algorithms

MDPI

Article

Forecast of Medical Costs in Health Companies Using Models Based on Advanced Analytics

Daniel Ricardo Sandoval Serrano [1,2,†], Juan Carlos Rincón [1,†], Julián Mejía-Restrepo [1,†], Edward Rolando Núñez-Valdez [2,*,†] and Vicente García-Díaz [2,†]

[1] Corporate Data Management, Keralty, Calle 100 # 11b-67, Bogotá 111001, Colombia; drsandoval@keralty.com (D.R.S.S.); juancrincon@keralty.com (J.C.R.); juemejia@keralty.com (J.M.-R.)
[2] Department of Computer Science, Oviedo University, 33003 Oviedo, Spain; garciavicente@uniovi.es
* Correspondence: nunezedward@uniovi.es; Tel.: +57-31-0586-9385
† These authors contributed equally to this work.

Abstract: Forecasting medical costs is crucial for planning, budgeting, and efficient decision making in the health industry. This paper introduces a proposal to forecast costs through techniques such as a standard model of long short-term memory (LSTM); and patient grouping through k-means clustering in the Keralty group, one of Colombia's leading healthcare companies. It is important to highlight its implications for the prediction of cost time series in the health sector from a retrospective analysis of the information of services invoiced to health companies. It starts with the selection of sociodemographic variables related to the patient, such as age, gender and marital status, and it is complemented with health variables such as patient comorbidities (cohorts) and induced variables, such as service provision frequency and time elapsed since the last consultation (hereafter referred to as "recency"). Our results suggest that greater accuracy can be achieved by first clustering and then using LSTM networks. This implies that a correct segmentation of the population according to the usage of services represented in costs must be performed beforehand. Through the analysis, a cost projection from 1 to 3 months can be conducted, allowing a comparison with historical data. The reliability of the model is validated by different metrics such as RMSE and Adjusted R^2. Overall, this study is intended to be useful for healthcare managers in developing a strategy for medical cost forecasting. We conclude that the use of analytical tools allows the organization to make informed decisions and to develop strategies for optimizing resources with the identified population.

Keywords: cost; LSTM neural networks; cluster; health; cohorts

Citation: Sandoval Serrano, D.R.; Rincón, J.C.; Mejía-Restrepo, J.; Núñez-Valdez, E.R.; García-Díaz, V. Forecast of Medical Costs in Health Companies Using Models Based on Advanced Analytics. *Algorithms* **2022**, *15*, 106. https://doi.org/10.3390/a15040106

Academic Editor: Ivan Kisel

Received: 9 February 2022
Accepted: 20 March 2022
Published: 23 March 2022

1. Introduction

Healthcare is one of the largest industries and services of the global economy, one that has been significantly increasing until it becoming one of the biggest challenges of our time [1]. According to the World Health Organization (WHO), healthcare represented 7.56% of Europe's gross domestic product (GDP) in 2015 [2]. In 2018, the total healthcare expenditure of the United States was 16.8% of its GDP (the highest in the world) (WHO-GDP) [2]. The national healthcare expenditure of the United States in 2018 was USD 3.8 trillion, but forecasts show that these costs will increase up to USD 6.2 trillion dollars by 2028 [3]. Among others, one reason for this increase is the misuse of medication and the duplication of procedures by doctors [4].

In Colombia, according to the National Government, public health issues have been prioritized to guarantee equality; thus, for 2020, the budget was USD 8 billion, with an 8.12% increase since 2019, when it was USD 7.45 billion [5,6]. In this sense, the public health sector became one of the national sectors with the highest allocation of resources in the national budget.

To be in line with the General Health and Social Security System (SGSSS), the Keralty organization, one of the main actors in the Colombian health system [7], designed the

integrated care model from a four-goal perspective: (i) the value generated by interventions relative to health results; (ii) the experience of assisting people; (iii) cost-efficiency (sustainability, adequate and smart use of resources); and (iv) the experience and involvement of health teams that perform interventions—all this within a framework in which the focus is the care provided to people and the individual and aggregated results obtained for the incurred costs (efficiency).

The Colombian health system has two regimes: public and private. The Public Social Security and Health Regime (General Health and Social Security System, SGSSS) provides universal health coverage to the entire Colombian population and access to basic quality healthcare through the payment of fair premiums. The efficiency and quality of the service are the foremost priorities: the regime intends to improve health conditions by allocating resources for primary care, prevention in rural and vulnerable areas, and making sure that all health services meet the highest possible standards based on the available resources [8]. The general social security system has two regimes: contributive and subsidized. The contributive regime covers formal workers, pensioners, and independent workers, while the subsidized plan covers any other person who cannot afford it [9,10].

