private insurers set their own benefits packages and cost-sharing structures within federal and state regulations, allowing healthcare organizations to negotiate rates and services based on healthcare outcomes and operational optimization.

World Health Organization estimates that 40 million deaths occurred due to noncommunicable diseases in the world in 2015, most of those deaths were caused by the cardiovascular disease with 17.7 million [5]. In the United States, cardiovascular disease is the leading cause of death despite the existence of effective and inexpensive treatments [6]. In 2015 the number of deaths in adults due to conditions of heart based on death certificates was 633,164 deaths in adults of 20 years old and over. Hypertension or high blood pressure is one of the most critical risk factors for cardiovascular disease among U.S. adults [7] and, it has significant public and economic implications. The health care expenditure associated with high blood pressure in 2011 in the US were $46 billion, and the projected total cost for 2030 is $274 billion [8]. The prevalence of hypertension in the U.S. adult's population is high and increasing in recent years as can be seen in the National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics. This survey has been the main source of tracking the burden of hypertension in the U.S. population [9]. This increment in hypertensive population probably shows us that the current approaches to predict cardiovascular disease fail to identify people who would benefit from preventive hypertensive treatment. The use of machine learning tools in medicine has proven significant outcomes in clinical decision support and patient care, specifically in hypertension diagnosis [10]. Due to the impact of hypertension in the population's health, we decided to build and train a machine learning model to predict hypertension using NHANES data with the collaboration of the head of the cardiovascular department in a healthcare institution in the USA.

Neonatal sepsis is another leading cause of death, and a severe problem for neonates. It is estimated to affect more than 3 million newborns worldwide every year [11]. World Health Organization estimates that one in ten deaths associated with pregnancy and 5 childbirth is due

to maternal sepsis with over 95% of deaths due to maternal sepsis occurring in low and middle-income countries. Despite the advances in antibiotic therapy and the awareness of risk factors in neonatal intensive care units, neonatal sepsis continues to be a severe complication and cause of severe illness and deaths of hospitalized neonates. The amount of data routinely collected in electronic medical records (EMR) and bed-monitors allow us to obtain quality data to build useful predicting models and tools like smart systems to generate continuous risk-assessments for neonatal sepsis from big clinical data and machine learning to produce an earlier diagnosis and improve sepsis management in neonates. Early diagnosis has been shown to reduce delays in treatment, increase appropriate care and reduce mortality [12].

For this reason, this research in collaboration with clinical experts will present the use of integrated clinical data to create an accessible and interpretable machine learning prediction model to classify hypertensive patients and a non-invasive prediction model that can be used as an inference engine of a smart system to provide decision support for health care clinicians at neonatal intensive care units delivering antibiotic administration when sepsis is detected.

This research also presents a model of a healthcare platform and its architectural design where organizations can obtain valuable insights from large amounts of heterogeneous data generated by different data sources from transactional healthcare systems. Correspondingly, the development of proper advanced data analytics methods such as machine learning and big data analytics to perform meaningful real-time analysis on the data to predict clinical complications before it happens to support the decision-making process, and to handle the complexity of the data-driven problems healthcare organizations are currently facing. Machine learning models were developed in two different institutions, Keralty organization in Colombia and Englewood Health in the USA.

Keralty organization is shaped by a group of insurance and health services companies with a global presence, which together develops an integral health model, whose purpose is to produce health and well-being for people throughout their lives. The organization is committed to keeping its users healthy and autonomous, focusing on prevention, identification, and management of health risks, control, and care of disease and dependency [13]. Keralty organization is a leader in

Colombia by providing integrated health services and is recognized for their human, scientific, technical and ethical approach [14]. On the other hand, Englewood Health is one of New Jersey's leading hospitals and healthcare networks. Composed of Englewood Hospital, the Englewood Health Physician Network, and the Englewood Health Foundation, their health system delivers nationally recognized care in a community setting to residents of northern New Jersey and beyond [15].

The previous mentioned healthcare organizations know they need to integrate, normalize, store and analyze their data. The problem is that they do not know where to start, and what is required to accomplish this. In addition, these organizations have been discussing about implementing machine learning models, and again the same question arises, where to start, how to be implemented, what needs to be done. This research, through their objectives shows clearly the development of machine learning models from feature selection to model interpretation, and a conceptual design of a healthcare platform model and its components that will provide useful insights to healthcare organizations to start this endeavor.

2. BACKGROUND

In this section, we discuss the basics of big data, population health management, and machine learning concepts, especially in the healthcare field, to correctly understand the role of every component in the research.

2.1. Population health management

The simple definition of population health management is the aggregation of patient data across multiple health information technology resources, the analysis of this information, and the actionable events performed by healthcare clinicians and healthcare facilities to improve clinical and financial outcomes. Healthcare institutions handle big and complex amounts of data generated from electronic health records, laboratory information systems, radiology information systems and financial billing systems. The integration of all these diverse data sources can provide useful information for modeling patient population health outcomes. Many activities and measurable results can be achieved by healthcare facilities and provider's networks to improve these outcomes, such activities can be quality reporting, care coordination, case management, chronic disease management, medication management, and health and financial risk status.

The expected outcomes from the implementation of the proposed health data platform model to support facilities transitioning to outcome-based reimbursement models for population health management are:

- Obtain a 360-degree view of the patient, to provide the clinicians with a specific and granular view of the health and financial cost of the patients.
- Care clinicians can easily describe the interactions the provider had with the patient over time and identify gaps in care occurring in the transition of care process.
- Healthcare clinicians have the possibility of identifying patient prescription trends and consumption based on pharmacy records.

- Healthcare clinicians can compare themselves to other providers in the network or program to receive performance bonuses. Reports can visibly show clinicians quality and performance metrics to allow the comparison.
- A measurable impact on preventive care with results in early healthcare conditions detection.

Another important piece for the research is the type and nature of the data we collect, transform and prepare for population health management. To have a good understanding of the complexity and importance of the data will allows us to implement better pipelines for data integration and to prepare the data used to train the machine learning models. The next section will comment on the data used for analysis in population health management.

2.1.1. The data in population health management

In this research, population health management is the beneficiary of implementing machine learning models, and data is the enabler. In medical domains, data come in many forms, but machine learning models rely on well-structured and normalized data sets. The data used in population health management by nature is diverse, big and complex. And, usually correspond to clinical, and financial business processes generated from transactional systems utilized to manage health and financial outcomes of the population. The data integration from this transactional sources should be at patient-level detail, and it should include, patient demographics, visits or encounters level information, telephone and web encounters in different settings, progress notes, prescription records, problems, vital signs, immunizations, laboratory data and past medical history. Correspondingly, financial data, commonly named claim data, and it typically contains, type of visit, admission date, discharge date, diagnosis related group (DRG Codes), visit charges, diagnosis codes, procedure codes, payer and reimbursement data. The goal of integrating this data is to aggregate and analyze data at patient level thus the resultant attributes can successfully describe the interactions that healthcare clinicians have with patients over a period. Another important feature of data in population health management is the data store and the volume. Storing and assembling clinical, administrative, programs and claims data requires substantial resources, including time and infrastructure. In average, 10 million of claims

are submitted each month at Keralty. This consumes nearly 2 terabytes of storage, and the data must be stored for over 7 years for auditing purposes. The volume of data shows that Important resources are needed to manage and maintain the infrastructure, security, data acquisition, aggregation, linkage, cleaning and analysis. As a result, the correct management and processing of the data for population health management produces better care management, corporate reporting and continuous quality improvement initiatives.

In the next section, we attempt to provide details on the impact of big data in the transformation of the healthcare sector, population health management and its impact in our daily lives. It is important for this research to review the big data concept because the implementation of machine learning algorithms would be necessary to generate useful insights from the large amount of data present in the healthcare field.

2.2. Big data

Healthcare is a complex scheme established with the only purpose for the prevention, diagnosis, and treatment of health-related matters or impairments in human beings. A healthcare system is formed by key components such as health professionals, health facilities, and a financing institution to manage funds and claims for medical resources. A big challenge in healthcare is to aggregate and normalize the enormous amount of data across different health records platforms, laboratory information systems, radiology information systems, and external sources like wearables and sensors. The data in healthcare is typically so large and complex that it is difficult to manage with traditional data management methods [16], and when the data exceeds the traditional capacity of conventional database systems, and it is generated and collected too fast, the concept of big data appears in the scene. Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information [17]. Big data in Healthcare comprises massive amounts of generated genomic data, clinical data, behavior data, business, financial data, and data generated from sensors and wearables devices. Population health management usually requires the collection of heterogeneous data from multiple sources and the application of advanced analytics models to improve clinical operations

and public health. These advanced analytics models denote predictive modeling and the transformation of this data information into actionable information.

2.2.1. Applications of big data in population health

Population health management is the aggregation of healthcare data across multiple health information technology systems and the analysis of this data to improve healthcare outcomes. Here, we list some of the numerous benefits big data can add to population health management.

- **Earlier disease detection**: Nowadays, a lot of sensors are being used to monitor patient health; these sensors produce a vast amount of information that can be aggregated and analyzed in real-time. This analysis can predict disease outbreaks and detect the early development of infections.
- **Fraud detection and prevention**: Big data predictive analytics can be used by healthcare payers to check the accuracy of claims and payments and verify compliance with medical policies [18].
- **Personalized patient care**: Healthcare is moving from a disease-centered model to a patient-centered model [19]. Data from the medical evidence and electronic health records are needed, and big data will facilitate the collection of personalized patient care data to improve clinical outcomes.
- **Genomics analytics**: A considerable number of diseases are genetic; genomic data can better understand these diseases, and conventional data-driven technologies can't treat this complex type of data.
- **Clinical outcomes analytics**: The data collected from clinical systems, clinical operations, and financial systems through big data provide improvements in clinical outcomes [20].
- **Smart health and wellbeing**: With the current sensor technologies, the adoption of wearable devices, and smart home environments, the amount of data collected from patients and their habitats require big data technologies to turn this massive amount of data into useful information that profoundly impacts population health.

As presented before, the use of big data generated in healthcare shows potential for improving healthcare outcomes and controlling operational and financial costs.

2.3. Machine learning

While massive data is becoming very familiar in healthcare, data mining and machine learning techniques can be implemented to generate knowledge from this data. In addition, machine learning has a significant opportunity to be applied in population health management. However, before investing in machine learning technologies, healthcare organizations should have a clear and concrete idea of how they will use it to improve health outcomes and better population health management. Machine learning algorithms were designed and used to analyze medical data sets [21]. Furthermore, still today, machine learning provides several tools for intelligent data analysis. Through the years, advanced machine learning methods has been used to analyze clinical data sets, and clinical prediction models are one of the essential branches of healthcare data analytics. The following is a list of the main learning methods that have been used successfully for clinical prediction tasks:

1. **Linear regression**: In linear regression, the dependent variable or outcome is assumed to be a linear combination of the attributes with corresponding estimated regression parameters as explicated in the elements of statistical learning by professors of statistics at Stanford University [22].
2. **Logistic regression**: A binary or multiclass classification method that assumes there is a linear relationship between the features and the log-odds of the probabilities as explained in a methodology review performed by a professor of the department of software engineering for medicine at University of Applied Sciences in Hagenberg, Austria [23].
3. **Bayesian models**: One of the essential principles in probability and mathematical statistics is the link between the posterior probability and the prior probability. It is possible to see the probability changes before and after accounting for a particular random event. This theorem is explained in detail in a very interesting article in the Journal of Applied Artificial Intelligence where the author explains inductive learning in medical diagnosis [24].

4. **Decision trees**: One of the most widely used clinical prediction models [25]. The predictions are made by asking a series of questions about a test record and based on the answers. The test record falls into a smaller subgroup where the individuals are similar to each other with respect to the predicted outcome as explained in a comparison of logistic regression to decision trees in medical domain initiated in the National Center for Biotechnology Information [26].
5. **Artificial Neural Networks**: Biological neural systems inspire this method. Simple artificial nodes called “neurons” are combined via a weighted connection to form a network that simulates a biological neural network as explained in early research at Cornell Aeronautical Laboratory in the fifties and redefined recently in many medical data classification models [24][27].

There are many other methods and future trends where machine learning in medical diagnosis could take place. The critical element here is that with all the advances in computer technology and devices to collect data, researchers have many opportunities to improve health outcomes and medical diagnoses that will positively affect population health management.

3. RELATED WORK

Big data and machine learning have become topics of special interest for the past two decades because of the great potential they have to improve healthcare outcomes. These elements require proper management and analysis in order to produce meaningful information for population health management. Otherwise, looking for a solution by just analyzing big data without proper preparation and structure becomes a problem. There are various challenges associated with the handling of big data and the utilization of machine learning that can only be exceeded by using high-end computing frameworks for big data analysis. For this reason, to provide relevant results for improving population health management, healthcare clinicians required the appropriate infrastructure, framework and machine learning models to systematically generate and analyze big data.

3.1. Analytics platforms

Remarkably, in recent years, several companies and start-ups have appeared to provide healthcare analytics platforms and solutions. Some of the vendors and platforms in the healthcare sector are shown in Table 1.

Table 1. Big data analytics platforms in healthcare sector

Platform	Description	Website
IBM Watson Health	Provides services on sharing clinical and health related data among hospital, researchers, and provider for advance researches.	https://www.ibm.com/watson
Ayasdi	Provides AI housed platform for clinical variations, population health, risk management and other healthcare analytics.	https://www.ayasdi.com/

Apixio	Provides cognitive computing platform for analyzing clinical data and digital health records to generate information.	https://www.apixio.com/
OptumHealth	Provides healthcare analytics, improve modern health system's infrastructure and comprehensive and innovative solutions for the healthcare Industry.	https://www.optum.com/
Digital Reasoning Systems	Provides cognitive computing services and data analytic solutions for processing and organizing unstructured data into meaningful data.	https://digitalreasoning.com/
Innovaccer	Intuitive healthcare analytics offering for population management health strategies in the industry.	https://innovaccer.com/
Health Catalyst DOS	Provider of data and analytics technology and services to healthcare organizations, committed to be the catalyst for massive, measurable, data-informed healthcare improvement.	https://www.healthcatalyst.com/
Google Health	Encourages collaboration amongst clinicians and supports them to visualize patient health trends, curate patient lists, spot signs of deterioration earlier, and receive notifications about preventable conditions.	https://health.google/

IBM Watson Health is one of the big data analytics platforms that utilizes machine learning and artificial intelligence-based algorithms to provide predictive insights from enormous quantities of structured and unstructured data. The architecture model of this platform includes a data collection layer for data consolidation and data integration, a data organization layer for master data management, and a governed data lake, an analyzing layer for business analytics and

machine learning deployment and a top layer called infuse layer where the platform orchestrates the cognitive services and business processes [28]. This platform shows clearly a way for extracting evidence and insights that can help inform clinical, operational and business decisions. However, at the time of this study, the platform lacks a robust population health insights layer for population health management, focusing on AI services for the understanding of human disease enabling approaches for diagnosis and treatment.

Google Health offers a healthcare analytics platform that helps process clinical and operational healthcare data to researchers, data scientists, IT teams, and business analysts. This platform harmonize data, monitor data pipelines, run analytics, and create visualizations for provider insights and data-driven decision-making [29]. However, this platform fails in satisfying essential requirements to privacy and a well define layer of health program management for population health management focusing on the computer vision diagnosis field and predictive analytics for cancer, eye disease and acute kidney injuries.

Health Catalyst data operative system (DOS) is a healthcare solution approach that combines data warehousing, clinical data repositories, and health information exchanges in a single platform allowing analysis of clinical, financial and patient data [30]. The platform architecture includes a data integration layer, a microservices architecture layer, a rapid response analytics layer that runs machine learning models and enables rapid development and utilization of these models, and an agnostic data lake for data integration. This is the strongest healthcare analytics platform in the market from the model and architecture perspective to improve population health. However, still this model lacks well define layers for data orchestration, business logic and care management.

Ayasdi is a healthcare platform that uses machine learning to capture revenue, minimize risks, and optimize operational efficiencies in healthcare [31]. It offers a framework that permits to rapidly build intelligent, automated applications for any size problem in theory. However, the Ayasdi AI layer absences of a clear machine learning approach and poor documentation on the tools and methods used to analyze the data and discover relationships within the variables.

Innovaccer is a population health management platform that offers seamless integration, unified patient records, interoperability and advanced analytics to identify at-risk patients, examine high-utilization measures and cost-drivers, track underlying utilization patterns and monitor adherence to medication [32]. This is a well-positioned platform for healthcare analytics, however, doesn't offer machine learning capabilities neither a clear business logic layer in the platform architecture for care program management.

In this research, we proposed a health platform model with well-defined layers for data integration and interoperability. Business logic and business integration for business processes and business operations. Foundational architectural layers such as security, governance, compliance and management, apps and services, a layer for business analytics and machine learning development that covers all the needs for an effective population health management analytics platform.

3.2. Machine learning models

Another important component of this research is the selection of hypertension and sepsis as the conditions to build the machine learning models. We review the state of art of several prediction models for hypertension and neonatal sepsis.

We reviewed previous studies of risk models to predict hypertension and other literature reviews. The number of people in these studies ranged from 637 to 11,407, with the age of participants ranging from less than 25 to 69 years or more and several of them only including a single gender. Age, sex, and smoking are present in almost all of them. For our model, gender was not statistically significant. However, we included it based on its clinical importance for our study.

Ahmad [33] reported strong correlation between age and blood pressure and the prevalence of hypertension increased with body mass index (BMI). It also shows that systolic response was a weaker predictor of hypertension and the identified risk factors were not the same in the studies conducted in different locations for this study.

Manandhar [34] showed that the hypertension difference between male and female was not statistically significant, and age, smoking, obesity, alcohol and family history of hypertension was statistically significant. The study includes features like occupation and religion that are not considered in other studies.

Zheng [35] includes age, fasting plasma glucose (FPG), alcohol and smoking as risks factors of hypertension. It also includes 102 more variables after feature selection, increasing the complexity and possibility of unintelligible results. For numeric variables the study cannot calculate their odds ratios. And, it includes variables that are not relevant for our study like the number of hours of sleep, the number of people for dinner, the times of eating pickles, illiterate or groundwater consumption.

Ramezankhani [36] includes low estimated glomerular filtration rate (eGFR), high fasting plasma glucose (FPG), body mass index (BMI), age and smoking status. The results of this study were useful for targeting efforts to promote strategies to reduce the risk of cardiovascular disease. However, the study excludes population between 20 and 30 years old.

For this research, in our machine learning model for predicting hypertension we used systolic blood pressure to calculate the class because of the strong significance with hypertension and the indication of the subject matter expert. Age, gender, ethnicity, BMI, smoking status, diabetes and kidney disease were included as well. We used estimated glomerular filtration rate and high fasting plasma glucose to calculate kidney disease presence as indicated in other studies.

Several papers were also reviewed that used artificial neural networks to predict sepsis neonatal. In every article, we analyzed the model building process, variable selection, ground truth, training and test datasets, overfitting avoidance, error estimate, and area under the curve information.

Subramani [37] presents several machine learning models including support vector machines (SVM), naive Bayes classifier (NB), tree augmented naive Bayes (TAN), averaged one-dependence estimators (AODE), K nearest neighbor (KNN), decision tree classifier and regression trees (CART), random forest (RF), logistic regression (LR) and Lazy Bayesian Rules (LBR). The dataset used consisted of 299 infants evaluated for late-onset sepsis. Several feature selection algorithms were used to select highly predictive features including SVM - (forward, backward, forward-

backward and recursive), HITON Markov blanket and HITON - parents and children algorithms. This study reported an area under the curve of 0.78 for naive Bayes.

Grifin [38] shows that the clinical diagnosis of neonatal sepsis is preceded by abnormal heart rate characteristics (HRC). This study reported an area under the curve of 0.82 for sepsis prediction using multivariable logistic regression. The dataset used consisted of 678 infants.

Honoré [39] presented a shallow feed-forward Neural Network model with 30 hidden nodes that used an imbalanced dataset composed by heart frequency signals and SpO2 signals after applying frame normalization and removed the mean of every signal. This study reported an area under the curve of 0.85 for sepsis prediction, but the author expressed that this model is based on the inaccurate modeling of a deficient number of training examples.

Calvert [40] developed a high-performance early sepsis prediction technology for the general patient population. This model reported an average area under the curve of 0.92. This model used nine vital sign variables, systolic blood pressure, pulse pressure, heart rate, temperature, respiration rate, white blood cell count, pH, blood oxygen saturation, and age.

Desautels [41] applied the InSight, machine learning classification model developed by Calvert, and used combinations of patient data such as vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age. This model reported an average area under the curve of 0.88.

Horng [42] presents a model that includes free text, vital signs, and 70 demographic data to identify patients with sepsis at the emergency department. This model reported an average area under the curve of 0.85.

The machine learning model presented in this research shows several improvements regarding the studies presented above. The type of risk variables included in our study such as sociodemographic, obstetric, neonatal and maternal infectious related variables were not used at the same time in the variable selection criteria in the other studies. Our classification results suggested that the combinatory use of these variables and the proposed an artificial neural network (ANN) is a potentially useful clinical model to classify the neonatal sepsis. Second, our

study shows a better calculated area under the curve than most of the work described before due to the performance of the ANN architecture even with a highly unbalanced dataset.

4. OBJECTIVES

The main objective of this research is to develop a conceptual model that clearly presents the components needed for an effective platform to provide integration, interoperability and management of big data in healthcare, and the implementation of machine learning models to improve population health management to support the decision-making process in healthcare facilities.

The main objective is divided into the following specific objectives:

1. **Develop a machine learning model to improve the decision-making process in population health management for hypertensive patients**. This research in collaboration with clinical experts presents the use of several non-invasive factors to create an accessible and highly interpretable logistic regression prediction model to classify hypertensive patients and study the relevance of each variable in the presence of the others using national health data from the National Health and Nutrition Examination Survey (NHANES). Risk variables were selected after performing a compressive review of studies describing equations to predict hypertension and the clinical suggestions of the subject matter expert that worked with us on the development of the machine learning model.
2. **Develop a machine learning model to provide decision support for health care clinicians at neonatal intensive care units**. This research presents a non-invasive prediction model that can be used as an inference engine of smart systems to provide decision support for health care clinicians at neonatal intensive care units to provide antibiotic administration when sepsis is detected. The necessity of this model was accessed with the clinical team that participates in our research and the importance of including this machine learning model as part of the models that will be integrated on the healthcare platform.
3. **A conceptual design of a healthcare platform model to integrate and normalize the data from the organizations and allow the machine learning models to take part in the advanced analytics modules for population health management**. This research presents

the conceptual design of a digital health platform model for a healthcare organization in the USA and Colombia to integrate operational, clinical, and business data repositories with advanced analytics to improve the decision-making process for population health management. The model designed in this research will be the input for the implementation of the healthcare platform as part of the healthcare strategy for Keralty organization toward achieving improvement in the health of the population.

5. PROPOSAL

In this research, the motivation is the design of a healthcare platform model and several machine learning models focused on specific healthcare problems evaluated with different clinical teams at Englewood Health and Keralty. This research proposes three machine learning models, two for predicting hypertensive patients with the development of a logistic regression and an artificial neural networks algorithm, and one for early neonatal sepsis detection with the development of an artificial neural network algorithm. Correspondingly, another goal of this research is that the platform model contains all the components needed for the healthcare platform to be aligned with the triple aim framework developed by the Institute for Healthcare Improvement that describes an approach to optimizing healthcare system performance, improving the patient experience of care, improving the health of populations and reducing the per capita cost of health care. Also, the organizations where this research was developed share their vision that everyone should have the best care and health possible, and they support their mission of improving health and health care worldwide.

The designing of this healthcare platform model intends to resolve several problems in healthcare services to assist patients and their families in managing their health by providing better access to healthcare services and clinicians to manage these healthcare services. The data process methodology used in this research is shown in Figure 1 and it is based on the Cross-Industry Process for Data Mining (CRISP-DM) methodology [43] and the Data Mining Methodology Extension for Medical Domain (CRISP-MED-DM) [44]. We decided to use this structured approach to plan the development and consumption of the machine learning models over other methodologies such as Sample, Explore, Modify, Model and Assess (SEMMA) [45] and(Process model, Predictive Model Markup Language (PMML) [46] because it is a robust and well-proven methodology and its easy adaption to medical domain.