In the private health regime, people voluntarily choose a private and supplemental health insurance policy once they have fulfilled their economical obligation to contribute to the SGSSS [9]. In the private health regime, prepaid health service companies finance the risk that a person may face when getting sick. This means that a person voluntarily selects a healthcare plan to pay in advance for any type of expenses related to an eventual sickness. This means that the client agrees to pay a fee for the service, and the company must issue a financing contract with the coverage conditions of the plan and the corresponding rate. Both the public and private regimes require the anticipation of medical costs to facilitate their planning.

The implementation of advanced analytics projects allows the different companies of the Keralty group to anticipate potential changes in medical costs [7]. This article presents two proposals based on the exploration of variables such as comorbidities, seniority, residence, age, gender, and economic situation, among others. It is critical to understand how to project costs. First, prediction through LSTM networks and the use of grouping by characteristics allows segmenting the population to project the costs of a particular population using LSTM networks.

LSTM stands for "long short-term memory", introduced as an improved RNN algorithm in 1997 [11]. LSTMs are an extension of previous RNNs which are able to retain a memory in the long term and use it to learn patterns in longer sequences of source data. Before LSTMs, RNNs were *forgetful*. They could retain a memory, but only about the steps of the process in its immediate past. LSTMs, however, introduce loops that can generate long-lasting gradients [12,13]. They can retain the long-term patterns they discover as they run along with their loops.

The other technique we used was clustering: it could also be considered an exploratory data analysis (EDA) technique that helps discover hidden patterns or data structures. The clustering technique may also work as an independent tool to obtain information about data distribution [11]. A cluster is the collection of data objects that resemble each other within the same group (class or category), and which are different from the objects of the other clusters [13]. Clustering is an unsupervised learning technique where there are predefined classes and previous pieces of information that define how data must be grouped or labeled in separate classes.

There are a variety of clustering algorithms, and the most popular ones include hierarchical clustering [14,15], Gaussian mixture models [16,17], and others within the Sklearn package [18]. In our case, we used k-means, an algorithm that consists of dividing the data points of x by a set of k clusters, where each data point is allocated to its closest cluster. This method is defined by the target function which tries to minimize the sum of all the squared distances within one cluster and for all clusters [19,20]. This work shows the process of grouping and the classification of accredited health entities using k-means [21].

This allowed the accredited health sector institutions to be grouped into two large clusters. The first was defined as institutions in the process of financial consolidation; and the second cluster was defined as large health institutions. The business profiles of the institutions under study were thus defined.

Specifically, we summarize our contribution as follows. First, we predict the medical cost of a healthcare organization using the described techniques and suggest an avenue of improvement in further work: namely that understanding how and why cost-drivers increase may provide information about the risk factors and the possible starting points for defining preventive measures and strategies.

This paper is structured as follows. In Section 2, we show related works. In Section 3, we describe the methodology and information about the data, data-processing operations, and the methods we used to evaluate the problem. In Section 4, we first present the results obtained with the LSTM networks and continue presenting the results obtained from combining cluster segmentation with LSTM networks. We then proceed in Section 5 to discuss the results to finally summarize the conclusions and directions for future research.

2. Related Work

The cost forecast is one of the main objectives of different time series methods when these methods are applied in diverse fields. A time series is a sequence of measurements over time rarely mapped in equal intervals. Time series forecasting can be applied to diverse sectors, and in this case, specifically to the prediction of medication costs as performed in papers by, e.g., Jaushic and Shruti [12,22], using different techniques such as ARIMA and LSTM. Another work by Kabir [23] using RL, RNN, and LSTM showed a sustainable approach to forecast the future demands of hospital beds, considering the hospital capacity and the population of the region in order to plan the future increase in required hospital beds. Scheuer [24] used electronic medical records for Finnish citizens over sixty-five years of age to develop a sequential deep learning model to predict the use of health services in the following year using RNN and LSTM networks. Another work which uses clustering techniques is that by Mahmoud [25]. This author studied hip fracture care in Ireland and, using k-means clustering, showed that elderly patients are grouped according to three variables: age, length of stay, and time to surgery. According to Mahmoud, the cost of treating a hip fracture was estimated to be approximately EUR 12,600. He identified hip fractures as one of the most serious injuries with long hospital admissions.

In addition, Miroslava [26] used k-means to find the most appropriate clinical variables between 23 and 26 variables capable of efficiently separating patients diagnosed with type 2 diabetes mellitus (T2DM) with underlying diseases such as arterial hypertonia (AH), ischemic heart disease (CHD), diabetic polyneuropathy (DPNP), and diabetic microangiopathy (DMA).

The following Table 1 provides a summary of the related papers and their input variables.

Table 1. Models and input variables from related papers.

Paper	Method	Cost Variables	No-Cost Variables
Kaushik (2017) [12]	Arima, LSTM	Medication cost	Demographic variables of patients (age, gender, region, year of birth) and clinical variables of patients (type of admission, diagnoses, and procedure codes)
Shruti (2020) [22]	Arima, LMP, LSTM	Medication cost	Predict the average weekly expenditure of patients on certain pain medications, selecting two medications that are among the ten most prescribed pain medications in the US
Kabir (2021) [23]	RL, RNN, LSTM	Bed cost	Number of beds, occupation, and patients
Scheuer (2020) [24]	Lasso, LightGBM, LSTM	Cost of visits by family doctor	Number of patients, number of visits, average visits per patient, procedure codes, and diagnoses

3. Materials and Methods

This study explores two different approaches to forecasting medical costs in the Colombian public health insurance. The steps of the methodology applied to meet the objectives of this paper are shown in Figure 1.

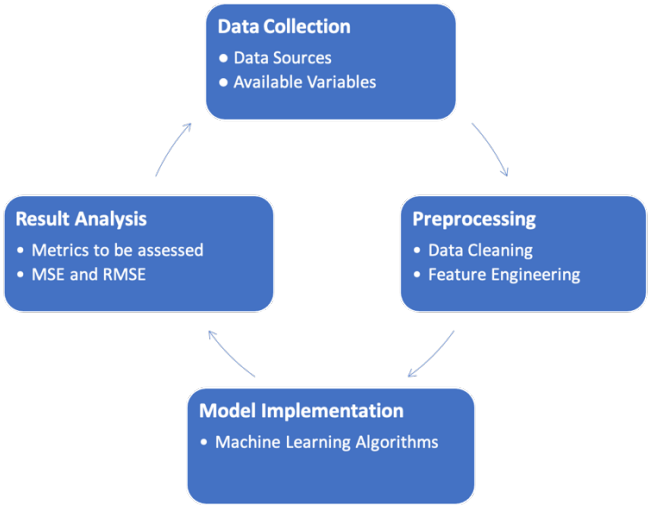

Figure 1. Steps of the implemented methodology.

3.1. Data Collection

In this research, we used datasets from the Keralty health company [7]. The data for this retrospective analysis were obtained from one of the modules of medical and affiliate accounts of the Core Beyond Health application developed by Sonda [27]. This includes invoices from medical services corresponding to patient assistance through the public health plan. We also used the Vacovid repository (Proprietary Source) to obtain the information of patients that are classified within any health conditions or cohorts. The dataset contains all the information available on the costs of services received by the users between 2017 and 2021. Figure 2 shows the datasets and the variables of each data source.

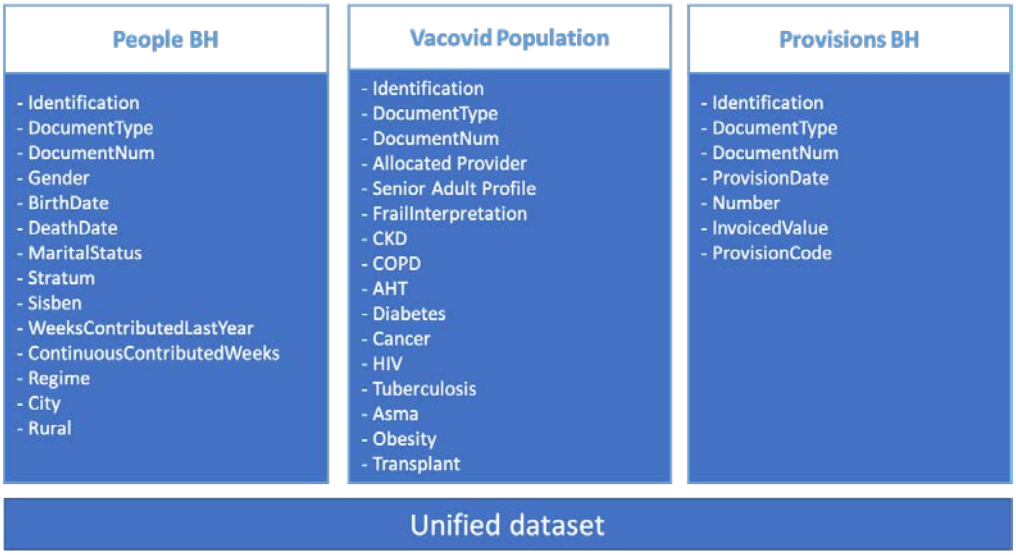

Figure 2. Variables by source.

3.2. Data Processing

In this step, we transformed raw data into an adequate and understandable format. In the real world, datasets contain errors. Therefore, this step solves errors and the datasets

become easy to manage [28]. Below, we briefly describe the most important data we followed in each dataset:

The following transformations from the data:

1. Dates are converted into DateTime Y%–M%–D% and thus dates are formatted;
2. Empty fields of dates are denoted by 1900–01–01;
3. Empty fields are mapped in 0 values;
4. The "TotalComorbidities" field is created, allowing to identify the number of diagnoses or cohorts of a patient;
5. Category values are encoded;
6. Mappings to a dictionary of types of documents;
7. Exceedingly small provision values of less than 1000 are disregarded;
8. DateTime Y%–M%–D% dates are formatted;
9. The "Number" and "InvoicedValue" fields are converted into int. format.