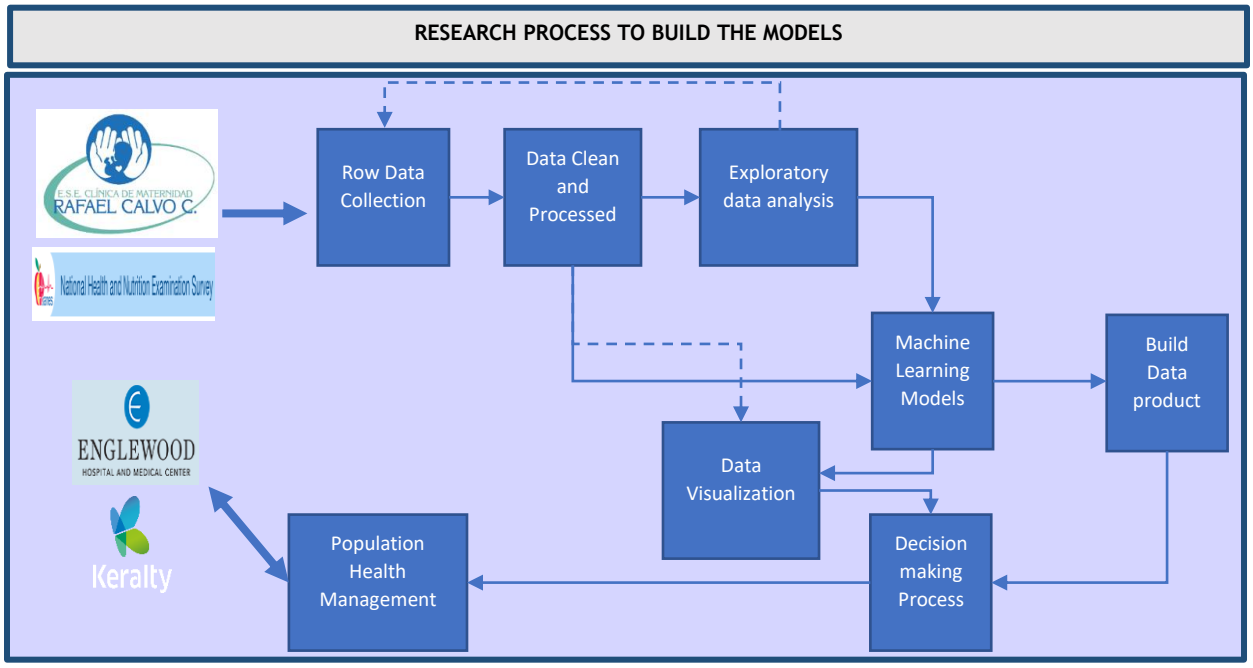

Figure 1 Data Methodology to build the models

- **Problem definition**: One of the most critical things to do, is to define what are the inputs and the expected outputs we will need to resolve the problem. After meeting with the clinical institutions, the following elements were defined: main objective, what are we trying to predict, the target features, the input data, kind of problem facing, expected improvement, current status of the target feature and how is going to be measured the target feature. This helped on defining the scope and objectives of this research.
- **Raw data collection**: Data collection is the process of gathering and measuring information from countless different sources. For the logistic regression classification model and the neural network model, The National Health and Nutrition Examination Survey (NHANES) datasets from NHANES 2007-2008 to NHANES 2015-2016 were used, and for the early neonatal sepsis detection, anonymous dataset from Rafael Calvo's clinic in the city of Cartagena, Colombia from 2016 to 2017.
- **Exploratory data analysis**: Understanding the data was very important for this research. In this work, summarizing, plotting and reviewing the data allow us to perform a data sanity check to identify the actions needed for the data cleansing process and how to handle the relationship between the features.

- **Data cleansing and feature selection**: Data cleaning is a critically important step in any machine learning project. In this project, activities were performed to detect and fix errors in the data. Some of the data cleaning activities were removal of undesirable observations, fixing of structural errors in clinical measure variables, removal of unwanted outliers and handling of missing data.
- **Machine learning model selection**: Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.
- **Data visualization**: Visualization is an important component of any analytics project and data is more than just a set of descriptive numbers, insights taken from data should be transformed into actionable understanding. It is important to shape the data in a more defined story, giving a better context for the information it is been served.
- **Build data product**: Managing a final machine learning model to build a product is more difficult than managing normal software products because it involves more uncertainties and requires not only technical but also cultural and organizational changes in the organization. This phase needs the participation of an interdisciplinary team and a well-defined strategy. The building of a machine learning product is a highly iterative process, so models need to be trained, tested, and tuned.
- **Decision making process**: Many machine learning models act like black boxes that takes an input and generates predictions. In a healthcare domain, the decision-making process is based on the explainability and interpretability of the model. Accountability and transparency are extremely important for our models designed for making decisions. To provide transparency to users, the code and hyperparameters are visible to who need to understand how the model is processing the data and explaining specifically which variables are driving the output of the model. Once the developed models have reached a confident threshold of correct results, it can accelerate the decision-making to adopt processes and workflows across the organization. Having the right information at the right time can make all the difference in the healthcare organizations, and It can literally save lives.

- **Population health management**: The goal of population health management (PHM) is to improve the health status of a group of patients, and by extension, the health outcomes for individual patients. Data aggregation and advance analytics, including machine learning models can augment the availability and usefulness of longitudinal patient risk predictions delivering better individual and population-level outcomes. Englewood Health and Keralty organizations decided to start successful population health management programs based on big data analytics to provide real-time insights to clinicians and administrators and to allow them to find and to manage gaps in care within their patient population. The organizations were in the middle of hiring vendors for implementing the platform without a clear path and how to ensure that the new platform accomplished the expected outputs. The organizations decided to create a team led by the researcher to propose the platform's model and develop a proof of concept about the development of machine learning models, the training, evaluation, and results in the interpretation process.

5.1. Machine Learning Models

We briefly present the predictive models implemented in this research:

- **Machine learning classification for a hypertensive population**: Logistic regression classification to evaluate the association between gender, race, BMI (Body Mass Index), age, smoking, kidney disease and diabetes using logistic regression. Data were collected from NHANES datasets from 2007 to 2016 to train and test the model. A sampling dataset of 19.709 with 83% non-hypertensive individuals and 17% hypertensive individuals. The results show a sensitivity of 77%, a specificity of 68%, precision on the positive predicted value of 32% in the test sample and a calculated area under the curve of 0.73 (95% Confidence Interval [0.70 - 0.76]).
- **Neural network approach to predict early neonatal sepsis:** A non-invasive neural network classification model for early neonatal sepsis detection. The data used in this study is from Crecer's Hospital center in Cartagena-Colombia. A dataset of 555 neonates with 66% of negative cases and 34% of positive cases was used to train and test the model.

The study results show a sensitivity of 80.32%, a specificity of 90.4%, precision on the positive predicted value of 83.1% in the test sample and a calculated area under the curve of 0.925 (95% Confidence Interval [91.4 - 93.06]).

- **An artificial neural network approach for predicting hypertension**: A neural network classification model to estimate the association among gender, race, BMI, age, smoking, kidney disease and diabetes in hypertensive patients. Data was obtained from the National Health and Nutrition Examination Survey (NHANES) from 2007 to 2016. This research utilized an imbalanced data set of 24,434 with 69.71% non-hypertensive patients, and 30.29% hypertensive patients. The results indicate a sensitivity of 40%, a specificity of 87%, precision of 57.8% and a measured area under the curve of 0.77 (95% Confidence Interval [75.01 - 79.01]). This research showed results that are to some degree more effectively than a previous study performed by the authors using a statistical model with similar input features that presents a calculated area under the curve of 0.73.

6. A BIG DATA AND MACHINE LEARNING MODEL TO IMPROVE MEDICAL DECISION SUPPORT IN POPULATION HEALTH MANAGEMENT

Big data and machine learning are redefining healthcare goals for the future. Healthcare data is impacting the way disease research is performed, and the level of complexity in population health management is increasing as the traditional fee for service approach is transformed into the value-based care model [47,48]. Population health management is basically the aggregation of patient health data from multiple data sources, and the analysis and transformation of this data into actionable insights to generate informed decisions to improve clinical and financial outcomes [49]. Big data technologies will allow to bring large amounts of structured and unstructured data from disparate data sources into data repositories to be examined and analyzed. Machine learning models assist in discovering insights from complex data sets with capabilities such as finding unseen patterns, making new predictions, and analyzing trends on health data. Machine learning is being used in a variety of clinical domains with the analysis of hundreds of clinical parameters resulting in effective and efficient models to improve the outcomes and quality of medical care models [50]. The design of this platform model for future implementation shows the enormous potential in using big data to individualize medical treatment, the opportunity for improving the lives of the patients, delivering better medical care, and reduced waste at an operational level [51]. We list some of the improvements for the healthcare organizations when implementing the platform model for population health management:

- A clinician would know before prescribing whether the patient is at high-risk to become dependent of this medication and different treatment plans can be selected based on this information.
- Psychosocial and clinical data could inform about the development of a chronic illness that can be properly diagnosed.

- The organization can use big data to understand how they are performing, what are the opportunities to improve clinical care and the capacity to redesign care delivery to their patients.
- Using the platform's analytics component can improve the quality of care and patient experience at the lowest possible cost to the organization.
- Capturing streaming data and wearable data can provide to health care clinicians real-time insights about a patient's health that will allow them to improve their decision-making process for treatment and medication.
- Big data analysis can help the organization to deliver information that is evidence-based and can improve the efficiency, understanding, and implementation of the best practices associated with any disease.

In addition to the big data technologies, another essential component is the advanced analytic module of the platform. This module contains several machine learning algorithms to support clinical diagnosis. However, the organization should feel confident in these models and how they can be applied to specific use cases. These first models will alert clinicians to changes in high-risk conditions such as sepsis and hypertensive patients.

One of the specific objectives of this research is to present the design of a platform model and its components to allow the organizations to drive better and more actionable insights from their data. To derive meaningful information from all this data in a way that allows them to improve care and lower costs needed for value-based reimbursement and business objectives while providing the highest quality care for population health management [52]. The future implementation of this platform intends to resolve several problems in healthcare services to assist patients and their families in managing their health by providing better access to these services [53].

6.1. Proposed platform model architecture

This research proposed a model with several layers to implement a healthcare platform for big data analytics and population health management. A corporate data repository to store massive

amounts of data generated from several sources, integration and interoperability capabilities, and analytic pipelines to gain smarter healthcare options for population healthcare management. Figure 2 shows the basic architectural layers.

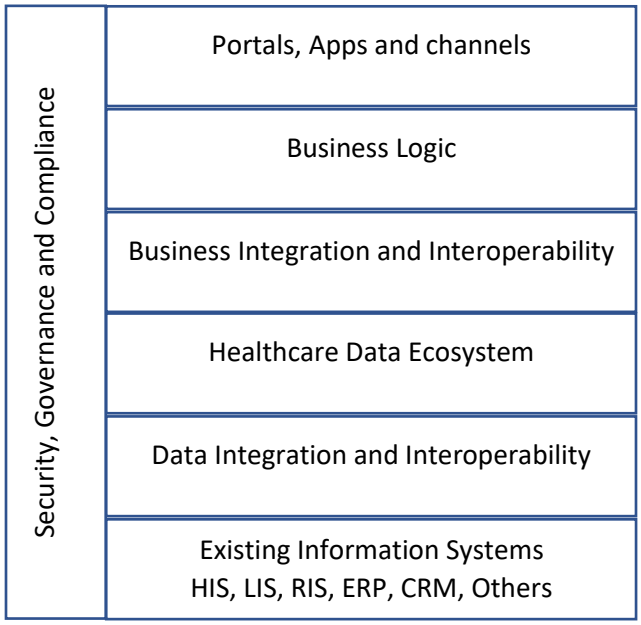

Figure 2. Healthcare Platform Architectural Layers

Then, we expanded each layer with the elements needed for an effective management of the population health and the data and integration ecosystems as shown in Figure 3.

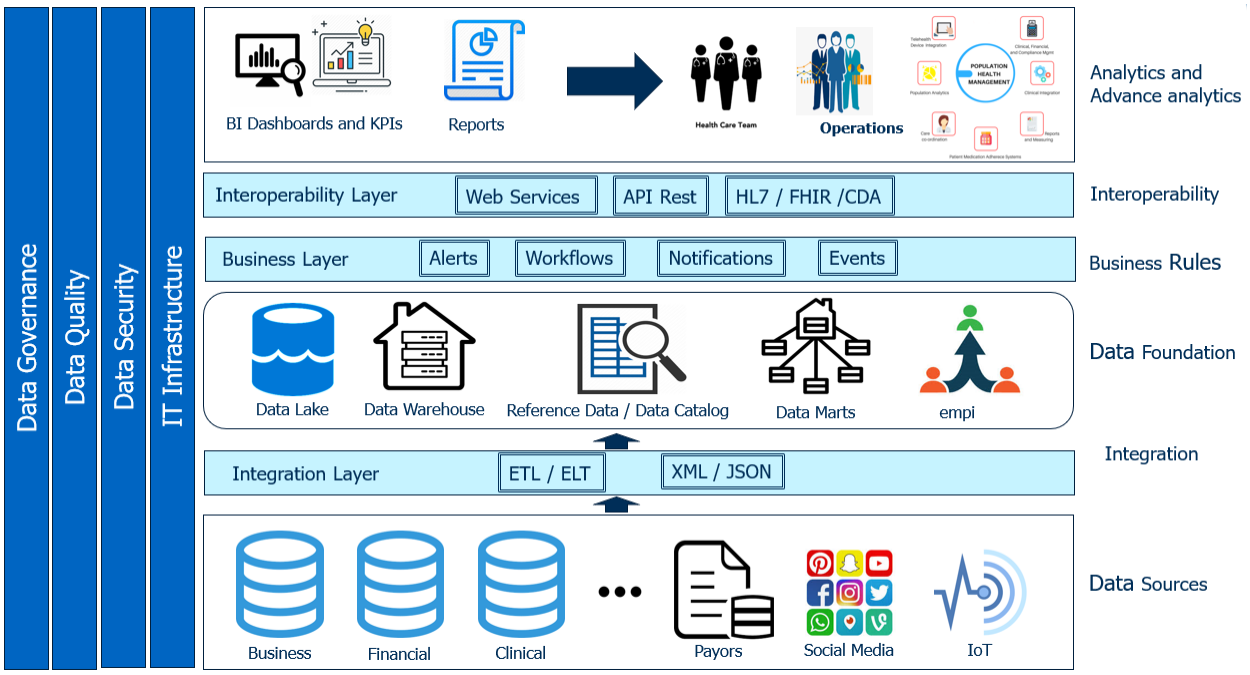

Figure 3. Expanded Architectural layers

Based on the expanded model, the initial approach was to build a proof of concept with a big data processing pipeline with lambda architecture to support real-time and batch analytics. The proof of concept architecture of the platform model is shown in Figure 4. This architecture's model has different mechanisms to consume data depending on the source and timing needed to generate insights. Also, with this approach, we can have professionals with different skills working in parallel to build the platform. The architecture contains a batch layer, a real-time layer, and a serving layer. The batch layer oversees persistent storage and can scale horizontally. The real-time layer process streaming data and performs dynamic computation. The serving layer query data on the repositories and consume the prediction models. From the infrastructure point of view, the platform offers the flexibility of being implemented in a hybrid environment, the cloud, and in the local data processing center using virtualization techniques, containers, and load balancing systems. The design of the infrastructure was prepared to provide a flexible set of resources that can be used on-demand and based on the specific workload requirements. The infrastructure deployment relied heavily on automation to provide fluid operations.

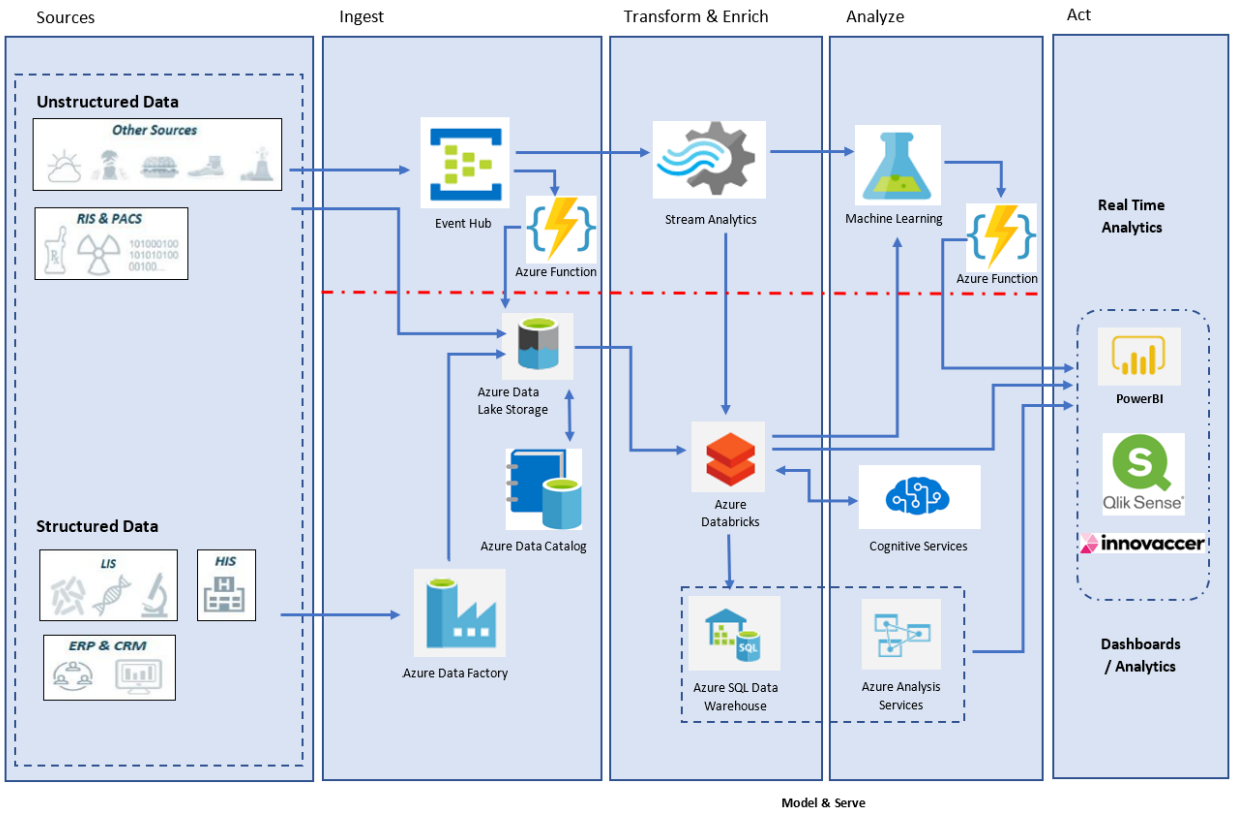

Figure 4 Azure Big Data and Machine Learning Lambda Architecture

6.2. Data Repository

An enterprise-wide staging repository for the big data analytics platform was considered. The data lake allows capturing data of any volume, type, and ingestion speed in one single place for storing heterogeneous data. This staging area included capabilities such as security, scalability, reliability, and availability. The data can be passed processed directly from the staging area or can be ingested to an enterprise data warehouse for historical load, preparation, and serve for business intelligence and machine learning needs. This data warehouse repository has a scale-out architecture and massively parallel processing (MPP) engine. Data models were developed to cover clinical, social, and health care program domains. Each model performs validations and processing on the data received, decoupling the processing and administration of the data from the source. These data models can also be extended to store additional attributes specific to the implementation, allowing these models to subscribe to certain types of messages, using the mapping and filtering options provided by the data processing pipelines. Once these

subscriptions are created, the model will be loaded with all relevant messages to those who are subscribed and stored in the data lake (repository of data stored in its natural format). For data storage, the data is loaded into a data warehouse with a daily refresh. This healthcare data repository contains a highly normalized data model for fast and efficient querying and analysis. This repository is read-only.

6.3. Integration and Interoperability

The platform provides a mechanism to integrate data from heterogeneous sources, define workflows to ingest data from different data stores, and transform and process data to data stores to be consumed by BI applications. A cloud-based data integration service is used to create these data-driven workflows and orchestrate all automation, transformation, and data movement in the platform. The main tasks this integration service should perform are creation and scheduling of data pipelines to ingest data from different data sources, processing and transformation of the data, and store data in data stores such data lakes or data warehouses. Azure Data Factory automates and orchestrates the entire data integration process from end to end in the platform. We built the extract, transform, load (ETL) pipelines with this Azure component. The data is extracted from the source locations, data is transformed from its source format to the target Azure data lake's schema, and the data is loaded into Azure data lake and the data warehouse, where it can be used for analytics and reporting. Azure Data Factory defines control flows that execute various tasks in the transform and load process. We used the mechanism called mapping data flows, combining control flows and data flows to build the data transformations with an easy-to-use visual user interface. These data flows are then executed as activities within Azure Data Factory pipelines. Data Factory is certified by HIPAA (Health Insurance Portability and Accountability Act), which protects the data while it is in use with Azure. In the data flow we created transformation streams where we define the source data and create the graph with the transformations, schema operations such as derived column, aggregate, surrogate keys and selects, and the output settings.

6.4. Data Security and Privacy Model

In terms of security, the platform guarantee authentication, access control, and encryption capabilities. The security mechanisms of the platform can provide protection, alerting monitoring, and support the OAuth 2.0 protocol for authentication with REST interfaces. Access control lists (ACLs) are enabled on folders, subfolders, and files. The platform also provides encryption mechanisms to protect the data. All these capabilities are accompanied by the implementation of enterprise security policies and regulatory compliance requirements.

6.5. Stream Analytics

The platform can handle mission-critical real-time data and to offer end to end streaming pipelines with continuous integration and continuous delivery (CI-CD) services. Other capabilities such as in-memory processing, data encryption, and support of international security standards like HIPAA (Health Insurance Portability and Accountability Act), HITRUST (Health Information Trust Alliance), and GDPR (General Data Protection Regulation).

6.6. Advanced Analytics

The analytic data component consists of two areas; one area is the BI models we develop for tactical, operational, and strategic decisions. The second area of analytic comprehends several prediction models that need to be developed. Currently, there are two prediction models developed by the author of this research to support population health management, specifically the diagnosis of sepsis and hypertension prediction [54][55]. These insights assist clinicians in the detection and tracking of chronic diseases. The machine learning component is used to build, test, consume, and deploy predictive analytic models on-demand and as requested for the organization. The platform provides self-service dashboards and visualizations that use data from the repositories to drive the decision-making process. The machine learning application layer is one of the essential layers of this platform. Once the data is integrated, aggregated, and normalized in the system, the platform offers a tool to provide knowledge management through the business intelligence interface providing data analysis, design, and training of machine

learning models, as well as development, and management of results-based care indicators or population health management. The platform provides a tool where clinicians, researchers, and scientists can dig the data and get valuable information. Machine learning models can be trained and customized in preconfigured data domains, allowing the storage of the results for future use. Data researchers and scientists can develop advanced tools to obtain information and value of the data stored in the solution, taking advantage of the model design, training, and validation component.

6.7. Platform Model Benefits

The implementation of the proposed platform model could become the healthcare data ecosystem for the organization, providing critical decision-making insights, accurate and reliable healthcare data that significantly increased the value of the healthcare outcome to patients and care clinicians. The platform once built can deliver significant benefits to the organization, such as clinicians having an intelligent application that can be configured to their preferences and optimized to their disciplines, patients receiving more personalized care, an improvement in healthcare workflow and patient care, and personalized care for clinicians and patients. The following subsections presents several use cases that effectively drive change and digital transformation for the organization with the implementation of the proposed platform model.