After unifying and cleaning the dataset, we ended up with a total of 160,463,128 entries about the invoices for the provided medical services. Table 2 shows the variables selected to work in the simulators with a 5% sample corresponding to 3,202,610 services with 34 different attributes. The output variable in this study is "InvoicedValue".

Table 2. Selected attributes.

Id	Column	Entries	Description
0	ProvisionDate	3,202,610	Service provision date
1	Identification	3,202,610	Affiliate identification
2	ProvisionCode	3,202,610	Provision identification
3	Number of services	3,202,610	Number of invoiced services
4	InvoicedValue	3,202,610	Invoice value
5	Principal_Group_id	3,202,610	Principal grouping, e.g., surgery
6	Group_1_id	3,202,610	e.g., hospital surgery
7	Group_2_id	3,202,610	e.g., abdominal/neck/neurosurgery
8	Group_3_id	3,202,610	e.g., bariatric appendicectomy
9	Gender	0—84,011 1—1,965,111 2—1,153,488	Gender 0—no data 1—men 2—women
10	BirthDate	3.202,610	Date of birth of the affiliate
11	DeathDate	3,202,610	Date of death of the affiliate
12	MaritalStatus	3,202,610	Marital status (married/single/divorced)
13	Stratum	0—1,168,949 1—22,069 2—19,375 3—1,937,099 4—26,785 5—7088 6—21,275	Socioeconomic stratum 0—no data 1—low–low 2—low 3—medium–low 4 —medium 5—medium–high 6–high
14	Sisben	3,202,610	Marks if a beneficiary of social programs
15	WeeksContributedLastYear	3,202,610	Weeks contributed to the last year
16	ContinuousContributedWeeks	3,202,610	Weeks contributed since first affiliation
17	Regime	3,202,610	Contributive or subsidized
18	City	3,202,610	City where the service was provided
19	Rural	3,202,610	People living in the countryside. not in cities
20	CKD	No—3,060,478 Yes—142,132	If patient has a chronic kidney disease
21	COPD	No—3,058,721 Yes—143,889	If patient has COPD

Table 2. *Cont.*

Id	Column	Entries	Description
22	AHT	No—2,318,889 Yes—883,721	If patient has arterial hypertension
23	Diabetes	No—2,841,842 Yes—360,768	If patient has diabetes
24	Cancer	No—3,047,414 Yes—155,196	If patient has cancer
25	HIV	No—3,180,665 Yes—21,945	If patient has HIV
26	Tuberculosis	No—3,201,777 Yes—833	If patient has tuberculosis
27	Asma	No—3,139,088 Yes—63,522	If patient has asthma
28	Obesity	No—2,404,289 Yes—798,321	If patient has obesity
29	Transplant	No—3,190,156 Yes—12,454	If patient has transplant
30	SeniorAdultProfile_id	3,202,610	Marks if a person is a senior adult
31	FrailInterpretation_id	3,202,610	Score to measure frailty diagnosis
32	AllocatedProvider_id	3,202,610	Provider allocated for vaccination
33	TotalComorbidities	0—1,842,012 1—588,168 2—439,766 3—237,582 4—75,496 5—17,284 6—2183 7—119	Number of cohorts of a person 0—no cohorts 1—with one cohort 2—with two cohorts 3—with three cohorts 4—with four cohorts 5—with five cohorts 6—with six cohorts 7—with seven cohorts
34	Age	3,202,610	Age

In Table 3, we show the Spearman correlation coefficients between the selected variables and the invoiced values for patients who are not marked with any morbidity: "Without comorbidity" means that it is not classified under any health cohort, as well as those marked with at least one morbidity "With one morbidity" designates patients belonging to at least one or more health cohorts. Similarly, in Table 4, we show the Pearson correlation coefficients between the listed variables and the invoiced value for patients within each cohort or pathology. This process allowed us to identify the most statistically significant variables that can be associated with the medical cost.

Table 3. Correlation of variables with or without morbidity with the InvoicedValue.

Variable	Without Comorbidity	With One Morbidity
Gender	−0.007411	−0.001028
Principal_Group_id	−0.002513	0.056446
Stratum	0.017053	0.042861
City	0.072799	0.034980
SeniorAdultProfile_id	0.003306	0.002666
FrailInterpretation_id	0.000963	−0.000554
AllocatedProvider_id	0.081264	0.072986
Age_Provision	−0.049595	0.043494
WeeksContributedLastYear	0.003423	0.022014
ContinuousContributedWeek	0.002380	0.038922
Number of services	0.400405	0.443571

Table 4. Correlation with the InvoicedValue field.