6.7.1. Reduce of Total Cost of Care for Care Coordination

With a robust data analytic component, the organization will be able to prioritize opportunities for improvement and to improve the way care is coordinated and delivered throughout its network of hospitals and medical facilities. The results can include a considerable increase in financial results. The organization can use the platform to generate timely, meaningful, and actionable data to drive change and improve the quality of care for patients, risk-stratification of the network's population, prioritization of the care coordination activities. Risk stratification will enable care managers to identify individuals at various risk levels for unnecessary services and high-cost utilization, improving patient outcomes and experience. The analytical component also

will reduce unnecessary visits, facilitate access to specialty care, community-based services, and the achievement of healthcare outcomes.

6.7.2. Self-Service Analytics

As described before, the healthcare platform model proposed integrate and standardized data across different source systems to provide actionable insights from a single source of truth. The platform will integrate data from different sources, such as claims data, cost data, financial data, clinical data, and other patient data. With self-service analytics, the organization increases the number of users accessing the analytic component, improving data visibility, and providing actionable insights to improve patient outcomes.

6.7.3. Reduced Deaths from Sepsis

The organization will improve sepsis mortality rates and improve care outcomes by using the advanced analytic component of the platform. Sepsis impacts almost 1.7 million adults in the U.S. and is responsible for nearly 270,000 annual deaths. One-third of all hospital deaths are patients with sepsis [56]. It is still too early to mention the results of the utilization of this feature. Still, the goal of the organization is to reduce its sepsis mortality rate, the costs of the creation of its sepsis care transformation team, and the implementation of an evidence-based sepsis care practice.

6.8. Limitations of the Platform

The health platform will help the organization with closing the gaps between multiple datasets, improving clinical benefits, improving patient's lives, supporting better decision-making to manage larger populations, and improving overall health outcomes. However, the need for algorithms with high accuracy in medical diagnosis is still a challenge that needs to be improved precisely and efficiently [57]. The increasing complexity of building end-to-end platforms to integrate disparate systems and to applying machine learning techniques in specific areas such as computer vision, natural language processing, reinforcement learning, and other generalized

methods present many challenges when conforming the interdisciplinary team needed and the set of technological components used for the implementation.

Some challenges should be considered in the design and implementation of machine learning projects for healthcare. One of the most critical challenges requires algorithms that can answer causal questions. These questions are beyond classical machine learning algorithms because they require a formal model of interventions [58]. To address this type of question from the analytical component of the platform, the algorithms need to learn from the data differently, and to understand how machine learning algorithms need to be trained. Another challenge is to create reliable outcomes from heterogeneous data sources with the participation of subject matter experts (SME) that understand the disease, the machine learning predictive accuracy and correct clinical interpretation depends on the criteria and context of the disease. Clinicians and machine learning engineers should work together on model interpretability and applicability. Machine learning implementation is not an easy task; the selection of predictive features and optimization of hyperparameters is another challenge that needs to be mastered to implement models that provide useful insights [59]. The success and meaningful use of these algorithms, and their integration into the platform depends on the accuracy of the models and their interpretability.

7. RESULTS OF THE ADVANCED ANALYTICS

Analysis of such big data from medical and healthcare systems can be of huge benefit in providing new strategies for healthcare. The latest technological developments in data generation, collection and analysis, have elevated the expectations towards the development of machine learning models for predicting and identifying different diseases such as diabetes mellitus, hypertension, coronary artery disease, renal chronic disease, chronic kidney disease, sepsis and chronic obstructive pulmonary disease. For this research and based on the suggestion from the clinical teams that supported and collaborate on the research, hypertension and neonatal sepsis were the selected conditions to build the machine learning models to be presented as part of the advance analytics module of the platform once implemented.

7.1. Machine learning classification for a hypertensive population

After training and testing the logistic regression model for predicting hypertension, we generated some evaluation metrics to evaluate the classifier. Table 2 shows the confusion matrix with the classification results. A true positive value (730), a true negative value (3407), a false negative (216), and a false positive value (1575). The classification report as shown in Table 3, displays the calculated precision and sensitivity.

Table 2 Confusion Matrix Logistic Regression Model

		Predicted Labels	
		Non-Hypertensive	Hypertensive
True Label	Non-Hypertensive	3407	1575
	Hypertensive	216	730

The test sampling of 5,928 contains 4,982 (84%) non-hypertensive, and 946 (16%) hypertensive. The model shows a sensitivity of 730/946 = 77% and a specificity of 3407/4982 = 68%. The precision of the model was 730/2305 = 32% and the negative predicted value 3407/3623 = 94%. The false negative rate of the model was 216/946 = 22% and a calculated area under the curve of 0.73 (95% CI [0.70 - 0.76]). The model was better at identifying individuals who will not develop hypertension than those that will develop hypertension.

Table 3 Classification Report Logistic Regression Model

Classification Report				
	precision	recall	f1-score	support
Non-Hypertensive	0.94	0.68	0.79	4982
Hypertensive	0.32	0.77	0.45	946
avg / total	0.84	0.7	0.74	5928

7.2. Neural network approach to predict early neonatal sepsis

For the neural network approach to predict early neonatal sepsis, Table 4, shows the confusion matrix with the classification results. Actual class label vs. the predicted ones. True positive value (49), true negative value (95), false negative (12) and false positive value (10).

Table 4 Confusion Matrix Sepsis Model

		Predicted	
		Non-Sepsis	Sepsis
TRUE	Non-Sepsis	95	10
	Sepsis	12	49

And, the classification report showed in Table 5 shows the precision and sensitivity. The sensitivity of the model moderately acceptable due to the imbalanced testing dataset, and there is still a high number of false negatives.

Table 5 Classification Report Sepsis Model

Classification Report			
True Positive	**False Negative**	**Precision**	**Accuracy**
49	12	0.83	0.867
False Positive	**True Negative**	**Recall**	**f1-score**
10	95	0.803	0.817
Positive Label: 1		**Negative Label: 0**	

A sensitivity of 80.3% and a specificity of 90.4% shows that the model might be useful for detecting positive cases, and the true negative rate shows that the model is also efficient at identifying negative cases. The high precision value of 83.1% and the area under the curve of 0.925 confirm the adequacy of the model as a preliminary screening tool. The percentage of positive cases shows that the model works better than random guessing and the conditional probability of negative test results is considerably low. The accuracy of 86.74% shows that the model correctly identifies negative cases and positive cases based on the characteristics of the dataset and the small number of cases examined.

7.3. An Artificial Neural Network Approach for Predicting Hypertension

The results for the artificial neural network are shown in Table 6. Summary of the actual label vs. the predicted. True positive value (887), True negative value (4,477), false negative (1,318) and false positive value (648).

Table 6 Confusion Matrix Neural Network Model

		Predicted	
		Non-Hypertensive	Hypertensive
TRUE	Non-Hypertensive	4477	648
	Hypertensive	1318	887

Table 7 shows the classification report with the sensitivity, precision, and the harmonic mean. The low precision and sensitivity on the hypertensive label are caused because of the large presence of false positives, and imbalanced of the testing data set. The sensitivity of the model moderately acceptable due to the imbalanced testing data set, and this reveals a high number of false negatives.

Table 7 Classification report Neural Network Model

Classification Report			
True Positive	**False Negative**	**Precision**	**Accuracy**
887	1318	0.578	0.732
False Positive	**True Negative**	**Recall**	**f1-score**
648	4477	0.402	0.474
Positive Label: 1		**Negative Label: 0**	

The results indicate a sensitivity of 40%, a specificity of 87%, precision of 57.8% and a measured area under the curve of 0.77 (95% Confidence Interval [75.01 - 79.01]). This research showed results that are to some degree more effectively than a previous study performed by the authors using a statistical model with similar input features that presents a calculated area under the curve of 0.73.

8. CONCLUSIONS AND FUTURE WORK

In this section, the final conclusions of the research are presented including the achievement of the objectives to validate that the initial expectations were met and the future lines of work and research that are open based on the results.

8.1. Verification, Contrast, and evaluation of objectives

The main objective was to develop a conceptual model that presents the components needed for an effective platform to improve population health management, and the implementation of several machine learning models to support the decision-making process in healthcare. This research provides details of an optimized, and secure healthcare platform model that supports the transformation happening in the healthcare industry in Colombia by providing better information to patients and care teams, and in the USA supporting all initiatives for population health management towards a better value-based care model. The proposed health platform model allows to address population health challenges, to understand better the patient's health, and to find hidden patterns that traditional data analytics fail to find. Organizations can use unified patient-centered, financial, and socioeconomic data to detect patterns and to discover groups which share similar health behavior. In addition, the use of this type of platforms reduces the total costs associated with healthcare in all settings and improve the healthcare outcome of the population.

Machine learning algorithms have been developed to predict, evaluate and detect cardiovascular disease cases, increasing the number of individuals that could benefit from the healthy people initiative 2020 of the government of the United States. In this case, a logistic regression model allows a better understanding of several risk factors and how they are associated with the dependent variable. The use of non-invasive risk factors allows creating programs to identify individuals at high risk for hypertension to direct them for treatment. Logistic regression was selected initially because of the easy interpretation of the results for clinical purposes. Important conclusion that emerges from the first model is that removing gender, race and smoke factors

from the data do not affect the accuracy neither the calculated area under the curve of the model. The model showed the best calculated area under de curve value 0.73, indicating fair agreement with the final diagnosis, more work will continue with this model to improve the diagnosis accuracy. After developing the logistic regression model, this research presents a neural network approach to overthrow the non-linearity problem with the risk factors utilized as inputs for the model. The new proposed model improved the accuracy and performance of the logistic regression model that used the same input features. The multi-layer model confirmed the influence of the imbalanced data set to the class with more presence in the data. However, this research showed that the proposed model could be a guide to the design of other models and inference engines for expert systems. The artificial neural network model cannot yet be used to provide final diagnosis in patients due that it requires more clinician's involvement for validation with real patients to reach the desire threshold. Though, the model can be used to make the clinicians aware that there is a probability that the patients could be developing hypertension.

After conversations with clinicians in Colombia, the interest turns over another public health issue, and one of the leading causes of complications and deaths in neonatal intensive care units. Early identification of neonatal sepsis allows clinicians to implement treatments, determine proper antibiotic administration, and potentially reduce associated complications for neonates at neonatal intensive care units. The use of data extracted from the electronic medical records allowed us to create a model with good performance and results when compared with others that used more complex data such as bio-signal data, laboratory results of blood culture, electrocardiogram, and pulse oximeter data. This research presents the use of a neural network models that learn features and make predictions for the detection of early neonatal sepsis. This research also indicates that such model has some limitations in setting the dependent variable, having enough data, and adequate explanatory power. However, the research showed that mothers with premature rupture of membrane, maternal fever, and premature newborn make an evident causal association for early neonatal sepsis. Level of education and marital status showed significant evidence in the appearance of neonatal sepsis. Presence of maternal infectious pathology, such as vaginal infection was a determining factor to explain the cases of premature membrane rupture over the 18 hours.

8.2. Main contributions

The main benefits and contributions arising from this research:

1. **Developed a machine learning model to improve the decision-making process in population health management for hypertensive patients**. In collaboration with clinical experts presents the use of several non-invasive factors to create an accessible and highly interpretable logistic regression prediction model to classify hypertensive patients and study the relevance of each variable in the presence of the others using national health data from the National Health and Nutrition Examination Survey (NHANES). Risk variables were selected after performing a compressive review of studies describing equations to predict hypertension and the clinical suggestions of the subject matter expert that worked with us on the development of the machine learning model. In addition, this research presents a neural network approach to overthrow the non-linearity problem with the risk factors utilized as inputs for the model.
2. **Developed a machine learning model to provide decision support for health care clinicians at neonatal intensive care units**. This research presents a non-invasive prediction model that can be used as an inference engine of smart systems to provide decision support for health care clinicians at neonatal intensive care units. The need of this model was accessed with the clinical team that participates in this research and needed to improve decision support at neonatal intensive care units.
3. **A conceptual design of a healthcare platform model**. A conceptual design was presented to integrate and normalize the data from the organizations and allow the machine learning models to take part in the advanced analytics modules for population health management.

8.3. Derivative works

Research initiatives on predictive modelling based on integrated data are used to find and identify individuals that are most likely to develop risk conditions and therefore, engage them to participate in programs to improve their healthcare outcomes. The analysis of clinical and financial data allows managing patient's health with better accurateness. The healthcare

platform will also allow better health discoveries and actions based on treatment history for individuals and groups of patients. Some work has been completed and published derivate from the initial research.

1. **A Machine Learning Approach for Severe Maternal Morbidity Prediction at Rafael Calvo Clinic in Cartagena-Colombia**. Arrieta Rodríguez E., López-Martínez F., Martínez Santos J.C. (2020) A Machine Learning Approach for Severe Maternal Morbidity Prediction at Rafael Calvo Clinic in Cartagena-Colombia. In: Saeed K., Dvorský J. (eds) Computer Information Systems and Industrial Management. CISIM 2020. Lecture Notes in Computer Science, vol 12133. Springer, Cham. https://doi.org/10.1007/978-3-030-47679-3_18.
2. **Big Data and Machine Learning: A Way to Improve Outcomes in Population Health Management**. Martinez, Fernando Enrique Lopez and Edward Rolando Núñez-Valdez. "Big Data and Machine Learning: A Way to Improve Outcomes in Population Health Management." Protocols and Applications for the Industrial Internet of Things, edited by Cristian González García, et al., IGI Global, 2018, pp. 225-239. http://doi:10.4018/978-1-5225-3805-9.ch008.
3. **IoT and Big Data in Public Health: A Case Study in Colombia**. Martinez, Fernando Enrique Lopez, et al. "IoT and Big Data in Public Health: A Case Study in Colombia." Protocols and Applications for the Industrial Internet of Things, edited by Cristian González García, et al., IGI Global, 2018, pp. 309-321. http://doi:10.4018/978-1-5225-3805-9.ch011.

8.4. Research lines and future work

Organizations such as Keralty and Englewood health, recognized that better care coordination and healthcare management are essential for improving patient's care services and treatments. These organizations wanted to improve quality outcomes, provider engagement, recruitment, and its own economic health. To meet these objectives, these organizations focuses on clinician's engagement and organizational alignment, ensuring widespread access to meaningful, actionable data, and the use of a healthcare analytics platform to inform decisions and drive improvement. Keralty and Englewood health have confidence in the use of big data and machine learning as the most important, life-saving technologies ever introduced to the organization. We

believe the opportunities in these organizations are virtually immeasurable for the implementation of the platform to improve and accelerate clinical, workflow, and financial outcomes. At the end of the research, we were greatly informed that Keralty in partnership with Google Cloud and Google Health will build the healthcare platform based in our proposed model where the implementation approach is to execute an integration exercise lead by Google by reusing healthcare solutions components developed by google at every layer of the model. Keralty with the support of the researcher will continue implementing machine learning models for the advance analytics component and in addition to the diagnostic and predictive models, prescriptive analytics models will be developed to assist the organization in making smarter decisions in population health management. This healthcare platform based in our research will include diagnostic support, inpatient and outpatient care, patient monitoring, medication management, health development and intelligent surgical support, among others. This new platform is expected to be finalized by 2023 with an estimated investment of 12 million dollars in development and 8.8 million dollars in Google Cloud Platform consumption, Google professional services and Google services. The main goal is to provide a better population health management to support population management, high cost pathology prediction models, management of patient care programs, closure of gaps in care and to improve decision-making at the organization.

9. IMPACT FACTOR REPORT FOR JOURNALS

The impact factor report of the four articles used to present the doctoral thesis work in the form of a publication compendium is presented below. The full copy of the articles is contained in the annex to the document.

Title: Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors.

Research specific objective accomplished: Develop a machine learning model to improve the decision-making process in population health management for hypertensive patients.

Authors: Fernando López Martínez, Aron Schwarcz.MD, Edward Rolando Núñez-Valdez, Vicente García-Díaz.

Journal: Expert Systems with Applications

Impact factor JCR: 4.292 5/275 Q1

Volume and Publication date: Volume 110, 15 November 2018

Pages: 206-215

DOI: 10.1016/j.eswa.2018.06.006

Contribution: The results show a sensitivity of 77%, a specificity of 68%, a precision on the positive predicted value of 32% in the test sample and a calculated area under the curve of 0.73. The model also confirms that individuals with obesity, age range between 71 and 80 years old, race non-Hispanic black and male have higher odds of having hypertension. Diabetes, kidney disease and smoking habits do not affect odds of the outcome.

Title: A neural network approach to predict early neonatal sepsis.

Research specific objective accomplished: Develop a machine learning model to provide decision support for health care clinicians at neonatal intensive care units.

Authors: Fernando López Martínez, Jaime Lorduy Gomez, Edward Rolando Núñez-Valdez, Vicente García-Díaz

Journal: Computers & Electrical Engineering

Impact factor JCR: 2.189 30/206 Q1

Volume and Publication date: Volume 76, June 2019

Pages: 379-388

DOI: 10.1016/j.compeleceng.2019.04.015

Contribution: The study results show a sensitivity of 80.32%, a specificity of 90.4%, a precision on the positive predicted value of 83.1% in the test sample and a calculated area under the curve of 0.925 (95% Confidence Interval [0.914 – 0.930]). This neural network model can be used as a smart system's inference engine to support the detection of neonatal sepsis in neonatal intensive care units.

Title: An artificial neural network approach for predicting hypertension using NHANES Data.

Research specific objective accomplished: Develop a machine learning model to improve the decision-making process in population health management for hypertensive patients.

Authors: Fernando López Martínez, Edward Rolando Núñez-Valdez, Rubén González Crespo, Vicente García-Díaz

Journal: Natures Scientific Reports

Impact factor JCR: 4.011 8/111 Q1

Volume and Publication date: Volume 10, 30 June 2020

Pages: 1 - 13

DOI: 10.1038/s41598-020-67640-z

Contribution: This paper focus on a neural network classification model to estimate the association among gender, race, BMI, age, smoking, kidney disease and diabetes in hypertensive patients. It also shows that artificial neural network techniques applied to large clinical data sets may provide a meaningful data-driven approach to categorize patients for population health management, and support in the control and detection of hypertensive patients, which is part of the critical factors for diseases of the heart. Data was obtained from the National Health and Nutrition Examination Survey from 2007 to 2016.

Title: A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population health management.

Research specific objective accomplished: A conceptual design of a healthcare platform model to integrate and normalize the data from the organizations and allow the machine learning models to take part in the advanced analytics modules for population health management.

Authors: Fernando López Martínez, Edward Rolando Núñez-Valdez, Vicente García-Díaz, Zoran Bursac

Journal: Algorithms - Special Issue "Algorithms in Decision Support Systems"

Impact factor SJR: 1.46 20/51 Q3

Volume and Publication date: Volume 13, Issue 4, 23 April 2020

Pages: 102 - 121

DOI: 10.3390/a13040102

Contribution: This paper presents the design of the data health platform and its components in a healthcare organization in Colombia to integrate operational, clinical, and business data repositories with advanced analytics to improve the decision-making process for population health management.

Title: A Machine Learning Approach for Severe Maternal Morbidity Prediction at Rafael Calvo Clinic in Cartagena-Colombia

Research specific objective accomplished: Develop a machine learning model to provide decision support for health care clinicians at neonatal intensive care units.

Authors: Fernando López Martínez, Eugenia Arrieta Rodríguez, Juan Carlos Martínez Santos

Journal: Computer Information Systems and Industrial Management

Impact factor SJR: 1.17 Q2

Volume and Publication date: volume 12133, 22 May 2020

Pages: 208 - 219

DOI: 10.1007/978-3-030-47679-3_18

Contribution: This study presents the results of two machine learning algorithms, logistic regression and support vector machine for Severe Maternal Morbidity Prediction at Rafael Calvo Clinic in Cartagena-Colombia.

Title: Big Data and Machine Learning: A Way to Improve Outcomes in Population Health Management

Research specific objective accomplished: A conceptual design of a healthcare platform model to integrate and normalize the data from the organizations and allow the machine learning models to take part in the advanced analytics modules for population health management.

Authors: Fernando López Martínez, Edward Rolando Núñez-Valdez

Journal: Protocols and Applications for the Industrial Internet of Things

Volume and Publication date: 4 April 2018

Pages: 225-239

Editor: Business Science Reference

Collection: Advances in Business Information Systems and Analytics

DOI: 10.4018/978-1-5225-3805-9.ch008

ISBN: 978-1522538059

Contribution: This study shows a high-level implementation of a complete solution of IoT, big data, and machine learning implemented in the city of Cartagena, Colombia for hypertensive patients by using an eHealth sensor and Amazon Web Services components.

Title: IoT and Big Data in Public Health: A Case Study in Colombia

Research specific objective accomplished: A conceptual design of a healthcare platform model to integrate and normalize the data from the organizations and allow the machine learning models to take part in the advanced analytics modules for population health management.

Authors: Fernando Enrique Lopez Martinez, Maria Claudia Bonfante, Ingrid Gonzalez Arteta, Ruby Elena Muñoz Baldiris

Journal: Protocols and Applications for the Industrial Internet of Things

Impact factor SJR:

Volume and Publication date: 4 April 2018

Pages: 309 – 321

Editor: Business Science Reference

Collection: Advances in Business Information Systems and Analytics

DOI: 10.4018/978-1-5225-3805-9.ch011

ISBN: 978-1522538059

Contribution: This study utilizes data acquisition sensors, large medical datasets, and machine-learning methods to perform predictive analytics in a hypertensive population in Cartagena to assist public health organizations to create proactive care programs to prevent the increase of this disease in Cartagena.

10. REFERENCES

1. Glassman, A.; Giuffrida, A.; Escobar, M.L.; Giedion, U. Chapter 1 Colombia: After a Decade of Health System Reform.; Inter-American Development Bank; Colombia; 2010; p. 188.

2. Ruíz, F.; Gaviria A.; Norman, J. Plan Decenal de Salud Pública; https://www.ins.gov.co/Normatividad/Resoluciones/RESOLUCION%201841%20DE%202013.pdf (accessed on 16 Jan 2020).

3. Legido, H.; Lopez, P.A.; Balabanova, D.; Perel, P.; Lopez-Jaramillo, P.; Nieuwlaat, R.; Schwalm, J.D.; McCready, T.; Yusuf, S.; McKee, M. Patients' knowledge, attitudes, behavior and health care experiences on the prevention, detection, management and control of hypertension in Colombia: A qualitative study. PLoS ONE 2015, 10, doi: 10.1371/journal.pone.0122112.

4. Lopez, F.E; Bonfante, M.C.; Arteta, I.G.; Baldiris, R.E. IoT and big data in public health: A case study in Colombia. In Protocols and Applications for the Industrial Internet of Things; IGI Global, 2018; pp. 309–21, ISBN 978-1-5225-3806-6.

5. World Health Organization World Health Statistics 2017: Monitoring Health for The SDGs; 2017; ISBN 9788578110796.

6. National Center for Health Statistics Health, United States, 2016: With Chartbook on Long-term Trends in Health; May. Report No.: 2017-1232. PMID: 28910066. 2017

7. Nwankwo, T.; Yoon, S.S.; Burt, V.; Gu, Q. Hypertension among adults in the United States: National Health and Nutrition Examination Survey, 2011-2012. NCHS Data Brief 2013, 1–8, doi:10.1017/CBO9781107415324.004.

8. Mozaffarian, D.; Benjamin, E.J. Heart disease and stroke statistics-2016 update a report from the American Heart Association. 2016, 133, e38–e48, doi:10.1161/CIR.0000000000000350.

9. Committee on Public Health Priorities to Reduce and Control Hypertension in the U.S. Population. A Population-Based Policy and Systems Change Approach to Prevent and Control Hypertension; 2010; ISBN 9780309148092.

10. Kublanov, V.S.; Dolganov, A.Y.; Belo, D.; Gamboa, H. Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics. Applied bionics and biomechanics. 2017, 2017, 1–13, doi:10.1155/2017/5985479.

11. Fleischmann-Struzek, C.; Goldfarb, D.M.; Schlattmann, P.; Schlapbach, L.J.; Reinhart, K.; Kissoon, N. The global burden of pediatric and neonatal sepsis: a systematic review. The Lancet Respiratory medicine. 2018, 6, 168–170, doi: 10.1016/S2213-2600(18)30063-8.

12. Nguyen, H.B.; Corbett, S.W.; Steele, R.; Banta, J.; Clark, R.T.; Hayes, S.R.; Edwards, J.; Cho, T.W.; Wittlake, W.A. Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality. Critical care medicine. 2007, 35, 1105–1112, doi: 10.1097/01.CCM.0000259463.33848.3D.

13. Dennis, R.J.; Caraballo, L.; García, E.; Rojas, M.X.; Rondon, M.A.; Pérez, A.; Aristizabal, G.; Peñaranda, A.; Barragan, A.M.; Ahumada, V. Prevalence of asthma and other allergic conditions in Colombia 2009-2010: a cross-sectional study. BMC Pulmonary Medicine 2012, 12, 1–9, doi:10.1186/1471-2466-12-17.

14. About Keralty; Available online: https://www.keralty.com/en/about-keralty (accessed on 27 Jan 2020).

15. About Englewood Health | Englewood Health Available online: https://www.englewoodhealth.org/about-englewood-health (accessed on Jun 13, 2017).

16. Saranya, P.; Asha, P. Survey on Big Data Analytics in Health Care. In Proceedings of the Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019; 2019; pp. 46–51, doi: 10.1109/ICSSIT46314.2019.8987882.

17. Demystifying Big Data: TechAmerica Foundation paper. 2012, 2012, https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf (accessed on 14 February 2018).

18. Remus, S.; Kennedy, M.A. Innovation in transformative nursing leadership: nursing informatics competencies and roles. Nursing Leadership. (Toronto. Ontario). 2012, 25, 14–26, doi:10.12927/cjnl.2012.23260.

19. Yang, S.; Njoku, M.; Mackenzie, C.F. “Big data” approaches to trauma outcome prediction and autonomous resuscitation. British journal of hospital medicine. 2014, 75, 637–641, doi: 10.12968/hmed.2014.75.11.637.

20. Helm-Murtagh, S.C. Use of big data by Blue Cross and Blue Shield of North Carolina. North Carolina Medical Journal. 2014, 75, 195–197, doi:10.18043/ncm.75.3.195.

21. Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine. 2001, 23, 89–109, doi:10.1016/S0933-3657(01)00077-X.

22. Hastie, T. et. all. Springer Series in Statistics the Elements of Statistical Learning; 2009; Vol. 27; ISBN 9780387848570.

23. Dreiseitl, S.; Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. Journal of Biomedical Informatics. 2002, 35, 352–359, doi:10.1016/S1532-0464(03)00034-0.

24. Kononenko, I. Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence. 1993, 7, 317–337, doi:10.1080/08839519308949993.

25. Aspinall, M.J. Use of a decision tree to improve accuracy of diagnosis. Nursing research. 1979, 28, 182–5.

26. Long, W.J.; Griffith, J.L.; Selker, H.P.; D’Agostino, R.B. A comparison of logistic regression to decision-tree induction in a medical domain. Computers and biomedical research. 1993, 26, 74–97, doi:10.1006/cbmr.1993.1005.

27. Rosenblatt, F. The perceptron: A theory of statistical separability in cognitive systems (Project Para). Buffalo, N.Y., Cornell Aeronautical Laboratory. 1958, 1–59.

28. Hoyt, R.E.; Snider, D.; Thompson, C.; Mantravadi, S. IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics. JMIR Public Health Surveillance. 2016, 2, e157, doi:10.2196/publichealth.5810.

29. Google Health Available online: https://health.google/ (accessed on Jan 16, 2020).

30. Grandia, L.; Grandia, B.L. Healthcare Information Systems: A Look at the Past, Present, and Future. Health Catalyst 2017, 1–6. Available online: https://downloads.healthcatalyst.com/wp-content/uploads/2014/05/A-Look-at-the-Past-Present-and-Future-Healthcare-Information-Systems.pdf (accessed on Dec 21, 2019).

31. Arslanian, H.; Fischer, F.; Arslanian, H.; Fischer, F. Future Trends in Artificial Intelligence. In the Future of Finance; Springer International Publishing, 2019; pp. 231–247, doi: 10.1007/978-3-030-14533-0_18

32. Nash, D.B. Predicting Success in Population Health. American health & drug benefits 2019, 12, 323–324. PMID: 32055280; PMCID: PMC6996616

33. Ahmad, W.M.A.W.; Nawi, M.A.B.A.; Aleng, N.A.; Halim, N.A.; Mamat, M.; Pouzi, M. Association of hypertension with risk factors using logistic regression. Appl. Math. Sci. 2014, 8, 2563–2572, doi:10.12988/ams.2014.42130.

34. Manandhar, N. Risk factors of Hypertension: Logistic regression analysis. SCIREA Journal of Health. 2016. doi: 10.3126/njs.v1i0.18818

35. Zheng, Z.; Li, Y.; Cai, Y. The Logistic Regression Analysis on Risk Factors of Hypertension among Peasants in East China & Its Results Validating. Scientific World Journal 2013, 10, 416–420. doi: 10.1155/2014/761486

36. Ramezankhani, A.; Azizi, F.; Hadaegh, F.; Eskandari, F. Sex-specific clustering of metabolic risk factors and their association with incident cardiovascular diseases: A population-based prospective study. Atherosclerosis 2017, 263, 249–256, doi:10.1016/j.atherosclerosis.2017.06.921.

37. Mani, S.; Ozdas, A.; Aliferis, C.; Varol, H.A.; Chen, Q.; Carnevale, R.; Chen, Y.; Romano-Keeler, J.; Nian, H.; Weitkamp, J.-H. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J. Am. Med. Inform. Assoc. 2014, 21, 326–36, doi:10.1136/amiajnl-2013-001854.

38. Griffin, M.P.; Lake, D.E.; Moorman, J.R. Heart Rate Characteristics and Laboratory Tests in Neonatal Sepsis. Pediatrics 2005, 115, 937–941, doi:10.1542/peds.2004-1393.

39. Honoré, A. Machine Learning for Neonatal Early Warning Signs, KTH, Information Science and Engineering, 2017. ISSN 1653-5146

40. Calvert, J.S.; Price, D.A.; Chettipally, U.K.; Barton, C.W.; Feldman, M.D.; Hoffman, J.L.; Jay, M.; Das, R. A computational approach to early sepsis detection. Computers in Biology and Medicine. 2016, 74, 69–73, doi:10.1016/J.COMPBIOMED.2016.05.003.

41. Desautels, T.; Calvert, J.; Hoffman, J.; Jay, M.; Kerem, Y.; Shieh, L.; Shimabukuro, D.; Chettipally, U.; Feldman, M.D.; Barton, C.; et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR medical informatics 2016, 4, e28, doi:10.2196/medinform.5909.

42. Horng, S.; Sontag, D.A.; Halpern, Y.; Jernite, Y.; Shapiro, N.I.; Nathanson, L.A. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017, 12, e0174708, doi:10.1371/journal.pone.0174708.

43. Chapman, Pete, Clinton, Julian, Kerber, Randy, Khabaza, Thomas, Reinartz, Thomas, Shearer, Colin and Wirth, Rudiger CRISP-DM 1.0 Step-by-step data mining guide, The CRISP-DM consortium (2000).

44. Niakšu, O. CRISP Data Mining Methodology Extension for Medical Domain. Balt. J. Modern Computing 2015, 3, 92–109.

45. Azevedo, A.; Santos, M.F. KDD, SEMMA and CRISP-DM: A parallel overview. In Proceedings of the MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008; 2008; pp. 182–185. uri: hdl.handle.net/10400.22/136

46. Park, J.; Lechevalier, D.; Ak, R.; Ferguson, M.; Law, K.H.; Lee, Y.T.T.; Rachuri, S. Gaussian process regression (GPR) representation in predictive model markup language (PMML). Smart and sustainable manufacturing systems. 2017, 1, 121, doi:10.1520/SSMS20160008.

47. Farooqi, M.M.; Shah, M.A.; Wahid, A.; Akhunzada, A.; Khan, F.; ul Amin, N.; Ali, I. Big Data in Healthcare: A Survey. Applications of Intelligent Technologies in Healthcare 2019; pp. 143–152, doi:10.1007/978-3-319-96139- 21_4.

48. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From big data to precision medicine; Frontiers in Medicine 2019, 6, doi:10.3389/fmed.2019.00034.

49. Hatzigeorgiou, M.N.; Joshi, M.S. Population Health Systems: The Intersection of Care Delivery and Health Delivery. Population Health Management 2019, 22, 467-469, doi:10.1089/pop.2019.0066.

50. Koti, M.S.; Alamma, B.H. Predictive analytics techniques using big data for healthcare databases. In Proceedings of the Smart Innovation, Systems and Technologies; Springer Science and Business Media 2019; 105, pp. 679–686, doi:10.1007/978-981-13-1927-3_71.

51. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, Analysis and Future prospects. Journal of Big Data 2019, 6, 54, doi:10.1186/s40537-019-0217-0.

52. Puaschunder, J.M. Big Data, Algorithms and Health Data. SSRN Electronic Journal 2019, doi:10.2139/ssrn.3474126.

53. Moreira, M.W.; Rodrigues, J.J.; Korotaev, V.; Al-Muhtadi, J.; Kumar, N. A Comprehensive Review on Smart Decision Support Systems for Health Care; Institute of Electrical and Electronics Engineers Inc 2019, 13, 3536-3545, doi:10.1109/JSYST.2018.2890121.

54. López-Martínez, F.; Núñez-Valdez, E.R.; Lorduy Gomez, J.; García-Díaz, V. A neural network approach to predict early neonatal sepsis. Computers and Electrical Engineering 2019, 76, 379–388, doi:10.1016/j. compeleceng.2019.04.015.

55. López-Martínez, F.; Schwarcz.MD, A.; Núñez-Valdez, E.R.; García-Díaz, V. Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors. Expert Systems with Applications 2018, 110, 206–215, doi: 10.1016/j.eswa.2018.06.006.

56. Rhee, C.; Dantes, R.; Epstein, L.; Murphy, D.J.; Seymour, C.W.; Iwashyna, T.J.; Kadri, S.S.; Angus, D.C.; Danner, R.L.; Fiore, A.E.; et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. JAMA - Journal of the American Medical Association 2017, 318, 1241–1249, doi:10.1001/jama.2017.13836.

57. Mahindrakar, P.; Hanumanthappa, M. Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges. Int. Journal of Engineering Research and Applications 2013, 3, 937–41, ISSN 2248-9622.

58. Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A.L.; Chen, I.Y.; Ranganath, R. A Review of Challenges and Opportunities in Machine Learning for Health 2018 , https://arxiv.org/abs/1806.00388.

59. Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare; Elsevier Artificial Intelligence in Medicine 2020, 104, 101822, doi:10.1016/j. artmed.2020.101822.

11. PUBLICATIONS

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

An artificial neural network approach for predicting hypertension using NHANES data

Fernando López-Martínez[1,3], Edward Rolando Núñez-Valdez[1], Rubén González Crespo[2]✉ & Vicente García-Díaz[1]

This paper focus on a neural network classification model to estimate the association among gender, race, BMI, age, smoking, kidney disease and diabetes in hypertensive patients. It also shows that artificial neural network techniques applied to large clinical data sets may provide a meaningful data-driven approach to categorize patients for population health management, and support in the control and detection of hypertensive patients, which is part of the critical factors for diseases of the heart. Data was obtained from the National Health and Nutrition Examination Survey from 2007 to 2016. This paper utilized an imbalanced data set of 24,434 with (69.71%) non-hypertensive patients, and (30.29%) hypertensive patients. The results indicate a sensitivity of 40%, a specificity of 87%, precision of 57.8% and a measured AUC of 0.77 (95% CI [75.01–79.01]). This paper showed results that are to some degree more effectively than a previous study performed by the authors using a statistical model with similar input features that presents a calculated AUC of 0.73. This classification model can be used as an inference agent to assist the professionals in diseases of the heart field, and can be implemented in applications to assist population health management programs in identifying patients with high risk of developing hypertension.

Currently, the use of neural network models for disease classification is increasing rapidly, not only because of a significant amount of data available that is being generated by healthcare devices and systems but also for the magnitude of computational resources available for data calculation and processing[1,2]. This immense volume of data is utilized to train models importantly, and facilitates the use of expert systems, NLP techniques[3,4] and classification techniques for finding trends and patterns in the evaluation and classification of several diseases. Hypertension is considered in the group of risk factors for cardiovascular disease that caused 17.7 million deaths in the world in 2015[5–7]. In the United States, hypertension is contemplated as the primary determinant of decease among U.S. adults even with the existence of practical and low-cost treatments[8–10], with significant public health risks and economic implications for U.S. population. The National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics is one of the principal sources for tracking hypertension in the U.S. population[11] with vast amounts of features and data related to cardiovascular diseases.

In this paper, we develop a neural network classification model to predict hypertension with non-invasive risk factors applying healthcare data from the NHANES. A multi-layer neural network architecture was used to identify hypertensive patients at risk of developing hypertension. Our primary goal in this paper was to train a classifier that will identify hypertensive patients in a highly imbalanced NHANES data set. Additionally, we aspire to achieve lower error rate with a neural network architecture compared to the logistic regression model developed in a previous paper[12] by using the same set of input features. The motivation to develop a new model was the non-linearity of the input features, and neural networks are usually trained to treat non-linearity due to the non-linear nature of them[13], making the model more flexible compared to logistic regression.

This paper is organized along these sections. Second section describes related work and literature research of various models that implemented neural networks for cardiovascular disease classification. Third section introduces the elaboration of the model, population, data source, and validation. Fourht section combines statistical and clinical analysis. Fifth section presents our results and limitations. Finally, sixth section presents conclusions and future work.

[1]Department of Computer Science, Oviedo University, C/ Federico Garca Lorca, 33007 Oviedo, Spain. [2]Department of Computer Science and Technology, Universidad Internacional de La Rioja, Av. de la Paz, 137, 26006 Logroño, La Rioja, Spain. [3]Sanitas, 8400 NW 33rd St, Doral, FL 33122, USA. ✉email: ruben.gonzalez@unir.net

Author	Input features	n Total	Type of model	AUC (%)
Artifical neural network models comparison				
LaFreniere et al.[21]	Age, gender, BMI, sys/diast BP, high and low density lipoproteins, triglycerides, cholesterol, microalbumin, and urine albumin creatinine ratio	379,027	Backpropagation neural network	82
Polak and Mendyk[22]	Age, sex, diet, smoking and drinking habits, physical activity level and BMI	159,989	backpropagation (BP) and fuzzy network	75
Tang et al.[23]	Sys/diast BP, fasting plasma glucose, age, BMI, heart rate, gender, WC, diabetes, renal profile	2,092	Feed-forward, back-propagation neural network	76
Ture et al.[24]	Age, sex, hypertension, smoking, lipoprotein, triglyceride, uric acid, total cholesterol, BMI	694	Feed-forwardneural network	81
Lynn et al.[25]	Sixteen genes, age, BMI, fasting blood sugar, hypertension medication, no history of cancer, kidney, liver or lung	22,184 genes, 159 cases	One-hidden-layer neural network	96.72
Sakr et al.[6]	Age, gender, race, reason for test, stress, medical history	23,095	Backpropagation neural network	64
López-Martínez et al.[12]	Age, gender, ethnicity, BMI, smoking history, kidney disease, diabetes	24,434	Three-hidden layer neural network	77

Table 1. Related work.

Related work

We reviewed a few published papers adopting neural networks models to infer hypertension, and some other studies that compared classification performance and accuracy with logistic regression[14–16]. In every paper, the development process, feature selection, ground truth definition, training data sets, test data sets, overfitting prevention, error assessment, and accuracy information were reviewed. We also reviewed if the models were validated or not, either by an unseen data set or by a panel of experts in the domain[17–20].r Some of the models are shown in Table 1.

LaFreniere et al.[21] presented an artificial neural network (ANN) to predict hypertensive patients utilizing the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) data set. The independent features used were age, gender, BMI, systolic and diastolic blood pressure, high and low-density lipoproteins, triglycerides, cholesterol, microalbumin, and urine albumin–creatinine ratio. Confusion matrix and Receiver Operating Characteristic (ROC) curve were utilized to measure the accurateness of the model. This paper used an extensive data set to train the model compared with other studies.

Polak and Mendyk[22] improve and validate an artificial neural network for high blood pressure risk, using data from the Center for Disease Control and the National Center for Health Statistics (CDC-NCHS). The independent features used in this model were age, sex, diet, smoking and drinking habits, physical activity level and BMI index. ROC curve was utilized to measure the accurateness of the model, and they performed a comparison with a logistic regression classification model.

Tang et al.[23] presented an artificial neural network for classification of cardiovascular disease including hypertension; this paper used a Chinese population. Statistical analysis indicated that 14 risk factors showed statistical importance with cardiovascular disease. The ROC curve is utilized to measure the performance of the model.

Ture et al.[24] implemented a multilayer perceptron for the classification of hypertensive patients. The independent features utilized were age, sex, family history of hypertension, smoking habits, lipoprotein, triglyceride, uric acid, total cholesterol, and body mass index (BMI). ROC curve is utilized to measure the accuracy of the model.

Lynn et al.[25] constructed a neural network model to simulate the geneendophenotype-disease relationship for Taiwanese hypertensive males. Sixteen genes, age, BMI, fasting blood sugar, hypertension medication, no history of cancer, kidney, liver or lung. Classification accuracy is utilized to measure the performance of the model.

Sakr et al.[6] built an artificial neural network to compares the performance of it with different machine learning techniques on predicting the risk of developing hypertension. Age, gender, race, a reason for the test, stress tests and medical history used for classification. ROC curve is utilized to measure the accuracy of the model.

We identified other studies for predicting hypertension using ANN, and all of them have advantages and disadvantages. However, the above mentioned are the most relevant. Our paper did not use data from any medical facility as the studies mentioned earlier. However, our model used a data sample more significant that the majority of them, collected from a national examination survey. The number of predictors was small and non-invasive, in comparison with the cited studies that used lab and exam data.

The national examination survey was designed to assess the health and nutritional status of adults and children in the United States, the data on this survey is unique because it combines social determinants of health data such as smoking, alcohol consumption and dietary habits, and physical examinations. This survey emphasized data regarding the prevalence of major diseases and risk factors for diseases for a broader population than just data from a medical facility, which contains only data for a small subset of the population that does not represent the entire picture of significant disease. In addition, historically, disease trends in the United States have been assessed by surveys.

We achieve an AUC of 0.77 which is acceptable and close to all the studies, considering the imbalanced data used in our paper. The results in our paper and the cited studies could be successfully utilized in hypertension classification, and can be included as inference engines in expert systems for hypertension screening tools. Our paper also includes more hidden layers that the others and we determined the number of hidden layers through

cross-validation experiments. Not evidence of the number of layers and hidden nodes selection techniques are present in the studies.

Materials and methods

We present and discuss an alternative approach, using artificial neural network to classify hypertensive patients. We build, trained, and evaluated the model with the Sklearn of Python programing language package[26], Microsoft Cognitive Toolkit (CNTK) from Microsoft, and Azure Machine Learning Studio[27,28].

A cross-sectional analysis comes from the collection of health examination data for a representative sample of noninstitutionalized U.S. residents, questionnaires administered in the home of the residents. The interview collects demographic, health, and nutrition information, as well as information about the household. This examination includes physical measurements such as blood pressure, dental examination, and the collection of blood and urine specimens for lab testing.

Data source. We collected NHANES data sets from NHANES 2007–2008 to NHANES 2015–2016. This dataset is intended for public access and health care utilization. This datasets are prepared and published through the Centers for Disease Control and Prevention (CDC) to provide full access. Statistic characterizing human populations, laboratory data, blood pressure, body measures data and questionnaires linked to diabetes, smoking, and kidney conditions are part of the data set. The original data set consists of five folders from 2007 to 2016, each one of them contains a pdf file with statistics of the response rates of the NHANES survey and the SAS Transport files for all the survey measure variables. After imported the original data sets in python, data extraction and transformation were necessary to select and categorize the input features. We created a repository in Github with the original files from NHANES, the final data set used to run the model and the notebooks used for the data preparation[29].

Ethic review board approval. For the use of NHANES data, the Institutional Review Board (IRB) approval and documented consent was obtained from participants. The description of the survey name and data, and the NCHS IRB/ERB Protocol Number or Description can be found in Centers for Disease Control and Prevention[30]. In 2003, the NHANES Institutional Review Board (IRB) changed its name to the NCHS Research Ethics Review Board (ERB). The National Center for Health Statistics (NCHS) offered downloadable public-use data files through the Centers for Disease Control and Preventions (CDC) FTP file server. NHANES survey is a public-use data files prepared and disseminated to provide access to the full scope of the data. This allows researchers to manipulate the data in a format appropriate for their analyses. In our study, by using these data we signify our agreement to comply with the data use restrictions to ensure that the information will be used solely for statistical analysis or reporting purposes. The data use restrictions can be found in National Center for Health Statistics[31]. In this study, all experiments were performed in accordance with relevant guidelines and regulations.

Study population and analysis. Healthcare survey data collected in the course of 2007–2016 was used to train and evaluate the classification model. A neural network model was developed to assess the importance of several factors and their relation with prevalence of hypertension with a symbolical sampling of adults ≥ 20 years in the United States (n = 24,434). Table 2 shows the grouping of the hypertensive patients by race and gender.

Input features. Several studies in the US integrated healthcare system in cardiovascular research with incident hypertension showed association between race, age, smoking, BMI, diabetes, and kidney conditions with hypertension[32–34]. Among different cohorts of patients with hypertension, during the follow up, individuals present more kidney disease, diabetes problems and remarkable association with smoking habits. In addition, these studies shown that effective BMI management decrease the incidence of hypertension, hypertension prevalence increase with age, and race is a significant factor of prevalence of hypertension.

For this paper, and based on the previous analysis, the selected input features are race, age, smoking, body mass index (BMI), diabetes, and kidney conditions. Participants admit to have diabetes if the answer presents "Yes or "Borderline" to the question "Doctor told that you have diabetes?"[35]. Smokers defined as individuals having smoked ≥ 100 cigarettes during their lifetime, and currently smoke some days or every day[36]. Chronic kidney disease (CKD) defined as "yes" response to the question "Have you ever told by a health care provider you have weak or failing kidneys?" during the interview, and for NHANES 2015–2016, CKD defined as a glomerular filtration rate (GFR) ≤ 60 ml/min/1.73 m^2[37] and albumin–creatinine ratio ≥ 30 mg/g[38]. Body mass index and age transformed from continues features to Categorical features to understand the relationship among the features. Blood pressure is utilized to generate the dependent feature.Hypertension category designated as systolic blood pressure of ≥ 130 mmHg, previously define as ≥ 140 mmHg by the American Heart Association[39]. Table 3 show the independent features and the dependent feature.

Features selection. Clinical importance was pertinent plus the statistical significance of the features to choose the final inputs. For this paper, we utilized chi-square because previous work indicates that this statistical test performs well to evaluate sets of categorical features[40–43]. Some heuristic methods were investigated to confirm the usefulness and relevance of the features. Genetic algorithm with other machine learning methods probably generates better results[42], and produce adequate time complexity to find optimal solutions[44,45]. We will consider it and discuss it in forthcoming studies. For this paper, we utilized statistical methods between each feature due to the nature of the inputs. Table 4 shows all features with their p values and scores. Based on the clinical significance of all input features, our clinical expert decided not to exclude any variable from the paper

Class	Gender	Race	n
Hypertension, adults 20 and over: 2007–2016			
Hypertensive	Female	Mexican American	464
		Non-Hispanic black	925
		Non-Hispanic white	1,433
		Other Hispanic	368
		Other race—including multi-racial	277
	Male	Mexican American	575
		Non-Hispanic black	1,039
		Non-Hispanic white	1,582
		Other Hispanic	371
		Other race—including multi-racial	365
Non-hypertensive	Female	Mexican American	1,461
		Non-Hispanic black	1,676
		Non-Hispanic white	3,663
		Other Hispanic	1,084
		Other race—including multi-racial	1,038
	Male	Mexican American	1,275
		Non-Hispanic black	1,465
		Non-Hispanic white	3,585
		Other Hispanic	820
		Other race—including multi-racial	968
		Total	24,434

Table 2. n samples by hypertensive class, gender and race.