Variable	CKD	COPD	AHT	Diabetes	Cancer	HIV	Tuber	Asma	Obesity	Transplant
Gender	0.017622	0.000404	0.004497	0.004986	−0.009991	−0.002089	0.270194	0.046490	−0.023649	0.028032
Principal_Group_id	0.079713	0.212875	0.082054	0.095878	−0.009113	−0.416173	0.257335	0.161863	0.068553	−0.450470
Stratum	0.028784	0.049269	0.043887	0.052435	−0.022120	−0.037370	−0.112660	0.068190	0.043283	−0.029483
City	0.007663	−0.018505	0.027033	0.020959	0.003355	0.212705	0.224876	−0.027337	0.036065	0.065647
SeniorAdultProfile_id	0.025530	0.029828	0.000947	−0.005390	0.044711	0.017209	0.080920	0.006746	−0.008730	0.061342
FrailInterpretation_id	−0.018983	0.017184	−0.002330	−0.018542	0.030893	0.030642	0.063878	−0.001963	−0.013096	0.028646
AllocatedProvider_id	0.006603	0.084994	0.070379	0.083724	0.053718	0.108541	0.272037	0.107602	0.070681	0.055575
Age_Provision	0.046911	0.054943	0.079575	0.086535	−0.034936	−0.006842	−0.261186	0.120718	0.024812	−0.040887
WeeksContributed-LastYear	0.019178	0.016896	0.020329	0.020762	−0.004686	0.000072	0.202423	0.060753	0.018878	0.039546
ContinuousContr-ibutedWeeks	0.031204	0.039401	0.041397	0.051791	−0.018609	−0.045660	−0.072248	0.088931	0.038706	−0.046542
Number of services	0.469864	0.541773	0.456648	0.480466	0.425494	0.251777	0.706559	0.485456	0.439157	0.304911

The only variable that has a relationship with the cohorts with a correlation coefficient close to 0.5 which is "Number of services"; if a coefficient that is assigned is a substantial (negative or positive) number, it has influence on the prediction. Conversely, if the coefficient is zero, it has no impact on the prediction.

3.3. Model Implementation

The cost forecast was performed under two proposals: cost analysis by selecting the variables using LSTM neural networks, and finally, segmentation through the *Cluster* to analyze the cost of each cluster using the same techniques. Our deep learning LSTM regression model was developed, with Keras [29,30] and Sklearn [31], using *Python* programming language [32]. We also used *Streamlit* [33], which allowed us to create a web application to display our results, and Google Cloud Platform AI Platform [34], to train the automatic learning models, host the model in the Cloud and finally make the model available for the users on cloud storage. The usage of LSTM networks is motivated by the long and short-term seasonalities involved in the medical cost time series, such as Christmas, summer, and weekdays. This makes the usage of LSTM models more appropriate.

3.3.1. LSTM Networks

This neural network, over time, can connect three pieces of information: current input data; the short-term memory received from the preceding cell (the so-called hidden state); and the long-term memory of more remote cells (the so-called cell state)—from which the RNN cell produces a new hidden state [12]. Figure 3 shows an LSTM memory cell.

Machine learning algorithms work best when numerical inputs are scaled to a standard range. Normalization and standardization are the two most popular techniques for scaling numerical data before modeling. Normalization scales each input variable separately to the range of 0–1, which is the range for floating-point values where we have the highest accuracy. Standardization scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to change the distribution to have a mean of zero and a standard deviation of one.

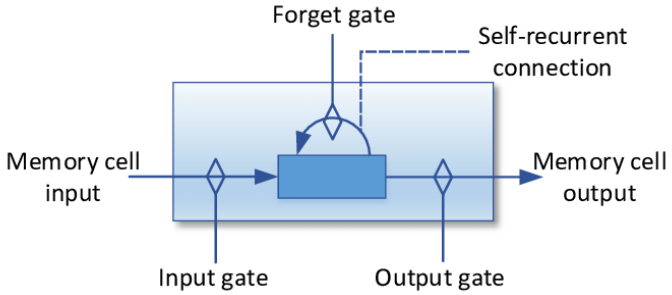

Figure 3. LSTM memory cell illustration [35,36].

To normalize the data and feed the LSTM, we used *MinMaxScaler* from *sklear.preprocessing* to scale our data between -1 and 1. The feature range parameter was used to specify the range of the scaled data. Then, we converted the training and test data into a time series problem; we must predict a value in time T based on the month data. To train the LSTM network with our data, we needed to convert the data into a 3D format in the form accepted by LSTM. This means that the input layer expects a 3D data matrix when fitting the model and making predictions, even if the specific dimensions of the matrix contain only one value, for example, a sample or a feature. When defining the input layer of your LSTM network, the network assumes that you have one or more samples and requires that you specify the number of time steps and the number of features.