Variable name	Description	Code	Meaning
Gender	Gender	1	Male
		2	Female
Agerange	Age at screening adjudicated—date of birth was used to calculate AGE	1	20-30
		2	31–40
		3	41–50
		4	51–60
		5	61–70
		6	71–80
Race	Race/Hispanic origin	1	Mexican American
		2	Other Hispanic
		3	Non-Hispanic white
		4	Non-Hispanic black
		5	Other race—including multi-racial
BMXBMI	Body mass index (kg/m^2)	1	Underweight = < 18.5
		2	Normal weight = 18.5–24.9
		3	Overweight = 25–29.9
		4	Obesity = BMI of 30 or greater
Kidney	Ever told you had weak/failing kidneys	1	Yes
		2	No
Smoke	Smoked at least 100 cigarettes in life	1	Yes
		2	No
Diabetes	Doctor told you have diabetes	1	Yes
		2	No
		3	Borderline
Hypclass	Systolic: blood pres (mean) mm Hg	0	Non-hypertensive
		1	Hypertensive

Table 3. Variables included in the model.

Feature	*p* value	Score
Gender	0.3988107	0.711909
Agerange	0.000000	1,965.607023
Race	0.008822	6.858521
BMIrange	0.0172385	5.67193
Kidney	0.3546428	0.856775
Smoke	0.0975246	2.745566
Diabetes	0.0012164	10.465222

Table 4. Chi-squared between each variable.

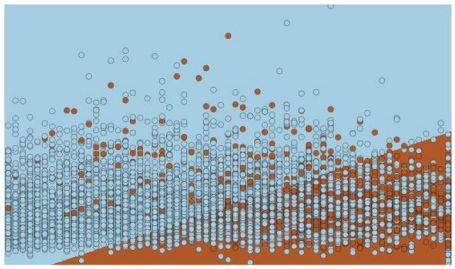

Fig. 1. Decision boundary.

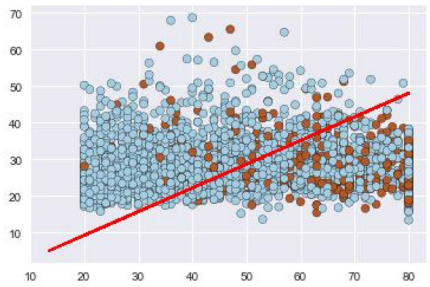

Fig. 2. Draw test points.

due to the relationship among features in previous studies. In addition, one of the options we used to get the feature importance or the influence of a given parameter in the classification model is to obtain and examine the coefficient of the parameters with the provided dataset as shown in López-Martínez et al.[12] Coefficients and Odds Ratio. Features using the same scale with larger coefficients are more important because they represent more significant changes in the dependent variable.

Neural network model. An artificial neural network describes a machine learning algorithms that are made of layers of nodes, usually, an input layer, hidden layers[46] and an output layer[47]. Figure 4 shows the form of the neural network architecture developed for this paper. The input nodes values are encoded and normalized with gaussian normalization[48] in order to improve the computation.

The non-linearity of the model. The motivation of developing this neural networks model is the ability to use non-linear activation functions to eliminate the non-linearity of the input features. The data used to train the model is not linearly separable, this means that there is no line that separates the data points easily as shown in Figs. 1 and 2 where we plot several input variables and the decision boundary using logistic regression.

This non separability can be seen also if we plot three input variables such as gender, age and bmi as shown in Fig. 3. The Neural network model learn a new representations of the data which makes the classification more approachable with respect to this representation. A non-linear activation function allows non-linear classification with a non-linear decision boundary which will be a hyperplane that is orthogonal to the line.

Model training. The generated probabilities need to be as proximate as possible to the observed features. The loss function calculated as the difference between the learned model against the generated by the training

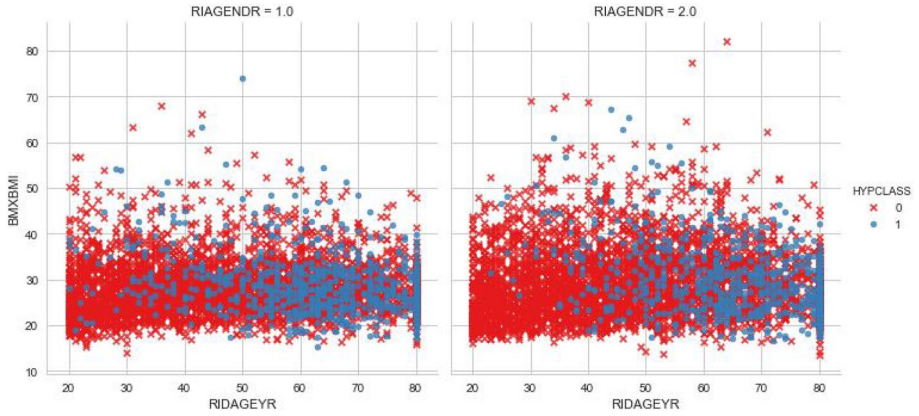

Fig. 3. Relation between BMI and age by gender and hypertension class.

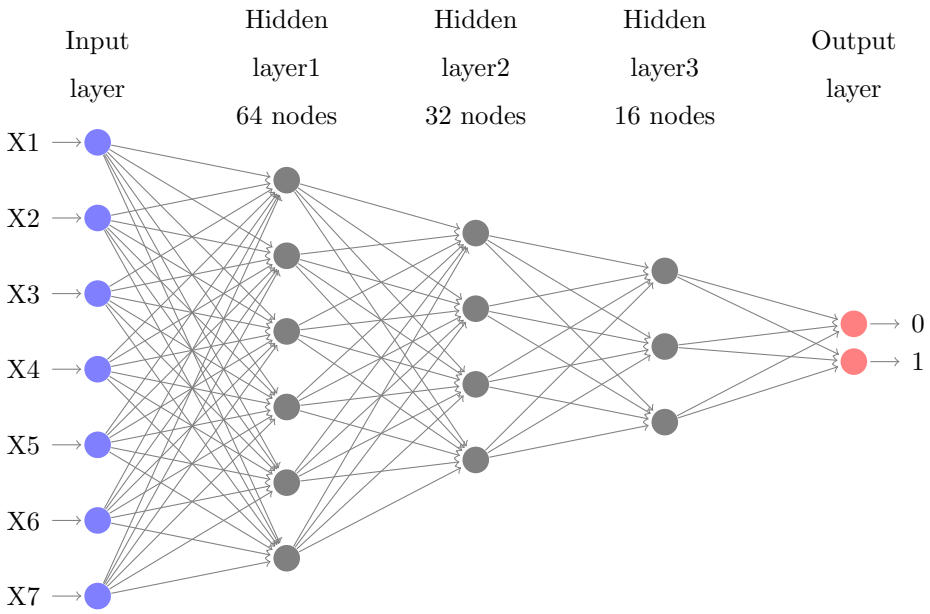

Fig. 4. Multilayer perceptron architecture.

set[49]. Cross-entropy with Softmax is utilized, but the mathematical computation of the derivative is not presented in this paper[50] (Fig. 4).

Random initialization of the parameters is the first step, and the network produces a new set of parameters after each evaluation. In this network, He initialization[51] is utilized. This type of initialization is comparable with the Xavier initialization excepting Xavier uses a different weights scaling factors W in layer l, and the author recommended for layers with ReLU activation. Mini-batches are utilized to train the model. Learning rate[52] is a factor that balances how much the parameters change in every iteration. Each iteration works on ten samples, and the model is trained on 70% (17,104) of the data set. Table 5 presents the parameters of the architecture, and Figs. 5 and 6 present the training loss and classification error of the mini-batch run for the model.

Model evaluation. To evaluate the classification model, computation of the average test error is utilized. The algorithm finds the position of the highest value in the output array, and compares it to the actual label. The evaluation of the network is performed on data never used for training, and coincide with the 30% (7,331) of the data set. The resulting error is compared with the training error and the results indicates that the model presents a useful generalization error. Our model can meritoriously deal with unseen observations, and this is one of the keys to avoiding overfitting[53].

For each observation, our model use softmax as the evaluation function that returns the probability distribution across all the classes. In our paper, it would be a vector of 2 elements per observation. The output nodes in our model convert the activations into a probability and map the aggregated activations across the network to probabilities across the two classes. Figure 7 shows the test classification error for our model.

Parameter	Value
Input dimension	7
Num output classes	2
Num hidden layers	3
Hidden layer1 dimension	64
Activation func layer1	Relu
Hidden layer2 dimension	32
Activation func layer2	Relu
Hidden layer3 dimension	16
Activation func layer3	Relu
Minibatch size	10
Num samples to train	17,104
Num minibatches to train	1,710
Loss function	Cross entropy with softmax
Eval error	Classification error
Learner for parameters	Momentum sgd
Learning rate	0.01
Momentum	0.9
Eval metrics	Confusion matrix, AUC

Table 5. Model architecture parameters.

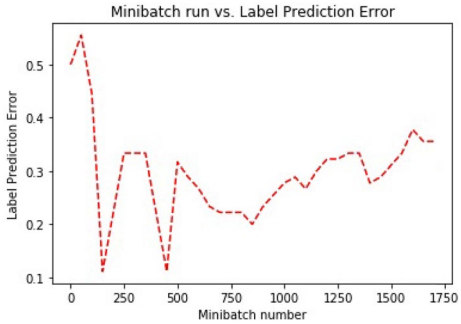

Fig. 5. Training error.

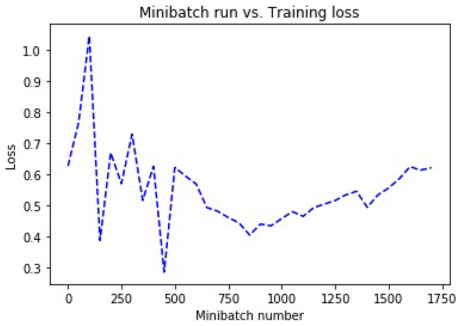

Fig. 6. Loss error.

In this paper, we utilized a few evaluation metrics to evaluate the model. Results are show in Table 6. Summary of the actual label versus the predicted. True positive value (887), True Negative value (4,477), False Negative (1,318) and False Positive value (648).

Table 7, shows the classification report with the sensitivity, precision, and the harmonic mean. The low precision and sensitivity on the hypertensive label is caused because of the large presence of false positives, and imbalanced of the testing data set.

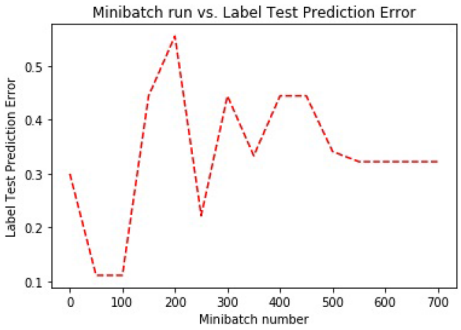

Fig. 7. Test prediction error.

	Predicted	
	Non-hypertensive	Hypertensive
True		
Non-hypertensive	4,477	648
Hypertensive	1,318	887

Table 6. Confusion matrix.

True positive	False negative	Precision	Accuracy
887	1,318	0.578	0.732
False positive	True negative	Recall	f1-score
648	4,477	0.402	0.474
Positive label: 1		Negative label: 0	

Table 7. Classification report.

The sensitivity of the model moderately acceptable due to the imbalanced testing data set, and this reveals a high number of false negatives.

Comparison of the model with alternative techniques

Conducting data analysis in a highly imbalanced data set is not trivial and often leads to obtain low sensitivity results. A comparison of our proposed ANN with other highly interpretable methods will allow us to compare the AUC curves of the models and to validate the performance and sensitivity of disease diagnostic. Five machine learning algorithms were identified and compared using the NHANES data. These algorithms not only accurately classify hypertensive patients, but were able to identify key features useful for hypertension diagnostic. A Two-Class Decision Jungle that represents an ensemble of decision directed acyclic graphs (DAGs), a Two-Class Boosted Decision Tree, an ensemble learning method. A Two-Class Bayes Point Machine, that efficiently approximates the Bayesian average of linear classifiers by choosing one "average" classifier, the Bayes Point. A Two-Class Support Vector Machine, and a Two-Class Logistic Regression.

Two-class decision jungle. This algorithm is a development of decision forest that lie on ensemble of decision directed acyclic graphs (DAGs), used to obtain accurate classifiers[54]. Decision jungles are very comparable to random forests, but it uses DAGs instead of trees as base learners. This structure is more memory-efficient because it eliminates the need for repeating leaf nodes, but it needs more computing time.

This method is selected because decision jungles are models that can express non-linear selection boundaries, and they are strong in the existence of noisy features. Table 8 shows the parameters and Table 14 shows the classification report.

Two-class logistic regression. Logistic regression is a well-known classification technique especially used for classification tasks. The algorithm tries to find the optimal values by maximizing the log probability of the parameters given the inputs. Maximization is performed by using a method for parameter estimation called Limited Memory BFGS[55]. Table 9 shows the parameters and Table 14 shows the classification report.

Parameter	Value
Two-class decision jungle parameters	
Resampling method	Bagging
Trainer mode	Single parameter
Number of decision DAGs	8
Maximum depth of the decision DAGs	32
Maximum width of the decision DAGs	128
Number of optimization steps per layer	2,048

Table 8. Decision jungle parameters.

Parameter	Value
Two-class logistic regression parameters	
Optimization tolerance	1.00E−07
L1 regularization weight	1
L2 regularization weigh	1
Memory size for L-BFGS	20

Table 9. Logistic regression parameters.

Parameter	Value
Two-class support vector machine parameters	
Lambda—weight for L1 regularization	1.00E−03
normalize features before training	Yes

Table 10. Support vector machine parameters.

Parameter	Value
Two-class boosted decision tree parameters	
Maximum number of leaves per tree	20
Minimum number of training instances	10
Learning rate	0.2
Number of trees constructed	100

Table 11. Boosted decision tree parameters.

Two-class support vector machine. The Support Vector Machine algorithm is a supervised learning model that evaluates input data in a multi-dimensional label zone called the hyperplane. The inputs are points in this zone or space, and are mapped to outputs that are divided as clear as possible[56]. Table 10 shows the parameters and Table 14 shows the classification report.

Two-class boosted decision tree. A boosted decision tree, ensemble learning method where the trees corrects for the errors of the previous trees. Predictions are established on the full ensemble of trees[57]. Table 11 shows the parameters and Table 14 shows the classification report.

Two-class Bayes point machine. This method approximates the optimal Bayesian average of linear classifiers by choosing the Bayes Point. This method created by Microsoft Research has shown that no external hyper-parameters are needed and the model can be trained in a single pass, without over-fitting, and without needing pre-processing steps such as data re-scaling[58]. Table 12 shows the parameters and Table 14 shows the classification report.

Synthetic minority oversampling technique. In addition to the previous methods, we have decided to solve the imbalance problem by using the Synthetic Minority Oversampling Technique (SMOTE) and compare the performances with all the methods. This statistical technique increase the number of underrepresented cases

Parameter	Value
Number of training iterations	30
bias to be added to each instance in training	Yes

Table 12. Bayes point machine parameters.

Parameter	Value
SMOTE percentage	200
Number of nearest neighbors	5

Table 13. Synthetic minority oversampling parameters.

Method	True positive	False negative	False positive	True negative	Precision	Accuracy	Recall	f1-score
Our model	887	1,318	648	4,477	0.578	0.732	0.402	0.474
Decision jungle	540	912	390	3,045	0.581	0.734	0.372	0.453
Logistic regression	557	895	389	3,046	0.589	0.737	0.384	0.465
Support vector machine	556	896	387	3,048	0.59	0.737	0.383	0.464
boosted decision tree	568	884	439	2,996	0.564	0.729	0.391	0.462
Bayes point machine	543	909	388	3,047	0.583	0.735	0.374	0.456
Synthetic minority oversampling	3,645	789	1,326	2,086	0.73	0.73	0.82	0.77
Positive label: 1					Negative label: 0			

Table 14. Classification report.

Method	Precision	Accuracy	f1-score	AUC
Our model	0.578	0.732	0.474	0.77
Decision jungle	0.581	0.734	0.453	0.769
Logistic regression	0.589	0.737	0.465	0.764
Support vector machine	0.59	0.737	0.464	0.759
Boosted decision tree	0.564	0.729	0.462	0.765
Bayes point machine	0.583	0.735	0.456	0.763
Synthetic minority oversampling	0.73	0.73	0.77	0.8

Table 15. Classification methods comparison.

in the dataset used in the study. This method returns a dataset that contains the original samples, plus an additional number of synthetic minority samples. In our case we have increase the number of cases 200% (module doubles the percentage of minority cases compared to the original dataset) and the number of nearest neighbors used was 5 (A nearest neighbor is a a case that is very similar to some target case. The distance between any two cases is measured by combining the weighted vectors of all features)[59]. Increasing the number of cases using this technique is not guaranteed to produce more accurate results. We experimented with different percentages, different feature sets, and different numbers of nearest neighbors to find the best results. Table 13 shows the parameters and Table 14 shows the classification report.

Table 15 presents the result of comparing six methods with our ANN proposed method. We presented the AUC and the corresponding accuracy rates. We observed that the accuracy of the methods are very similar with imbalanced dataset, but the AUC and f1-score of our method are slightly higher and competitive; except for the technique used to solve the imbalance of the dataset which shows a higher AUC. however, no one is statistically better than the others. We utilized cross-validation to measure the performance of the models, and performed several train-score-evaluate operations (10 folds) on different subsets of the input data. An statistical significance test, proposed by Giacomini and White[60] was utilized. Where the predictive ability of the presented model for the Cross-entropy loss function showed better performance than the others. We performed a pairwise test of predictive ability of the five models using the Cross-entropy loss function. Table 16 shows the results for the

	DJ	LR	SVM	BDT	BPM
ANN	0.001- (3.65)	0.035- (1.80)	0.001- (4.03)	0.036- (1.80)	0.011- (1.67)

Table 16. Predictive ability tests.

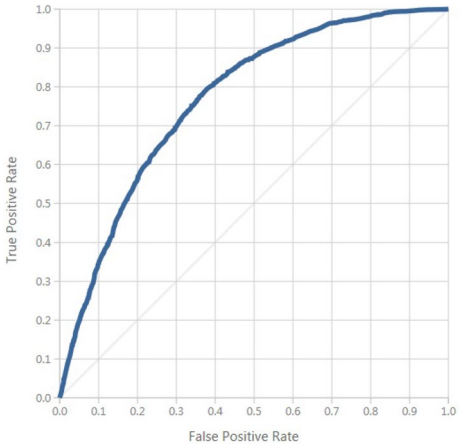

Fig. 8. ROC curve.

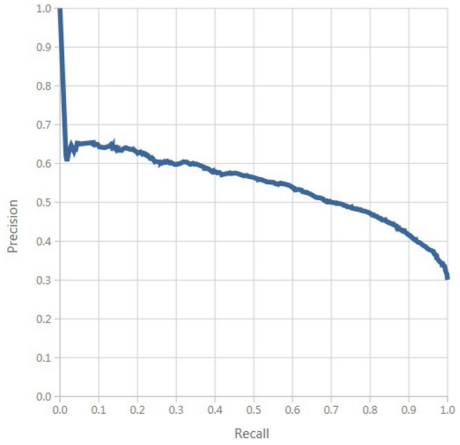

Fig. 9. Precision/recall.

cases. The minus sign indicates that the model under performs the method in the row at the 5% significance level, evidenced by the value in parenthesis (critical value) greater than 1.

Results and analysis

The classification of hypertensive patients was carried out by using artificial neural network with back-propagation. Statistical and clinical analysis were performed to explain the results.

Statistical analysis. In our paper, we used a non-linear model to determine the unidentified non-linearity of the input. The test data set of 7,330 includes 5,125 (69.9%) non-hypertensive samples and 2,205 (30.,1%) hypertensive samples. The model shows a sensitivity of 887/2,205 = 40% (positives properly classified), and a specificity of 4,477/5,125 = 87% (negatives properly classified).

A positive predicted value of 887/1,535 = 57%, and a negative predicted value of 4,477/5,795 = 77%. A false negative rate of 1,318/2,205 = 59%, and a false positive rate of 648/5,125 = 12%. A false alarm of 12%, and a likelihood ratio for negative patients of 0.68%. In this paper, the multi-layer neural network model exceed at classifying patients who will not develop hypertension than those that will develop hypertension. The area under the curve is shown in Fig. 8. Figure 9 shows the proportion of the true results of overall positives results in contrast with the fraction of all correct results.

Clinical analysis. With a sensitivity of 40%, and specificity of 87%, the artificial neural network model demonstrates that it might be ineffective for healthcare diagnosis in detecting positive occurrences, but the true negative rate demonstrates that the model is effective at finding non-hypertensive patients. The high negative predicted value of 77% shows that our model can be used as an examination tool. The positive cases of 57% shows that our model is superior to a random inference with a low probability for negative test results. This model correctly identifies non-hypertensive patients with an accuracy of 73%.

Discussion and limitations

Even though a multi-layer neural network with one layer can model a vast variety of problems in the clinical domain, in our paper, a model with three hidden layers was advantageous to approximate the highly non-linear behavior of the input features. The result of the model was affected by the imbalanced data set, but we did not balance it to maintain the real distribution of the samples.

The current model configuration and size of the data source captured the complexity of the data. We used data augmentation to generate more input data from the already collected data, but the model was over-fitting and learned too many specific details about the training set. We reduced the number of training iterations to prevent over-fitting, and the accuracy was the same as the model without data augmentation.

The paper has shown that the classification capability of the model improved (AUC—0.77), based on the results of the statistical model utilized in a previous paper (AUC—0.73) when applied to the input features gender, race, BMI, age, smoking, kidney conditions and diabetes. However, challenges in applying artificial neural networks to the clinical domain remain. The use of deep learning to analyze hypertension risk features can be considered as complementary for the traditional approach and might be considered to validate other statistical models.

Our model achieved an AUC of 0.77 and used a smaller network architecture than the architecture used by Polak and Mendyk[22] obtaining an AUC of 0.73 and[23] that achieves an AUC of 0.76. However, our model presents a bigger network than[24] that developed a network that achieves an AUC of 0.81, LaFreniere et al.[21] achieves an AUC of 0.82 and Lynn et al.[25] achieves an AUC of 0.96.

One of the significant limitations of our model is that it was developed using a highly imbalanced data set from the CDC to which a high prevalence of non-hypertensive patients was observed. There was no significant increase in accuracy. And we are not relying on this measure. We have a severe class imbalance, and the model will maximize accuracy by simply always picking the most common class.

Therefore, our model must be validated in other clinical settings, and further studies should include other neural network architectures and socio-demographic information[61] to improve the precision and recall of the model, and to consider the integration of this model to the clinical diagnostic scheme. Also, adequate training data volume will be needed to train a bigger model to improve the classification results.

Conclusions and future work

This paper presents a neural network approach to overthrown the non-linearity problem with the risk factors utilized as inputs for the model. This paper shows that the proposed model improves the accuracy and performance of a previous paper that used the same input features and the results were better than logistic regression in a small percentage. The main contribution of this paper is the developed model and based on results showed that ANN was the proper model compared with the previously developed LR model.

Our multi-layer model confirmed the influence of the imbalanced data set to the class with more presence in the data. This paper showed that the proposed model could be a guide to the design of other models and inference engines for expert systems. This model cannot be used yet to provide the final diagnosis in developing hypertension in patients due that it requires clinician's involvement for validation with real patients. However, this model can be used to make them aware that there is a probability that the patients could be developing hypertension.

Knowledge of the current model, and their parameters on risk-prediction models in general, is constructive to determine how to best approach the build of prediction models for hypertension, design the study, and interpret its results. When there is a realistic chance to find an applicable positive effect on decision-making and patient outcome, this model on a new setting could be potentially useful. This paper outlines the process for the development of a neural network risk prediction model, from choosing a data source and selecting features to assess model performance, performing validation, and assessing the impact of the model outcomes.

For future work, this model will be apply and re-train if necessary to a real balanced data set, and bigger network architectures will be considered. Also, new risk factors can be used to better handle the distribution and behavior of the input features for the model, and a sensitivity analysis will be performed to determine which inputs in our ANN model are significant with respect to the output. This paper will be the ground for the construction of a decision supporting tool that may be useful to healthcare practitioners for contributing to decision making about the risk of developing hypertension in typical or atypical patient screening circumstances.

Received: 12 October 2019; Accepted: 9 June 2020
Published online: 30 June 2020

References

1. Vijayarani, M. Liver disease prediction using SVM and Naïve Bayes algorithms. *Int. J. Sci. Eng. Technol. Res.* **4**, 816–820 (2015).
2. Lakshmanaprabu, S. K. *et al.* Online clinical decision support system using optimal deep neural networks. *Appl. Soft Comput.* **81**, 105487. https://doi.org/10.1016/j.asoc.2019.105487 (2019).
3. Sandoval, A. M., Díaz, J., Llanos, L. C. & Redondo, T. Biomedical term extraction: NLP techniques in computational medicine. *Int. J. Interact. Multimed. Artif. Intell.*https://doi.org/10.9781/ijimai.2018.04.001 *(2018)* **(in the press)**.

4. Bobak, C. A., Titus, A. J. & Hill, J. E. Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets. *Appl. Soft Comput. J.* **74**, 264–273. https://doi.org/10.1016/j.asoc.2018.10.005 (2019).
5. World Health Organization. *World Health Statistics 2017: Monitoring Health for The SDGs*. arXiv:1011.1669v3 (2017).
6. Sakr, S. *et al.* Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford exercise testing (FIT) Project. *PLoS ONE* **13**, e0195344. https://doi.org/10.1371/journal.pone.0195344 (2018).
7. Park, J. *et al.* Patient-level prediction of cardio-cerebrovascular events in hypertension using nationwide claims data. *J. Med. Internet Res.* **21**, e11757. https://doi.org/10.2196/11757 (2019).
8. National Center for Health Statistics. *Health, United States, 2016: With Chartbook on Long-term Trends in Health*. Technical Report (2017).
9. Gu, A., Yue, Y., Kim, J. & Argulian, E. The burden of modifiable risk factors in newly defined categories of blood pressure. *Am. J. Med.* **131**, 1349-1358.e5. https://doi.org/10.1016/j.amjmed.2018.06.030 (2018).
10. Li, Y. *et al.* Impact of healthy lifestyle factors on life expectancies in the us population. *Circulation* **138**, 345–355. https://doi.org/10.1161/CIRCULATIONAHA.117.032047 (2018).
11. David, F., Howard, K., Roux Ana, D. & Jiang, H. *A Population-Based Policy and Systems Change Approach to Prevent and Control Hypertension* (National Academies Press, Washington, DC, 2010).
12. López-Martínez, F., Schwarcz, M. D. A., Núñez-Valdez, E. R. & García-Díaz, V. Machine learning classification analysis for a hypertensive population as a function of several risk factors. *Expert Syst. Appl.* **110**, 206–215. https://doi.org/10.1016/j.eswa.2018.06.006 (2018).
13. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.*https://doi.org/10.1016/S1532-0464(03)00034-0 *(2002).*
14. Seidler, T. *et al.* A machine learning approach for the prediction of pulmonary hypertension. *J. Am. Coll. Cardiol.* **73**, 1589. https://doi.org/10.1016/s0735-1097(19)32195-3 (2019).
15. Ambale-Venkatesh, B. *et al.* Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ. Res.* **121**, 1092–1101. https://doi.org/10.1161/CIRCRESAHA.117.311312 (2017).
16. Mortazavi, B. . J. *et al.* Analysis of machine learning techniques for heart failure readmissions. *Circ. Cardiovasc. Qual. Outcomes* **9**, 629–640. https://doi.org/10.1161/CIRCOUTCOMES.116.003039 (2016).
17. Debray, T. P. A. *et al.* A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.* **68**, 279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018 (2015).
18. Tengnah, M. A. J., Sooklall, R. & Nagowah, S. D. A predictive model for hypertension diagnosis using machine learning techniques. In *Telemedicine Technologies* (eds Jude, H. D. & Balas, V. E.) 139–152 (Academies Press, Elsevier, 2019). https://doi.org/10.1016/b978-0-12-816948-3.00009-x.
19. Clim, A., Zota, R. D. & Tinica, G. The Kullback–Leibler divergence used in machine learning algorithms for health care applications and hypertension prediction: a literature review. *Procedia Comput. Sci.* **141**, 448–453. https://doi.org/10.1016/j.procs.2018.10.144 (2018).
20. Singh, N., Singh, P. & Bhagat, D. A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. *Expert Syst. Appl.* **130**, 188–205. https://doi.org/10.1016/j.eswa.2019.04.029 (2019).
21. LaFreniere, D., Zulkernine, F., Barber, D. & Martin, K. Using machine learning to predict hypertension from a clinical dataset. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7. https://doi.org/10.1109/SSCI.2016.7849886 (2016).
22. Polak, S. & Mendyk, A. Artificial neural networks based Internet hypertension prediction tool development and validation. *Appl. Soft Comput.* **8**, 734–739. https://doi.org/10.1016/j.asoc.2007.06.001 (2008).
23. Tang, Z.-H. *et al.* Comparison of prediction model for cardiovascular autonomic dysfunction using artificial neural network and logistic regression analysis. *PLoS ONE* **8**, e70571. https://doi.org/10.1371/journal.pone.0070571 (2013).
24. Ture, M., Kurt, I., Turhan Kurum, A. & Ozdamar, K. Comparing classification techniques for predicting essential hypertension. *Expert Syst. Appl.* **29**, 583–588. https://doi.org/10.1016/j.eswa.2005.04.014 (2005).
25. Lynn, K. S. *et al.* A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data. *Bioinformatics* **25**, 981–988. https://doi.org/10.1093/bioinformatics/btp106 (2009).
26. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2 (2012).
27. Team, A., Dorard, L., Reid, M. . D. & Martin, F. . J. AzureML: anatomy of a machine learning service. *JMLR Workshop Conf. Proc.* **50**, 1–13 (2016).
28. Seide, F. & Agarwal, A. Cntk. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD '16*, 2135–2135. https://doi.org/10.1145/2939672.2945397 (ACM, New York, NY, USA, 2016).
29. López-Martínez, F. Deep learning hypertension model repository. https://github.com/sysdevelopment/phd (2018). Accessed July 2019.
30. Centers for Disease Control and Prevention. NHANES—NCHS Research Ethics Review Board Approval.
31. National Center for Health Statistics, C. Data Access—Data User Agreement (2017).
32. Daugherty, S. L. *et al.* Age-dependent gender differences in hypertension management. *J. Hypertens.* **29**, 1005–1011. https://doi.org/10.1097/HJH.0b013e3283449512 (2011).
33. Dye, B. A., Thornton-Evans, G., Li, X. & Iafolla, T. J. *Key findings Data from the National Health and Nutrition Examination Survey, 2011–2012*. Technical Report, Vol. 197 (2011).
34. Ong, K. L., Tso, A. W., Lam, K. S. & Cheung, B. M. Gender difference in blood pressure control and cardiovascular risk factors in Americans with diagnosed hypertension. *Hypertension* **51**, 1142–1148. https://doi.org/10.1161/HYPERTENSIONAHA.107.105205 (2008).
35. HSS. Awareness of Prediabetes—United States, 2005–2010. *Centers for Disease Control & Prevention Source: Morbidity and Mortality Weekly Report Centers for Disease Control & Prevention*, Vol. 62, 209–212 (2005).
36. CDC. *Current Cigarette Smoking Prevalence Among Working Adults–United States , 2004–2010*. Technical Report. Morbidity and Mortality Weekly Report (MMWR) (2016).
37. Miller, W. G. Estimating glomerular filtration rate. *Clin. Chem. Lab. Med.* **47**, 1017–1019. https://doi.org/10.1515/CCLM.2009.264 (2009).
38. CDC. *Percentage with CKD stage 3 or 4 who were aware of their disease by stage and age 1999–2012*. Technical Report (2015).
39. Whelton, P. K. *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *J. Am. Coll. Cardiol.*https://doi.org/10.1016/j.jacc.2017.11.006 *(2017).*
40. Feizi-Derakhshi, M.-R. & Ghaemi, M. Classifying different feature selection algorithms based on the search strategies. In *International Conference on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014* 17–21. https://doi.org/10.15242/IIE.E0114032 (2014).
41. Razmjoo, A., Xanthopoulos, P. & Zheng, Q. P. Online feature importance ranking based on sensitivity analysis. *Expert Syst. Appl.* **85**, 397–406. https://doi.org/10.1016/j.eswa.2017.05.016 (2017).
42. Uysal, A. K. & Gunal, S. Text classification using genetic algorithm oriented latent semantic features. *Expert Syst. Appl.* **41**, 5938–5947. https://doi.org/10.1016/j.eswa.2014.03.041 (2014).

43. Seret, A., Maldonado, S. & Baesens, B. Identifying next relevant variables for segmentation by using feature selection approaches. *Expert Syst. Appl.* **42**, 6255–6266. https://doi.org/10.1016/j.eswa.2015.01.070 (2015).
44. Jiang, S., Chin, K. S., Wang, L., Qu, G. & Tsui, K. L. Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. *Expert Syst. Appl.* **82**, 216–230. https://doi.org/10.1016/j.eswa.2017.04.017 (2017).
45. Wu, Y.-L., Tang, C.-Y., Hor, M.-K. & Wu, P.-F. Feature selection using genetic algorithm and cluster validation. *Expert Syst. Appl.* **38**, 2727–2732. https://doi.org/10.1016/j.eswa.2010.08.062 (2011).
46. Huang, G. .-B. *et al.* Extreme learning machine: theory and applications. *Neurocomputing* **70**, 489–501. https://doi.org/10.1016/j.neucom.2005.12.126 (2006).
47. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387. https://doi.org/10.1098/rsif.2017.0387 (2018).
48. Jain, S., Shukla, S. & Wadhvani, R. Dynamic selection of normalization techniques using data complexity measures. *Expert Syst. Appl.* **106**, 252–262. https://doi.org/10.1016/j.eswa.2018.04.008 (2018).
49. Singh Gill, H., Singh Khehra, B., Singh, A. & Kaur, L. Teaching–learning-based optimization algorithm to minimize cross entropy for selecting multilevel threshold values. *Egypt. Inform. J.*https://doi.org/10.1016/j.eij.2018.03.006 *(2018).*
50. Bendersky, E. *The Softmax Function and Its Derivative* 1–9. https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/ (2018). Accessed November 2018.
51. He, K., Zhang, X., Ren, S. & Sun, J. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*. Technical Report https://doi.org/10.1109/ICCV.2015.123 (2015). arXiv:1502.01852.
52. Takase, T., Oyama, S. & Kurihara, M. Effective neural network training with adaptive learning rate based on training loss. *Neural Netw.* **101**, 68–78. https://doi.org/10.1016/j.neunet.2018.01.016 (2018).
53. Subramanian, J. & Simon, R. Overfitting in prediction models—is it a problem only in high dimensions?. *Contemp. Clin. Trials* **36**, 636–641. https://doi.org/10.1016/j.cct.2013.06.011 (2013).
54. Shotton, J., Sharp, T. & Kohli, P. Decision jungles: compact and rich models for classification. *Adv. Neural Inf. Process. Syst.* **26**, 234–242 (2013).
55. Asl, A. & Overton, M. L. Analysis of Limited-Memory BFGS on a Class of Nonsmooth Convex Functions. arXiv:1810.00292 (2018).
56. Son, Y. J., Kim, H. G., Kim, E. H., Choi, S. & Lee, S. K. Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthc. Inform. Res.* **16**, 253–259. https://doi.org/10.4258/hir.2010.16.4.253 (2010).
57. Friedman, J. . H. Greedy function approximation : a gradient boosting machine 1 function estimation 2 numerical optimization in function space. *North* **1**, 1–10. https://doi.org/10.2307/2699986 (1999).
58. Lazic, N., Bishop, C. & Winn, J. Structural Expectation Propagation (SEP): Bayesian structure learning for networks with latent variables. In *16th International Conference on Artificial Intelligence and Statistics* Vol. 31, 379–387 (2013).
59. Barua, S., Islam, M. M. & Murase, K. A novel synthetic minority oversampling technique for imbalanced data set learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7063 LNCS, 735–744. https://doi.org/10.1007/978-3-642-24958-7_85 (2011).
60. Giacomini, R. & White, H. *Tests of conditional predictive ability*. https://doi.org/10.1111/j.1468-0262.2006.00718.x (2006).
61. Elvira, C., Ochoa, A., Gonzalvez, J. C. & Mochon, F. Machine-learning-based no show prediction in outpatient visits. *Int. J. Interact. Multimed. Artif. Intell.* https://doi.org/10.9781/ijimai.2017.03.004 (2018).

Acknowledgements

This document presents independent study funded by Sanitas USA. The points of view expressed are those of the authors and not necessarily those of the NIHR, the NHS, the NHANES or the department of health. We thank Ivan Javier Murcia Muõz, Healthcare Services Director at Sanitas USA and Martha Duarte, Epidemiologist at Sanitas USA. Their skills and competence remarkably support the study.

Author contributions

R.G.C.: worked on the global and methodological review of the paper. F.L.M.: worked on the implementation, research and eld tests. E.R.N.V.: worked on methodological part. V.G.D.: worked in the development and field tests.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.G.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

algorithms

MDPI

Article

A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management

Fernando López-Martínez [1,*,†,‡], Edward Rolando Núñez-Valdez [1,‡], Vicente García-Díaz [1,‡] and Zoran Bursac [2,‡]

[1] Department of Computer Science, Oviedo University, 33003 Oviedo, Spain; nunezedward@uniovi.es (E.R.N.-V.); garciavicente@uniovi.es (V.G.-D.)
[2] Department of Biostatistics, Florida International University, Miami, FL 33199, USA; zbursac@fiu.edu
* Correspondence: uo259897@uniovi.es; Tel.: +1-551-587-0112
† Current address: Sanitas Medical Center, 8400 NW 33rd St, Doral, FL 33122, USA.
‡ These authors contributed equally to this work.

Received: 28 February 2020; Accepted: 21 April 2020; Published: 23 April 2020

Abstract: Big data and artificial intelligence are currently two of the most important and trending pieces for innovation and predictive analytics in healthcare, leading the digital healthcare transformation. Keralty organization is already working on developing an intelligent big data analytic platform based on machine learning and data integration principles. We discuss how this platform is the new pillar for the organization to improve population health management, value-based care, and new upcoming challenges in healthcare. The benefits of using this new data platform for community and population health include better healthcare outcomes, improvement of clinical operations, reducing costs of care, and generation of accurate medical information. Several machine learning algorithms implemented by the authors can use the large standardized datasets integrated into the platform to improve the effectiveness of public health interventions, improving diagnosis, and clinical decision support. The data integrated into the platform come from Electronic Health Records (EHR), Hospital Information Systems (HIS), Radiology Information Systems (RIS), and Laboratory Information Systems (LIS), as well as data generated by public health platforms, mobile data, social media, and clinical web portals. This massive volume of data is integrated using big data techniques for storage, retrieval, processing, and transformation. This paper presents the design of a digital health platform in a healthcare organization in Colombia to integrate operational, clinical, and business data repositories with advanced analytics to improve the decision-making process for population health management.

Keywords: decision support systems; population health management; big data; machine learning; deep learning; personalized patient care

1. Introduction

Colombia's health system is formed by the public sector and the private sector. The general social security system has two plans, contributory and subsidized. The contributory regime covers salaried workers, pensioners, and independent workers, with the subsidized plan covering anyone who cannot pay. Enrollment coverage increased from 96.6% in 2014 to 97.6% in 2015 [1].

The National Health Authority's primary purpose in Colombia is to improve the quality of healthcare and strengthening supervision, surveillance, and control of the health system. The 2015 Statutory Health Law No. 1751 places the responsibility for guaranteeing the right to health with the

health system and recognizes health as a fundamental social right and makes it the state's responsibility to pursue an approach in health promotion and disease prevention [2].

The health sector in Colombia supports all initiatives for implementing new technologies to prevent cardiovascular diseases, disabilities, and high-cost hospitalization cases [3]. There is a remarkable need to improve the prediction of the risk of conditions for the population through the integration and unification of massive volumes of data and the implementation of effective advance analytic solutions to improve the decision-making process and population health management in Colombia's population [4].

Keralty organization is formed by a group of insurance and health services companies with a global presence, which together develops an integral health model, whose purpose is to produce health and well-being to people throughout their lives. The organization is committed to keeping its users healthy and autonomous, focusing on prevention, identification, and management of health risks, control, and care of disease and dependency [5]. The organization is a leader in Colombia by providing integrated health services and is recognized for their human, scientific, technical, and ethical approach [6].

This paper presents how we can obtain value from a large volume of heterogeneous data generated by different data sources in healthcare, and the architecture implemented. The development of proper advanced data analytics methods such as machine learning and big data analytics to perform meaningful real-time analysis on the data to predict clinical complications before it happens and to support the decision-making process are challenging but much needed to handle the complexity of the data-driven problems we are currently facing.

1.1. Related Work

Several initiatives in Europe, Asia, and North America aim to develop healthcare digital platforms with collaborative access tools to allow the exchange and sharing of information and knowledge wherever and whenever needed throughout the attention process. This type of frameworks and architectures will allow maximum quality and efficiency for patient's care, and to provide appropriate attention to the patient's condition and risks.

Castilla and Leon, for example, implemented a digitalization of health services as a tool to increase the efficiency of the services and increase the security in the attention to patient [7]. A healthcare cyber-physical system assisted by cloud and big data is being developed in the department of computer science at Pace University in New York [8]. This system consists of a data integration layer, a data management layer, and a data-analytics service layer to improve the functioning of the healthcare system. In France, a group of researchers implemented a wearable knowledge as a service platform to cleverly manage heterogeneous data coming from wearable devices to assist the physicians in supervising the patient health [9]. Another interesting work was presented at the International Conference on Computational Intelligence and Data Science (ICCIDS 2018). The authors proposed a hybrid four-layer healthcare model to improve disease diagnostic [10]. In India, a centralized architecture for an end to end integration of healthcare systems deployed in the cloud environment was developed using fog computing [11].

Medical organizations are investing more and more in developing a healthcare platform that integrates data, applications, business processes, and user interfaces to gain knowledge and useful insights for clinical decisions, drug recommendation systems, and better disease diagnoses. Some other examples of big data applications in healthcare can be found in healthcare monitoring, where data captured from wearable devices can assist providers in managing symptoms of patients online and adjust their prescriptions [12]. An analytical platform called "MedAware" has been developed to detect errors in medical prescriptions and clinical errors, reducing the hospital admission and readmission in real-time [13]. In the healthcare prediction field, a healthcare system called "Gemini (Generalizable Medical Information analysis and Integration system)" was developed to collect, process, and analyze large volumes of clinical data and apply machine learning algorithms for performing predictive

analytics [14]. Other platforms have been implemented for genomics data analytics to generate predictions based on DNA molecular changes and mutations [15]. Another type of healthcare platform is related to the healthcare knowledge system, defined as the combination of clinical data and physician expertise to support clinical decision-making and diagnosis [16].

1.2. Why Big Data and Machine Learning?

Big data and machine learning are redefining healthcare goals for the future. Healthcare data are impacting the way disease research is performed, and the level of complexity in population health management is increasing as the traditional fee for service approach is transformed into the value-based care model [17,18].

Population health management is basically the aggregation of patient health data from multiple data sources, and the analysis and transformation into actionable insights to generate informed decisions to improve clinical and financial outcomes [19].

Big data technologies will allow us to bring large volumes of structured and unstructured data from disparate data sources into a data repository to be examined and analyzed. Machine learning models will assist in discovering insights from complex datasets with capabilities such as finding unseen patterns, making new predictions, and analyzing trends on health data. Machine learning is being used in a variety of clinical domains with the analysis of hundreds of clinical parameters resulting in effective and efficient models to improve the outcomes and quality of medical care models [20].

The implementation of this platform shows the enormous potential in using big data to individualize medical treatment, the opportunity for improving the lives of the patients, delivering better medical care, and reduced waste at an operational level [21]. Other chances for big data in healthcare for Keralty organization are:

- A physician would know before prescribing whether the patient is at high-risk to become dependent and different treatment plans can be selected based on this information.
- Psychosocial and clinical medical data could inform about the development of a chronic illness that can be properly diagnosed.
- The organization can use big data to understand how they are performing, the opportunities to improve clinical care, and their capacity to redesign care delivery to their patients.
- Using the platform's analytics component to improve the quality of care and patient experience at the lowest possible cost is core to the organization.
- Capturing streaming data and wearable data can provide to healthcare providers real-time insights about a patient's health that will allow them to improve their decision-making process for treatment and medication.
- Big data analysis can help the organization to deliver information that is evidence-based and can improve the efficiency, understanding, and implementation of the best practices associated with any disease.

In addition to the big data technologies used to build the platform, another essential component is the advanced analytic module of the platform. This module contains several machine learning algorithms to support clinical diagnosis. However, the organization should feel confident in these models and how they can be applied to specific use cases. These first models will alert providers to changes in high-risk conditions such as sepsis and hypertensive patients.

The main objective of this paper is to present the developed platform and its components to allow Keralty organization to derive better and more actionable insights from their data, i.e., to derive meaningful information from all these data in a way that allows them to improve care and lower costs needed for value-based reimbursement and business objectives while providing the highest quality care for population health management [22]. The goal is to be aligned with the triple aim framework developed by the Institute for Healthcare Improvement that describes an approach to optimizing healthcare system performance.The implementation of this platform intends to resolve

several problems in health services to assist patients and their families in managing their health by providing better access to healthcare services [23].

2. Proposed Digital Health Platform

Keralty organization currently have several information systems such as Health Information Systems (HIS), Lab Information Systems (LIS), Radiology Information Systems (RIS), Enterprise Resource Planning (ERP), and Customer Relationship Management (CRM), among others, in their ambulatory care centers, hospitals, and home care, which support their integrated health model. The information from these systems was not consolidated on a single platform, and its access and availability generated an operative load, which obstructs all health management processes and the support of clinical decisions for physicians. Consequently, we proposed the design and implementation of a healthcare, clinical, and business data repository with advanced analytic capabilities to consume machine learning prediction models to improve the decision-making process and population health management at the organization. The digital health platform conceptual framework is shown in Figure 1.

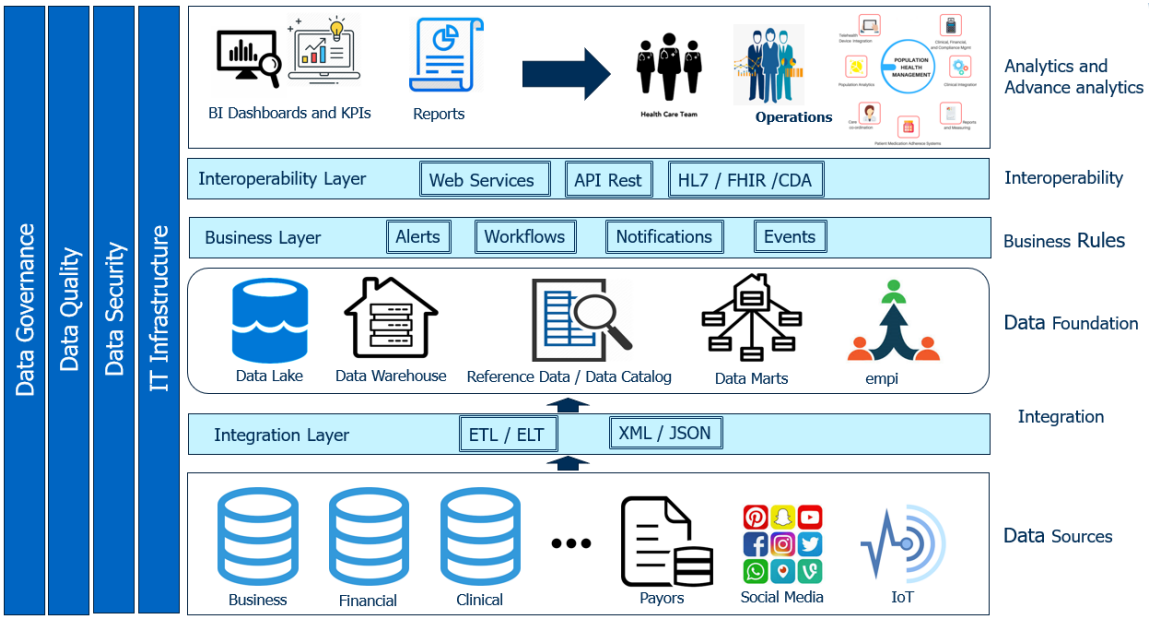

Figure 1. Conceptual Framework—Keralty Health Portal.