There is not a general rule as to how many nodes or how hidden layers must be elected, and very often a trial-and-error approach may yield the best results for each problem [37]. As this is a simple network, we started trying with four neurons, then with eight, and finally, a test was performed with sixteen neurons, which was the first parameter of the LSTM layer. The second parameter was "return sequences", which was established in false, as we did not add more layers to the model. The last parameter was the number of indicators [12]. We also added an exclusion layer to our model to prevent overfitting. Finally, we added a dense layer at the end of the model; the number of neurons on the dense layer was established at 1, as we wanted to predict a single number value in the output. In this paper, we used the Adam optimizer [38] and we used mean squared error as the loss metric [39] to show the implementation of the LSTM network.

Some of the parameters that can be modified and which are very important to achieving the good performance of the model are the activation function and the cost function. Activation functions largely control what information is propagated from one layer to the next. By combining non-linear activation functions with multiple layers, network models are able to learn non-linear relationships. The most commonly used activation functions are relu and sigmoid. The activation function relu will generate an output equal to zero when the input is negative, and an output equal to the input when the input is positive. As such, the activation function retains only the positive values and discards the negative ones, giving them an activation of zero. The sigmoid activation function takes any range of values at the input and maps them to the range of 0–1 at the output.

Another parameter is the cost function, also called the loss function, which quantifies the distance between the actual value and the value predicted by the network. In other words, it measures how incorrect the network is when making predictions. In most cases, the cost function returns positive values. The network's predictions are improved when the cost value is close to zero.

An epoch corresponds to the number of times that the algorithms will be executed. In each cycle (epoch), all the training data pass through the neural network so that it learns about them:

```
model = Sequential()
model.add(LSTM(2))
model.add(Dropout(0.2))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mean_squared_error')
model.fit(batch_size=1, verbose=0, epochs = 20, shuffle = False)
```
(1)

A long short-term memory network (LSTM) is one of the most popular neural networks for analyzing time series. The ability of an LSTM to remember previous information makes it ideal for such tasks [40].

3.3.2. Clusters

In this case, we use it to try to identify patients with the same characteristics, as shown in Figure 4.

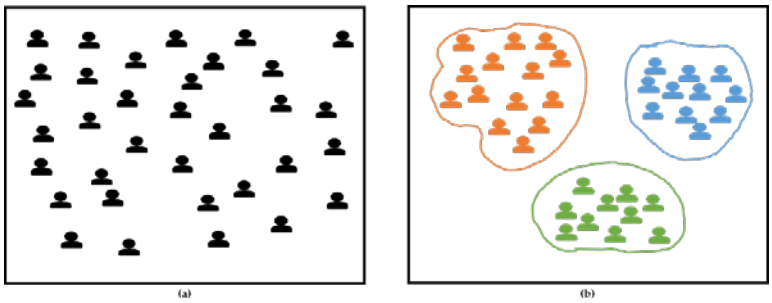

Figure 4. Cluster structure: (**a**) before cluster; (**b**) after cluster.

To implement the k-means clustering algorithm, one must first choose a k value, i.e., the number of clusters to be formed. Then, one must randomly select k data points from the dataset as the centers/initial centers of the clusters. Then, the distance between the data point and the cluster's centroid is calculated; as such, each datum is assigned to the cluster with the closest centroid. For each cluster, the new mean is estimated based on the data points of the conglomerate. This does not end until the mean of the clusters remains stable under a predetermined variation limit or until the maximum number of iterations is reached.

For the clustering process carried out in this paper, we considered the related co-morbidities, namely "Age"; "WeeksContributedLastYear", corresponding to the weeks contributed to in the last year; "ContinuousContributedWeeks", corresponding to the weeks contributed since first affiliation—in addition to two new variables which are "frequency", corresponding to the number of services provided a to patient; and "recency", corresponding to the last time they received medical assistance. In addition, the cohort variables are used for CKD, COPD, AHT, diabetes, cancer, HIV, tuber, asthma, obesity, and transplant.

We determined the most suitable number of clusters through the elbow method [41,42]. To this end, we varied the number of clusters from 1 to 20 and calculated the WCSS (within-cluster sum of squares). This designates the sum of squared distances between each point and the centroid in the calculated clusters. The point after which the curve does not decrease quickly is the appropriate value for K, as shown in Figure 5.

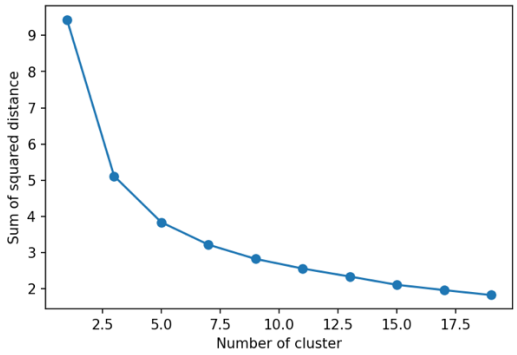

Figure 5. The elbow method is used to determine the number of clusters [41,42].