The implementation of the platform was an ambitious project that required integrating health information from disparate sources, building numerous technological and functional components, and the definition of IT management processes robust enough to support interoperability with other systems. The digital health information platform included patient-related data, Electronic Health Records (EHR), diagnostic reports, prescriptions, medical images, pharmacy records, research data, operational data, financial data, and human resources data.

This project was innovative and pioneered the designing and building of a comprehensive health digital platform for a healthcare organization in Colombia, with the patient being at the center of it and all of its information aggregated and summarized based on the standardized enterprise data repository. This information can be accessed quickly and intuitively when and where it is needed, hiding all technical complexity and providing longitudinal process management tools, as well as tools for decision support for professionals. The difference of this platform with other implementations was the development of a medical portal with a patient 360 view that uses data from the enterprise data repository to generate real-time early warning scores, patient surveillance, open API for hospitals integration, prediction of health risk patterns, high-risk markers, co-morbidity

detection to predict critical diseases, early diagnosis of diseases, treatment comparison with medical guidelines, and measurement of efficiency of specific drugs to provide the best quality of care.

The Digital Healthcare platform architecture can ingest data from over 50 different source systems at the granular level, including claims, clinical, financial, administrative, wearables, genomics, and socioeconomic data. Few platforms today can integrate that many heterogeneous data sources successfully. The platform can consume machine learning models on-demand without the need for further development. The data logic models are on top of the raw data and can be accessed, reused, and updated through open APIs without the need for clinical and business logic changes. The platform was able to integrate successfully structured and unstructured data. It is commonly seen that this type of platforms in the market is built to either integrate structured data or unstructured but few cases successfully integrate both. Open microservices APIs were created for operations such as authorization, identity management, interoperability, and data pipeline management. These microservices enable the development of third-party applications to interoperate with the platform.

2.1. Platform Architecture

The initial approach was to build a big data processing pipeline with a Microsoft Azure lambda architecture to support real-time and batch analytics. This approach is shown in Figure 2. This architecture has different mechanisms to consume data depending on the source and timing needed to generate insights. In addition, with this approach, we can have professionals with different skills working in parallel to build the platform.

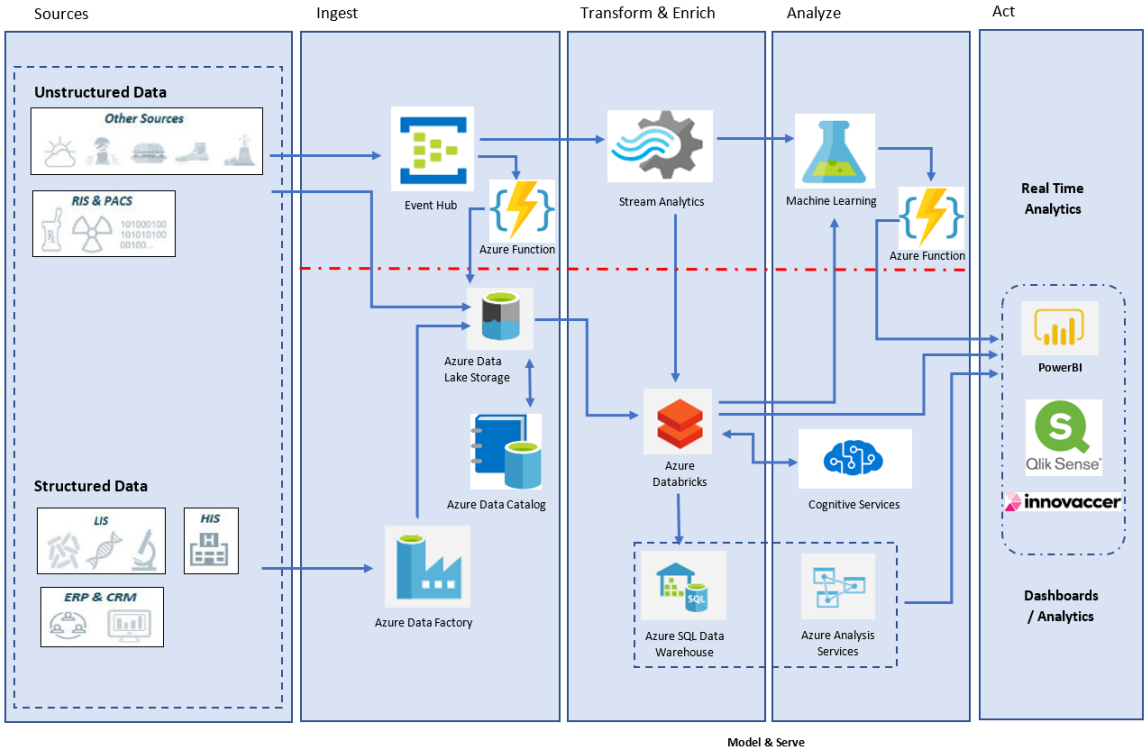

Figure 2. Azure Big Data and Machine Learning Lambda Architecture.

The architecture contains a batch layer, a real-time layer, and a serving layer. The batch layer is in charge of persistent storage and is able to scale horizontally. The real-time layer process streaming data and performs dynamic computation. The serving layer query data on the repositories and consume the prediction models.

From the infrastructure point of view, the platform offers the flexibility of being implemented in a hybrid environment, namely the cloud and the local data processing center, through the use of virtualization techniques, containers, and load balancing systems. The design of the infrastructure was

prepared to provide a flexible set of resources that can be used on-demand and based on the specific workload requirements. The infrastructure deployment relied heavily on automation to provide fluid operations.

2.2. Data Repository

An enterprise-wide staging repository for the big data analytics platform was considered. The data lake allows capturing data of any volume, type, and ingestion speed in one single place for storing heterogeneous data. This staging area included capabilities such as security, scalability, reliability, and availability. The data can be passed, processed directly from the staging area, or can be ingested to an enterprise data warehouse for historical load, preparation, and serve for BI and machine learning needs. This data warehouse repository has a scale-out architecture and massively parallel processing (MPP) engine.

Data models were developed to cover clinical, social, and healthcare program domains. Each model performs validations and processing on the data received, decoupling the processing and administration of the data from the source. These data models can also be extended to store additional attributes specific to the implementation, allowing these models to subscribe to certain types of messages, using the mapping and filtering options provided by the data processing pipelines. Once these subscriptions are created, the model will be loaded with all relevant messages to those who are subscribed and stored in the data lake.

For data storage, the data are loaded into a data warehouse with a daily refresh. This healthcare data repository contains a highly normalized data model for fast and efficient querying and analysis. This repository is read-only.

2.3. Integration and Interoperability

The platform provides a mechanism to integrate data from heterogeneous sources, define workflows to ingest data from different data stores, and transform and process data to data stores to be consumed by BI applications. A cloud-based data integration service is used to create these data-driven workflows and orchestrate all automation, transformation, and data movement in the platform. The main tasks this integration service should perform are: creation and scheduling of data pipelines to ingest data from different data sources, processing and transformation of the data, and store data in data stores such as data lakes or data warehouses.

Azure Data Factory automates and orchestrates the entire data integration process from end to end in the platform. We built the ETL (extract, transform, and load) pipelines with this Azure component. The data are extracted from the source locations, transformed from its source format to the target Azure data lake's schema, and loaded into Azure data lake and the data warehouse, where they can be used for analytics and reporting. Azure Data Factory defines control flows that execute various tasks in the transform and load process.

We used the mechanism called mapping data flows, combining control flows and data flows to build the data transformations with an easy-to-use visual user interface. These data flows are then executed as activities within Azure Data Factory pipelines. Data Factory is certified by HIPAA (Health Insurance Portability and Accountability Act), which protects the data while they are in use with Azure. In the data flow, we created transformation streams where we define the source data and create the graph with the transformations, schema operations such as derived column, aggregate, surrogate keys and selects, and the output settings.

2.4. Data Security and Privacy Model

In terms of security, the platform guarantees authentication, access control, and encryption capabilities. The security mechanisms of the platform can provide protection, alert monitoring, and support the OAuth 2.0 protocol for authentication with REST interfaces. ACLs are enabled on folders, subfolders, and files. The platform also provides encryption mechanisms to protect the data.

All these capabilities are accompanied by the implementation of enterprise security policies and regulatory compliance requirements.

2.5. Stream Analytics

The platform can handle mission-critical real-time data and offer end to end streaming pipelines with continuous integration and continuous delivery (CI-CD) services. Other capabilities such as in-memory processing, data encryption, and support of international security standards including HIPAA (Health Insurance Portability and Accountability Act), HITRUST (Health Information Trust Alliance), and GDPR (General Data Protection Regulation).

2.6. Advanced Analytics

The analytic data component consists of two areas: The first area is the BI models we develop for tactical, operational, and strategic decisions. The second area comprehends several prediction models that need to be developed. Currently, there are two prediction models developed by the authors of this paper to support population health management, specifically the diagnosis of sepsis and hypertension prediction [24,25]. These insights assist providers in the detection and tracking of chronic diseases. The machine learning component is used to build, test, consume, and deploy predictive analytic models on-demand and as requested for the organization. The platform provides self-service dashboards and visualizations that use data from the repositories to drive the decision-making process. The machine learning application layer is one of the essential layers of this platform.

Once the data are integrated, aggregated, and normalized in the system, the platform offers a tool to provide knowledge management through the business intelligence interface providing data analysis, design, and training of machine learning models, as well as development and management of results-based care indicators or population health management. The platform provides a tool where clinicians, researchers, and scientists can mine the data and get valuable information.

Machine learning models can be trained and customized in preconfigured data domains, allowing the storage of the results for future use. Data researchers and scientists can develop advanced tools to obtain information and value of the data stored in the solution, taking advantage of the model design, training, and validation component. We briefly present the predictive models implemented in the platforms.

- **Machine Learning Classification for a Hypertensive Population:** This prediction model evaluates the association between gender, race, BMI (Body Mass Index), age, smoking, kidney disease, and diabetes using logistic regression. Data were collected from NHANES datasets from 2007 to 2016 to train and test the model, a dataset of 19,709 samples with (83%) non-hypertensive individuals and (17%) hypertensive individuals. The results show a sensitivity of 77%, a specificity of 68%, precision on the positive predicted value of 32% in the test sample, and a calculated AUC of 0.73 (95% CI [0.70–0.76]). The model used to estimate the probability that a person will belong to the hypertensive or non-hypertensive class is:

$$p = \frac{e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}{1+e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}$$

We used the logistic regression classification model in this experiment to evaluate the importance of the risk factor variables and their relationship with the prevalence of hypertension among a nationally representative sample of adults ≥20 years in the United States (n = 19,759). The distribution of the samples by hypertensive patients, gender, and race is shown in Table 1.

Table 1. Number of samples by hypertensive class, gender, and race.

Hypertension, Adults 20 and over—2007–2016			
Class	**Gender**	**Race**	**n**
Non Hypertensive	**Female**	**Mexican American**	1269
		Non-Hispanic Black	1674
		Non-Hispanic White	**3674**
		Other Hispanic	951
		Other Race—Including Multi-Racial	864
	Male	**Mexican American**	1255
		Non-Hispanic Black	1599
		Non-Hispanic White	**3714**
		Other Hispanic	774
		Other Race—Including Multi-Racial	843
Hypertensive	**Female**	**Mexican American**	205
		Non-Hispanic Black	420
		Non-Hispanic White	**662**
		Other Hispanic	149
		Other Race—Including Multi-Racial	114
	Male	**Mexican American**	214
		Non-Hispanic Black	478
		Non-Hispanic White	**670**
		Other Hispanic	138
		Other Race—Including Multi-Racial	132
		Total	**19,799**

We computed chi-square test between each independent variable and the dependent variable to indicate the strength of evidence that there is some association between the variables. Chi-square was selected due to the categorical form of the data used in the model, and it is considered one of the best methods to estimate the dependency between the class and the features when the feature can take a fixed number of possible values that belong to a group or nominal category.

Table 2 shows the p-value for each variable; the null hypothesis is reject for any $p \leq 0.05$, while the null hypothesis is not rejected when $p > 0.05$. p-values for the variables GENDER, BMIRANGE_1, BMIRANGE_3, and KIDNEY_2 are not statistically significant at 0.05 alpha level; the clinical importance of these variables in the model for interpretation allows us to include them. We ran the model with and without the variables, and there were no significant changes in the accuracy score, positive predicted value rate, and true positive rate.

The training dataset was derived from a random sampling of 70% (13,831) of the extracted study population and the test sampling the remaining 30% (5928) to evaluate the model on the ground-truth that was never used for training. We ran the logistic regression model on the entire dataset to verify the accuracy score of the model.

Table 2. Chi2 test and *p*-value for the independent variables.

Chi-Squared between Each Indicator Variable and the Baseline for the Model				
Feature	**Description**	**Dummy**	***p*-Value**	**Score**
GENDER	**Male**	**GENDER_1**	0.1416446	2.160001
	Female	**GENDER_2**	0.1450268	2.123795
AGERANGE	**20–30**	**AGERANGE_1**	0.0000001	560.890568
	31–40	**AGERANGE_2**	0.0000001	299.675698
	41–50	**AGERANGE_3**	0.0000001	98.221463
	51–60	**AGERANGE_4**	0.0000035	21.520345
	61–70	**AGERANGE_5**	0.0000001	342.879412
	71–80	**AGERANGE_6**	0.0000001	1037.137074
RACE	**Mexican American**	**RACE_1**	0.0067797	7.330429
	Other Hispanic	**RACE_2**	0.0275756	4.854409
	Non-Hispanic White	**RACE_3**	0.0455912	3.996636
	Non-Hispanic Black	**RACE_4**	0.0000001	91.264812
	Other Race	**RACE_5**	0.0000278	17.562718
BMIRANGE	**Underweight = <18.5**	**BMIRANGE_1**	0.6730361	0.178071
	Normal weight = 18.5–24.9	**BMIRANGE_2**	0.000033	17.234712
	Overweight = 25–29.9	**BMIRANGE_3**	0.9174572	0.010741
	Obesity = BMI of 30 or greater	**BMIRANGE_4**	0.0006362	11.666854
KIDNEY	**Yes**	**KIDNEY_1**	0.0000001	58.963059
	No	**KIDNEY_2**	0.1872889	1.738816
SMOKE	**Yes**	**SMOKE_1**	0.0021759	9.394891
	No	**SMOKE_2**	0.0053461	7.758468
DIABETES	**Yes**	**DIABETES_1**	0.0000001	217.214128
	No	**DIABETES_2**	0.0000001	39.351672
	Borderline	**DIABETES_3**	0.0000051	20.798905

The Logistic Regression model uses the logit function to express the relationship of the risk factors as:

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

The probability of success can be expressed as:

$$p = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i)}}$$

where p is the predicted probability of having hypertension, X_i are the risk factors or independent variables, and β_i are the coefficients that are estimated by using the method of maximum likelihood and allow us to calculate the odds that, for every unit increase in X_i, the odds of having hypertension changes by e^{β}.

- **A neural network approach to predict early neonatal sepsis:** We developed a non-invasive neural network classification model for early neonatal sepsis detection. The data used in this study are from Crecer's Hospital center in Cartagena-Colombia. A dataset of 555 neonates with (66%) of negative cases and (34%) of positive cases was used to train and test the model. The study results show a sensitivity of 80.32%, a specificity of 90.4%, precision on the positive predicted value of 83.1% in the test, sample and a calculated area under the curve of 0.925 (95% Confidence Interval [91.4–93.06]). The neural network architecture can be seen in Figure 3.

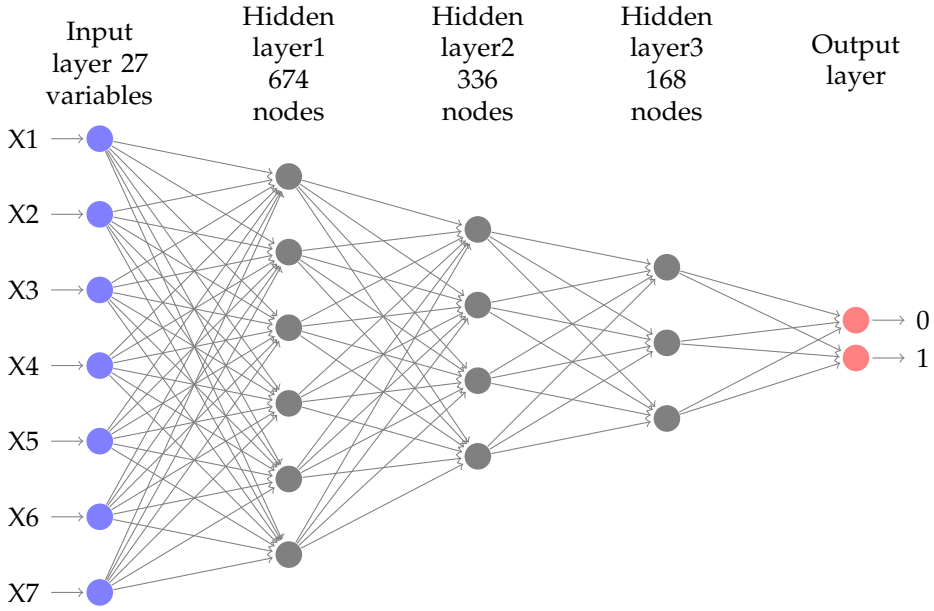

Figure 3. Multilayer Perceptron Architecture.

Table 3 shows the parameters of the architecture. Labels X1–X7 are informative only, and the input size is 27 variables.

Table 3. Model architecture parameters.

Model Architecture Parameters	
Parameter	**Value**
Input Dimension	27
Num Output classes	2
Num Hidden Layers	3
Hidden Layer1 Dimension	674
Activation Func Layer1	Relu
Hidden Layer2 Dimension	336
Activation Func Layer2	Relu
Hidden Layer3 Dimension	168
Activation Func Layer3	Relu
Minibatch size	8
Num samples to train	388
Num minibatches to train	48
Loss Function	cross entropy with softmax
Eval Error	Classification error
Learner for parameters	momentum sgd
Eval Metrics	Confusion Matrix, AUC

The model used an anonymous dataset from a private medical institution in Cartagena, Colombia, from 2016 to 2017. Demographic, laboratory data, blood pressure, and body measures data were part of the dataset. This dataset includes cases of live newborns of ages inferior to 72 h with a diagnosis of early neonatal sepsis by clinical criteria and laboratory blood cultures. Control cases were part of the dataset including all newborns healthy by clinical diagnosis and who returned healthy for a follow up at 72 h.

This retrospective study includes 186 cases and 368 controls based on a case-control relationship of 1:2 with a 95% trust factor and power of 80%. Bivariate analysis and logistic regression were performed to detect the variables associated with early sepsis, and the statistical significance was considered at the alpha level of 0.05.

This model considered nine sociodemographic, fourteen obstetric, nine neonatal, and four maternal infectious related pathology variables. Table 4 shows the quantitative sociodemographic variables, Table 5 shows the qualitative sociodemographic variables, Table 6 shows the quantitative neonatal variables, Table 7 shows the qualitative neonatal variables, Table 8 shows the quantitative obstetric variables, Table 9 shows the qualitative obstetric variables, and Table 10 shows the qualitative maternal infections of the cases and controls.

A bivariate chi-square test with correction was performed to the qualitative variables to find a statistical association between the independent variable and the possibility to develop early neonatal sepsis. For continuous variables, the Mann–Whitney U test was performed. From this statistical analysis, it is essential to show that we did not find significant statistical evidence for the variables age, start of marital status at younger than 18 years old, gender, APGAR (Appearance, Pulse, Grimace, Activity, and Respiration) value less than 7 after 1 and 5 min, the number of pregnancies, and the type of birth. Prenatal control is not associated with the case of sepsis; however, assisting to five prenatal controls are associated with the protection to avoid the appearance of early neonatal sepsis. There was no evidence with the variables IUGR (Intrauterine Growth Restriction) background and multiple pregnancies. Twenty-seven (27) variables were selected as input variables for our artificial neural network architecture.

Table 4. Quantitative sociodemographic variables in cases (186) and controls (369).

Quantitative Socio Demographic Variable	Cases				Controls				
	Mean	Median	SD	RIQ	Mean	Median	SD	RIQ	*p*-Value
Age	23.93	23.5	4.99	20–26	24.22	23	6.19	19–28	0.793
Onset of sexual activity	16.06	16	0.945	15–17	15.6	16	0.971	15–16	0.0001

Table 5. Qualitative sociodemographic variables in cases (186) and controls (369).

Qualitative Socio Demographic Variable	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Teen Mother	Yes	15	8.1	69	18.7	10.88	0.001
	No	171	91.9	300	81.3		
Health Regimen	Government	183	98.4	349	94.6	4.51	0.041
	Commercial	3	1.6	20	5.4		
Origin	Rural	42	22.6	5	1.4	71.87	0.00001
	Urban	144	77.4	364	98.6		
Marital Status	Married or in common law married	128	68.8	101	27.4	87.64	0.00001
	Single, divorced or widow	58	31.2	268	72.6		
Level of education	Elementary School	86	46.2	80	21.7	35.57	0.00001
	High School	100	53.8	289	78.3		
Start of Marital status life younger than 18 yo	Yes	178	95.7	357	96.7	0.39	0.531
	No	8	4.3	12	3.3		
Start of Marital status life younger than 16 yo	Yes	47	25.3	147	39.8	11.54	0.001
	No	139	74.7	222	60.2		

Table 6. Quantitative Neonatal variables in cases (186) and controls (369).

Quantitative Neonatal Variable	Cases				Controls				
	Mean	Median	SD	RIQ	Mean	Median	SD	RIQ	*p*-Value
New born weight in grams	2639.9	2768.5	546.5	2500–3020	3202.4	3224	412.1	2950–3500	0.0001
APGAR after 1 min of birth	7.73	8.0	0.611	8.0	8.09	8.0	0.598	8.0	0.0001

Table 7. Qualitative Neonatal variables in cases (186) and controls (369).

Qualitative Neonatal Variable	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Premature	Yes	100	53.8	25	6.8	156.4	0.0001
	No	86	46.2	344	93.2		
Gender	Male	109	58.6	202	54.7	0.748	1.672
	Female	77	41.4	167	45.3		
Less than 1500 grams	Yes	11	5.9	2	0.5	15.6	0.00001
	No	175	94.1	367	99.5		
Less than 2500 grams	Yes	44	23.7	9	2.4	64.44	0.00001
	No	142	76.3	360	97.6		
APGAR less than 7 after 1 min of birth	Yes	2	1.1	3	0.8	0.095	0.999
	No	184	98.9	366	99.2		
APGAR less than 7 after 5 min	Yes	4	2.2	9	2.4	0.045	0.999
	No	182	97.8	360	97.6		
Respiratory distress	Yes	89	47.8	27	7.3	122.8	0.0001
	No	97	52.2	342	92.7		

Table 8. Quantitative Obstetric variables in cases (186) and controls (369).

Quantitative Obstetric Variable	Cases				Controls				
	Mean	Median	SD	RIQ	Mean	Median	SD	RIQ	*p*-Value
Gestational age at the time of birth	35.6	36.0	3.47	34–39	38.4	39.0	1.62	38–39	0.0001
Number of prenatal controls	4.08	5.0	1.83	3.75–5.0	4.32	5.0	1.83	4–5.0	0.002
Number of pregnacies	1.77	1.0	1.15	1.0–2.0	1.6	1.0	1.15	1–2.0	0.076
Number of births	1.04	1.0	1.03	0–1	0.7	1.0	1.03	0–1	0.0001
Numbers of C-sections	0.65	1.0	0.68	0–1	0.76	1.0	0.68	0–1	0.029

Table 9. Qualitative Obstetric variables in cases (186) and controls (369).

Qualitative Obstetric Variable	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Type of birth	Vaginal	98	52.7	162	43.9	3.833	0.05
	C-Section	88	47.3	207	56.1		
IUGR Background	Yes	5	2.7	13	3.5	0.275	0.6
	No	181	97.3	356	96.5		
Assistance for prenatal control	Yes	165	88.7	318	86.2	0.702	0.402
	No	21	11.3	51	13.8		
Assistance for at least 4 prenatal control	Yes	140	75.3	301	81.6	3.01	0.083
	No	46	24.7	68	18.4		
Assistance for at least 5 prenatal control	Yes	105	56.5	254	68.8	8.301	0.004
	No	81	43.5	115	31.2		
Premature rupture of membrane with more than 18 hours	Yes	95	51.1	17	4.6	165.7	0.00001
	No	91	48.9	352	95.4		
Chorioamnionitis	Yes	23	12.4	3	0.8	36.96	0.00001
	No	163	87.6	366	99.2		
Premature membrane rupture with more than 6 hours	Yes	161	86.6	194	52.6	61.96	0.0001
	No	25	13.4	175	47.4		
Multiple Pregnacies	Yes	2	1.1	10	2.7	0.39	0.353
	No	184	98.9	359	97.3		

Table 10. Qualitative maternal infections variables in cases (186) and controls (369).

Qualitative Maternal Infections Variables	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Maternal Fever	Yes	67	36.0	40	10.8	50.38	0.0001
	No	119	64.0	329	89.2		
Yeast Infections	Yes	31	16.7	15	4.1	25.83	0.0001
	No	155	83.3	354	95.9		
Sexualy transmitted disease history	Yes	27	14.5	7	1.9	34.24	0.0001
	No	159	85.5	362	98.1		
Urinary Tract Infections	Yes	11	5.9	9	2.4	4.29	0.0381
	No	175	94.1	360	97.6		

In terms of computational timing, It is difficult to evaluate the complexity and timing of a machine learning algorithm. However, based on the algorithmic complexity, we can measure the time performance in terms of its training time complexity using big O notation because the classification time of the models can vary depending on the stress in the computational performance and power. In terms of timing, the classification prediction with the trained models is less than 1 s. The time complexity of the logistic regression could be expressed as $O((f+1)csE)$, where f is the number of features (+1 because of bias), c is the number of possible outputs, s is the number of samples, and E is the number of epochs to run. For the neural network approach, $O(pnl_1 + nl_1nl_2 + \dots)$, where p is the number of features and nl_i is the number of neurons at layer i in a neural network [26].

3. Actual Platform Benefits

The implementation of the platform became the digital healthcare ecosystem for the organization. The organization can populate workflow information systems with critical decision-making insights, accurate and reliable healthcare data that significantly increased the value of the healthcare outcome to patients and care providers. This platform delivers significant benefits to the organization, such as physicians having an intelligent application that can be configured to their preferences and optimized to their disciplines, patients receiving more personalized care, an improvement in healthcare workflow and patient care, and personalized care for physicians and patients.

We describe in the following subsections several use cases that effectively present the change and digital transformation of the organization with the implementation of the platform.

3.1. Reduce Total Cost of Care for Care Coordination

With a robust data analytic component, the organization was able to prioritize opportunities for improvement and to improve the way care is coordinated and delivered throughout its network of hospitals and medical facilities. The results include a considerable increase in financial results in just six months.

The organization uses the platform to generate timely, meaningful, and actionable data to drive change and improve the quality of care for patients. The organization uses the data for risk-stratification of the network's population, prioritization of the care coordination activities, and prevention activity's interventions. Risk stratification was completed for all patients, enabling care managers to identify individuals at various risk levels for unnecessary services and high-cost utilization, improving patient outcomes and experience. The analytical component also reduces unnecessary visits, facilitates access to specialty care and community-based services, and achieves healthcare outcomes. Other benefits include 3% increase in the detection of high-risk patients with primary care, 20% increase in the number of patients with ongoing care managed, and 10% percent reduction in emergency department utilization per member among care managed patients.

3.2. Self-Service Analytic

As described in this paper, the healthcare platform combines and standardizes data across different source systems to provide actionable insights in a single platform. The platform integrates data from different sources, such as claims data, cost data, financial data, clinical data, and other patient data. With self-service analytics, the organization increases the number of users accessing the analytic component, improving data visibility and providing actionable insights to improve patient outcomes.

3.3. Reduce Deaths from Sepsis

The organization improved sepsis mortality rates and improving care outcomes by using the advanced analytic component of the platform. Sepsis impacts almost 1.7 million adults in the U.S. and is responsible for nearly 270,000 annual deaths. One-third of all hospital deaths are patients with sepsis [27]. The machine learning prediction model used in the platform was developed by one of the authors of this paper, as described before. It is still too early to mention the results of the utilization of this feature. However, the goal of the organization is to reduce its sepsis mortality rate, the costs of the creation of its sepsis care transformation team, and the implementation of an evidence-based sepsis care practice.

3.4. Discussion and Limitations

The digital health platform helps Keralty organization with closing the gaps between multiple datasets, improving clinical benefits, improving patient's lives, supporting better decision-making to manage larger populations, and improving overall health outcomes. However, the need for algorithms with high accuracy in medical diagnosis is still a challenge that needs to be improved precisely and efficiently [28]. The increasing complexity of building end-to-end platforms to integrate disparate systems and to apply machine learning techniques in specific areas such as computer vision, natural language processing, reinforcement learning, and other generalized methods present many challenges when forming the interdisciplinary team needed and the set of technological components used for the implementation.

Some challenges should be considered in the design and implementation of machine learning projects for healthcare. One of the most critical challenges requires algorithms that can answer causal questions. These questions are beyond classical machine learning algorithms because they require a formal model of interventions [29]. To address this type of question from the analytical component of the platform, we need to learn from data differently and to gain knowledge in causal models to understand how machine learning algorithms need to be trained. Another challenge is to create reliable outcomes from heterogeneous data sources with the participation of SME (Subject Matter Experts) who understand the disease; the machine learning predictive accuracy and correct clinical interpretation depend on the criteria and context of the disease. Providers and machine learning engineers should work together on model interpretability and applicability. Machine learning implementation is not an easy task; the selection of predictive features and optimization of hyperparameters is another challenge that needs to be mastered to implement models that provide useful insights [30]. The success and meaningful use of these algorithms, and their integration into the platform depends on the accuracy of the models and their interpretability.

4. Results of Advanced Analytics

After training and testing the logistic regression model for predicting hypertension, we generated some evaluation metrics to evaluate the classifier. Table 11 shows the confusion matrix with the classification results, include the true positive value (730), true negative value (3407), false negative (216), and false positive value (1575). The classification report in Table 12 shows the calculated precision and sensitivity.

Table 11. Confusion matrix.

		Predicted	
		Non-Hypertensive	**Hypertensive**
True	**Non-Hypertensive**	3407	1575
	Hypertensive	216	730

Table 12. Classification report.

Classification Report				
	Precision	**Recall**	**f1-Score**	**Support**
Non-Hypertensive	0.94	0.68	0.79	4982
Hypertensive	0.32	0.77	0.45	946
avg/total	0.84	0.7	0.74	5928

The test sampling of 5928 contains 4982 (84%) non-hypertensive and 946 (16%) hypertensive patients. The model shows a sensitivity of 730/946 = 77% and a specificity of 3407/4982 = 68%. The precision of the model was 730/2305 = 32% and the negative predicted value 3407/3623 = 94%. The false negative rate of the model was 216/946 = 22%. The model was better at identifying individuals who will not develop hypertension than those who will develop hypertension.

For the neural network approach to predict early neonatal sepsis, Table 13 shows the confusion matrix with the classification results of actual class label vs. the predicted ones, including the true positive value (49), true negative value (95), false negative (12), and false positive value (10).

Table 13. Confusion matrix.

		Predicted	
		Non-Sepsis	**Sepsis**
True	**Non-Sepsis**	95	10
	Sepsis	12	49

The classification report in Table 14 shows the precision and sensitivity. The sensitivity of the model is moderately acceptable due to the imbalanced testing dataset, and there is still a high number of false negatives.

Table 14. Classification report.

Classification Report			
True Positive	**False Negative**	**Precision**	**Accuracy**
49	12	0.83	0.867
False Positive	**True Negative**	**Recall**	**f1-score**
10	95	0.803	0.817
Positive Label: 1		**Negative Label: 0**	

A sensitivity of 80.3% and a specificity of 90.4% show that the model might be useful for detecting positive cases, and the true negative rate shows that the model is also efficient at identifying negative cases. The high precision value of 83.1% and the AUC of 0.925 confirm the adequacy of the model as a preliminary screening tool. The percentage of positive cases shows that the model works better than random guessing and the conditional probability of negative test results is considerably low.

The accuracy of 86.74% shows that the model correctly identifies negative cases and positive cases based on the characteristics of the dataset and the small number of cases examined.

5. Comparison with Other Platforms

A review of several healthcare platforms shows that the architecture presented in this paper covers all the categories from integration, interoperability, security care, and advanced analytics. Generally, other implementations only focused on one specific area, as shown in Table 15 and taken from the International Conference on Computational Intelligence and Data Science (ICCIDS 2018) and a healthcare frameworks review proposed in the Journal of King Saud University [31].

Table 15. Comparison of healthcare big data platforms.

Author and Year	Patient Centric	Predictive Analysis	Real Time Monitoring	Improve Treatment	Interoperability	Workflow and Rules	Pop Health	Patient 360
Our Health Platform	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Raghupathi et al. (2014) [32]	Yes	No	Yes	Yes	Partial	No	Partial	No
Patel et al. (2016) [33]	Yes	Yes	Yes	Yes	Partial	Partial	Yes	No
Chawla et al. (2013) [34]	Yes	Yes	No	Yes	Partial	Partial	Yes	No
Abinaya et al. (2015) [35]	Partial	Yes	Partial	Yes	Yes	Partial	No	No
Balladini et al. (2015) [36]	Yes	No	Yes	Yes	Partial	Yes	No	No
Belle et al. (2015) [37]	Partial	No	Yes	Yes	Partial	Yes	No	No
Mezghani et al. (2015)	Partial	No	Yes	Yes	Yes	Partial	No	No
Chen et al. (2017) [38]	Yes	Partial	Yes	Yes	Yes	Partial	Yes	No

We designed and implemented a healthcare platform using big data technologies with actionable insights to augment human decision-making at the organization impacting the population's health, public health, and to capture social determinants of health. This platform comprehends all the features we use in the comparison. Raghupathi et al. reported a conceptual architecture to present big data analytic outlines in healthcare with no predictive analytic capabilities and no patient 360 view. Patel et al. designed a big data architecture platform to improve data aggregation in the healthcare industry and to provide a reduction in healthcare cost, predicting analytic, preventive care, and drug discovery capabilities but without patient 360 view capabilities. Chawla et al. presented a patient-centric healthcare framework—Collaborative Assessment and Recommendation Engine (CARE)—to improve patient-centric treatment and diagnosis without real-time monitoring and 360 view capabilities. Abinaya et al. implemented a fascinating e-Health service application for diagnosing heart diseases. Balladini et al. designed a real-time architecture of big data for Francisco Lopez Lima Hospital in Argentina to process physiological data. This platform did not include predictive analytic and patient 360 view. Belle et al. implemented a genomic data processing platform that provides image analytic and signal processing of psychological data. Mezghani et al. designed a big data platform for integrating heterogeneous wearable data in healthcare for real-time monitoring and diagnosis. Lastly, Chen et al. presented a real-time big data platform to improve communication and collaboration between patients and providers, increasing the quality of care that clinical teams can provide.

6. Conclusions and Future Work

This paper provides details of an optimized and secure healthcare platform that revolutionizes the healthcare industry in Colombia by providing better information to patients and care teams. The use of this technology reduces the costs associated with healthcare.

The proposed digital health platform allows us to address population health challenges, to understand better patient's health, and to find hidden patterns that traditional data analytics fail to

find. The organization can use unified patient-generated data, financial data, and socioeconomic data to detect patterns and to discover a group of patients who share similar health behavior. The analysis of clinical and non-clinical data allows predicting patient's health with better accuracy. The platform also allows better health discoveries and actions based on treatment history for individuals and groups of patients.

Keralty organization recognized that better care coordination was required for patients receiving care. The organization wanted to improve quality outcomes, provider engagement and recruitment, and its own economic health. To meet these objectives, the organization focuses on clinician engagement and organizational alignment, ensuring widespread access to meaningful, actionable data, and the use of the healthcare analytics platform to inform decisions and drive improvement. Keralty believes the use of machine learning will be one of the most important, life-saving technologies ever introduced to the organization. We believe the opportunities are virtually limitless for the platform to improve and accelerate clinical, workflow, and financial outcomes.

More future work needs to be done on the platform to continue improving all the benefits for the entire organization. Tools for performing knowledge discovery process will be added to the ecosystem. The organization is planning to start the implementation of prescriptive analytics models to assist the organization in making smarter decisions in population health management. The architecture team will look at the possibility of implementing Map/Reduce-based computations for processing data with high scalability and to execute low latency and high concurrency analytical queries on top of Hadoop clusters.

Author Contributions: Conceptualization, F.L.-M., V.G.-D. and E.R.N.-V.; Methodology, F.L.-M., V.G.-D. and E.R.N.-V.; Software, F.L.-M.; Validation, F.L.-M., V.G.-D., Z.B. and E.R.N.-V.; Formal Analysis, F.L.-M.; Investigation, F.L.-M., V.G.-D., Z.B. and E.R.N.-V.; Resources, F.L.-M.; Data Curation, F.L.-M. and Z.B.; Writing—Original Draft Preparation, F.L.-M., V.G.-D. and E.R.N.-V.; Writing—Review and Editing, F.L.-M., V.G.-D., Z.B. and E.R.N.-V.; Visualization, F.L.-M., Z.B. and E.R.N.-V.; Supervision, F.L.-M. and V.G.-D.; Project Administration, V.G.-D. and E.R.N.-V.; Funding Acquisition, F.L.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This document presents an independent study supported by the company Sanitas USA. The points of view expressed are those of the authors and not necessarily those of Sanitas USA. We thank Ivan Murcia VP of Healthcare Services at Sanitas USA and Santiago Thovar, CIO at Keralty who provided insight and expertise that greatly assisted the study.

Conflicts of Interest: The authors declare no conflict of interest. The founders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACL	Access Control List
BI	Business Intelligence
CRM	Customer Relationship Management
EHR	Electronic Health Record
ERP	Enterprise Resource Planning
GDPR	General Data Protection Regulation
HIS	Hospital Information System
HIPAA	Health Insurance Portability and Accountability Act
HITRUST	Health Information Trust Alliance
LIS	Lab Information System
MPP	Massive Parallel Computing
RIS	Radiology Information System
REST	Representational State Transfer

References

1. Glassman, A.; Giuffrida, A.; Escobar, M.L.; Giedion, U. Chapter 1 Colombia: After a Decade of Health System Reform. In *From Few to Many*; Inter-American Development Bank: Washington, DC, USA, 2009; Volume 1, pp. 1–13.
2. Ruíz, F.; Gaviria, A.; Norman, J. Plan Decenal de Salud Pública. *Bogotá* **2020**, in press.
3. Legido, H.; Lopez, P.A.; Balabanova, D.; Perel, P.; Lopez-Jaramillo, P.; Nieuwlaat, R.; Schwalm, J.D.; McCready, T.; Yusuf, S.; McKee, M. Patients' knowledge, attitudes, behaviour and health care experiences on the prevention, detection, management and control of hypertension in Colombia: A qualitative study. *PLoS ONE* **2015**, *10*, e122112. [CrossRef]
4. Lopez, F.E.; Bonfante, M.C.; Arteta, I.G.; Baldiris, R.E. IoT and big data in public health: A case study in Colombia. In *Protocols and Applications for the Industrial Internet of Things*; IGI Global: Hershey, PA, USA, 2018; pp. 309–321, ISBN 978-1-5225-3806-6.
5. Dennis, R.J.; Caraballo, L.; García, E.; Rojas, M.X.; Rondon, M.A.; Pérez, A.; Aristizabal, G.; Peñaranda, A.; Barragan, A.M.; Ahumada, V. Prevalence of asthma and other allergic conditions in Colombia 2009–2010: A cross-sectional study. *BMC Pulm. Med.* **2012**, *12*, 12. [CrossRef]
6. About Keralty. Available online: https://www.keralty.com/en/about-keralty (accessed on 27 January 2020).
7. León, G.R. Digitalización de Historia Clínica. Available online: https://contrataciondelestado.es/wps/wcm/connect/3236c434-7ce1-484f-bb50-b8942bdc7d66/DOC20190314132936Estandar_digitalizacion_SACYL-+9.pdf?MOD=AJPERES (accessed on 27 January 2020).
8. Zhang, Y.; Qiu, M.; Tsai, C.W.; Hassan, M.M.; Alamri, A. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Syst. J.* **2017**, *11*, 88–95. [CrossRef]
9. Mezghani, E.; Exposito, E.; Drira, K.; Da Silveira, M.; Pruski, C. A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare. *J. Med. Syst.* **2015**, *39*. [CrossRef]
10. Kaur, P.; Sharma, M.; Mittal, M. Big Data and Machine Learning Based Secure Healthcare Framework. *Proc. Procedia Comput. Sci.* **2018**, *132*, 1049–1059. [CrossRef]
11. Thota, C.; Sundarasekar, R.; Manogaran, G.; Varatharajan, R.; Priyan, M.K. Centralized Fog Computing security platform for IoT and cloud in healthcare system. In *Fog Computing: Breakthroughs in Research and Practice*; IGI Global: Hershey, PA, USA, 2018; pp. 365–378, ISBN 978-1-5225-5650-3.
12. Edet, R.; Afolabi, B. Prospects and Challenges of Population Health with Online and other Big Data in Africa. *Adv. J. Soc. Sci.* **2019**, *6*, 57–63. [CrossRef]
13. MedAware—Using AI to Eliminate Prescription Errors—Digital Innovation and Transformation. Available online: https://digital.hbs.edu/platform-digit/submission/medaware-using-ai-to-eliminate-prescription-errors/ (accessed on 8 March 2020).
14. Ling, Z.J.; Tran, Q.T.; Fan, J.; Koh, G.C.H.; Nguyen, T.; Tan, C.S.; Yip, J.W.L.; Zhang, M. GEMINI: An integrative healthcare analytics system. *Proc. VLDB Endow.* **2014**, *7*, 1766–1771. [CrossRef]
15. Manogaran, G.; Thota, C.; Lopez, D.; Vijayakumar, V.; Abbas, K.M.; Sundarsekar, R. *Big Data Knowledge System in Healthcare*; Springer: Cham, Switzerland, 2017; pp. 133–157. [CrossRef]
16. Bates, D.W.; Saria, S.; Ohno-Machado, L.; Shah, A.; Escobar, G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* **2014**, *33*, 1123–1131. [CrossRef]
17. Farooqi, M.M.; Shah, M.A.; Wahid, A.; Akhunzada, A.; Khan, F.; ul Amin, N.; Ali, I. Big Data in Healthcare: A Survey. *Appl. Intell. Technol. Healthc.* **2019**, 143–152. [CrossRef]
18. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From big data to precision medicine. *Front. Media* **2019**, *6*, 34. [CrossRef]
19. Hatzigeorgiou, M.N.; Joshi, M.S. Population Health Systems: The Intersection of Care Delivery and Health Delivery. *Popul. Health Manag.* **2019**, *22*, 467–469. [CrossRef]
20. Koti, M.S.; Alamma, B.H. Predictive analytics techniques using big data for healthcare databases. In *Proceedings of the Smart Innovation, Systems and Technologies*; Springer Science and Business Media: Singapore, 2019; Volume 105, pp. 679–686. [CrossRef]
21. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, Analysis and Future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
22. Puaschunder, J.M. Big Data, Algorithms and Health Data. *SSRN Electron. J.* **2019**. [CrossRef]

23. Moreira, M.W.; Rodrigues, J.J.; Korotaev, V.; Al-Muhtadi, J.; Kumar, N. A Comprehensive Review on Smart Decision Support Systems for Health Care. *Inst. Electr. Electron. Eng.* **2019**, *13*, 3536–3545. [CrossRef]
24. López-Martínez, F.; Núñez-Valdez, E.R.; Lorduy Gomez, J.; García-Díaz, V. A neural network approach to predict early neonatal sepsis. *Comput. Electr. Eng.* **2019**, *76*, 379–388. [CrossRef]
25. López-Martínez, F.; Schwarcz.MD, A.; Núñez-Valdez, E.R.; García-Díaz, V. Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors. *Expert Syst. Appl.* **2018**, *110*, 206–215. [CrossRef]
26. Singh, A. Foundations of Machine Learning. *SSRN Electron. J.* **2019**, *486*. [CrossRef]
27. Rhee, C.; Dantes, R.; Epstein, L.; Murphy, D.J.; Seymour, C.W.; Iwashyna, T.J.; Kadri, S.S.; Angus, D.C.; Danner, R.L.; Fiore, A.E.; et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA J. Am. Med. Assoc.* **2017**, *318*, 1241–1249. [CrossRef]
28. Mahindrakar, P.; Hanumanthappa, M. Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges. *Int. J. Eng. Res. Appl.* **2013**, *3*, 937–941.
29. Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A.L.; Chen, I.Y.; Ranganath, R. A Review of Challenges and Opportunities in Machine Learning for Health 2018. *arXiv* **2018**, arXiv:1806.00388.
30. Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* **2020**, *104*, 101822. [CrossRef]
31. Palanisamy, V.; Thirunavukarasu, R. Implications of big data analytics in developing healthcare frameworks—A review. *J. King Saud Univ. Comput. Inf. Sci.* **2019**, *31*, 415–425. [CrossRef]
32. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*. [CrossRef] [PubMed]
33. Patel, S.; Patel, A. A Big Data Revolution in Health Care Sector: Opportunities, Challenges and Technological Advancements. *Int. J. Inf. Sci. Tech.* **2016**, *6*, 155–162. [CrossRef]
34. Chawla, N.V.; Davis, D.A. Bringing big data to personalized healthcare: A patient-centered framework. *J. Gen. Intern. Med.* **2013**, *28*. [CrossRef] [PubMed]
35. Abinaya, K. Data Mining with Big Data e-Health Service Using Map Reduce. *IJARCCE* **2015**, *4*, 123–127. [CrossRef]
36. Balladini, J.; Rozas, C.; Frati, F.; Vicente, N.; Orlandi, C. Big Data Analytics in Intensive Care Units: Challenges and applicability in an Argentinian Hospital. *J. Comput. Sci. Technol.* **2015**, *15*, 61–67.
37. Belle, A.; Thiagarajan, R.; Soroushmehr, S.M.R.; Navidi, F.; Beard, D.A.; Najarian, K. Big data analytics in healthcare. *BioMed Res. Int.* **2015**, *2015*. [CrossRef]
38. Chen, D.; Chen, Y.; Brownlow, B.N.; Kanjamala, P.P.; Arredondo, C.A.G.; Radspinner, B.L.; Raveling, M.A. Real-time or near real-time persisting daily healthcare data into HDFS and elasticsearch index inside a big data platform. *IEEE Trans. Ind. Inform.* **2017**, *13*, 595–606. [CrossRef]

Khalid Saeed
Jiří Dvorský (Eds.)

LNCS 12133

Computer Information Systems and Industrial Management

19th International Conference, CISIM 2020
Bialystok, Poland, October 16–18, 2020
Proceedings

Springer

Lecture Notes in Computer Science 12133