After choosing the number of clusters, a manual description of the characteristic of each cluster was made to be able to identify each group, as seen in Table 5.

To confirm the result of the optimal number of clusters indicated by the elbow technique, we ran the silhouette method, which is also a method for finding the optimal number of clusters, interpretation, and the validation of the consistency of data within clusters. See Table 6. The silhouette method calculates the silhouette coefficients of each point, which measure the extent to which a point resembles its own cluster compared to other clusters.

Table 5. Cluster manual description.

Cluster	Description
0	HighAge, COPD-AHT
1	YoungAdult, HEALTHY
2	Adult, AHT-OBESITY
3	SeniorAdult, AHT
4	Adult, OBESITY
5	SeniorAdult, AHT-DIABETES-OBESITY
6	Inactive
7	SeniorAdult OBESITY-AHT
8	SeniorAdult, HEALTHY
9	SeniorAdult, CANCER-AHT
10	HighAge, CKD-AHT
11	Young, HEALTHY, LittleUse
12	Adult, CANCER
13	HighAge, COPD-AHT-OBESITY
14	Young, HEALTHY, RecentUse

Table 6. Silhouette score for k (clusters).

K (Clusters)	Silhouette Score
4	0.41823
5	0.43770
6	0.30693
7	0.32616
8	0.333503
9	0.34014
10	0.31921
11	0.32706
12	0.33285
13	0.344021
14	0.30254
15	0.34314

In this case, the optimal number of clusters is 5; however, for a better differentiation of patients with different medical conditions in cohorts and according to suggestions from clinical experts inside in our organization, in the interest of observing, over a period of time, which of these groups did or did not have the expected outcome associated with mortality, higher fatality events and higher cost events, it was decided that a total of 15 clusters would be used.

4. Results

After applying the clustering and training predictive models using the LSTM network, we found a set of features that give the best performance. These features are shown in Table 7 below.

Table 7. Proposed models with specific parameters.

Method	Parameters
LSTM	16, batch_input_shape= (1, X_train. shape[1], X_train.shape[2]), stateful=True)
Clustering	n_cluster = 15, scale_method = 'minmax', max_iter = 1000

For both models, the LSTM network model and clustering were executed and the data were grouped into two variables, namely ProvisionDate and InvoicedValue, to predict the cost of services for more than 1,558,613 patients in the sample between 2017 and 2021. The first 80% were used to train the models, and the remaining 20% were used to assess them.

4.1. LSTM Networks

For a summary of the model run with sixteen hidden memory cells, see (2):

$$
\begin{array}{lll}
\text{Layer(type)} & \text{Output Shape} & \text{Param\#} \\
\hline
\text{lstm\_8(LSTM)} & (1, 16) & 1280 \\
\text{dropout\_8(Dropout)} & (1, 16) & 0 \\
\text{dense\_8 (Dense)} & (1, 1) & 17 \\
\hline
\text{Total params: 1297} & & \\
\text{Trainable params: 1297} & & \\
\text{Non-trainable params: 0} & &
\end{array}
\tag{2}
$$

For the result of the execution of the last three epochs, see (3):

$$
\begin{array}{l}
\text{Epoch } 18/20 \\
46/46\ [==========]\ -\ 0\ \text{s}\ 1\ \text{ms/step}\ -\ \text{loss: } 0.0739 \\
\text{Epoch } 19/20 \\
46/46\ [==========]\ -\ 0\ \text{s}\ 1\ \text{ms/step}\ -\ \text{loss: } 0.0643 \\
\text{Epoch } 20/20 \\
46/46\ [==========]\ -\ 0\ \text{s}\ 1\ \text{ms/step}\ -\ \text{loss: } 0.0690
\end{array}
\tag{3}
$$

Table 8 shows the RMSE for standard models with different numbers of memory cells. The lowest RMSE was obtained (=89.03) for a standard LSTM with 16 hidden memory cells.

Table 8. Results for the different memory cells.

No. of Layers	No. of Memory Cells	RMSE
1 standard LST	4	104,06
	6	93,12
	8	93,78
	10	92,12
	12	94,28
	14	95,99
	16	89,03

One of the features is the prediction of a particular population, showing the current and projected cost. This feature allows us to filter by conditions such as gender, healthcare regime, marital status, and whether they have a cohort or condition such as diabetes, CKD, hypertension. Additional cohort variables can be projected for one to three months. Figure 6 shows the result with the following filters: woman as gender and diabetes condition.

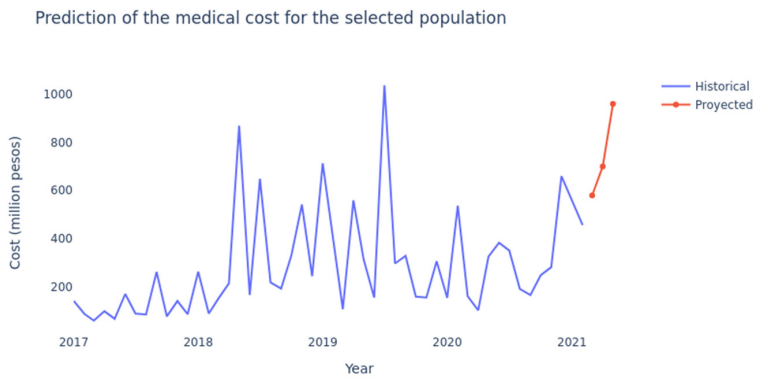

Figure 6. Projected cost using the LSTM network.

4.2. Clustering

In this section, we visually explore the discovered clusters to look for relations and insights. The clusters are examined with respect to patient characteristics, outcomes, and standards of care considering variables such as age, frequency, and recency. A discussion is then presented to better interpret these results.

4.2.1. Distribution by Age Cluster (in Years)

First, we explored the clusters discovered in terms of age. In the Figure 7, the behavior of age is represented by the identified clusters, in which we can see that the clusters (0, 3, 7, 8, 10, 12 and 13) show older people with some health condition, in comparison with the other clusters that show that the population is concentrated on younger people.

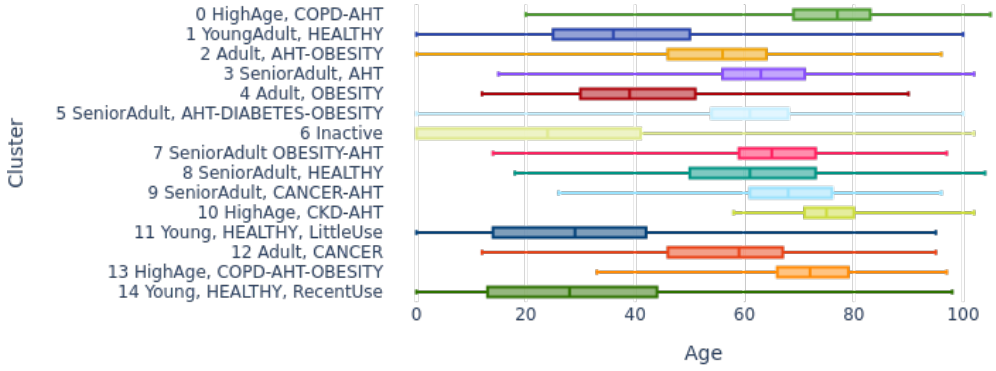

Figure 7. Distribution by age cluster (in years).

4.2.2. Distribution by Frequency of Use Cluster

Second, we explored the variable frequency, as shown in Figure 8, where it can be observed that all the people in the groups are attending medical consultations quite often.

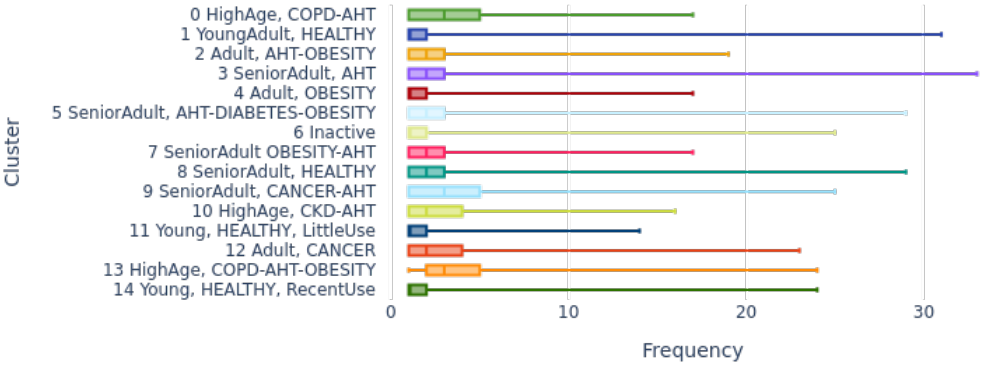

Figure 8. Distribution by frequency of use cluster.

4.2.3. Distribution by Cluster of Last Attention Time (Recency)

We also explored the users by the variable recency, as can be seen in Figure 9, that measures the time elapsed since the last medical service. All of them have recently seen a doctor, unlike cluster 11 which comprises young patients which have not seen a doctor for a long time. The rest of the clusters have had at least one visit recently.

4.2.4. Distribution by Cluster of Weeks Contributed since Last Year

This corresponds to the number of weeks contributed since the last year Figure 10, showing outliers in clusters 1 and 4. The rest of the clusters show that all patients are continuous since their affiliation date. Cluster 14 shows people who are newly enrolled.

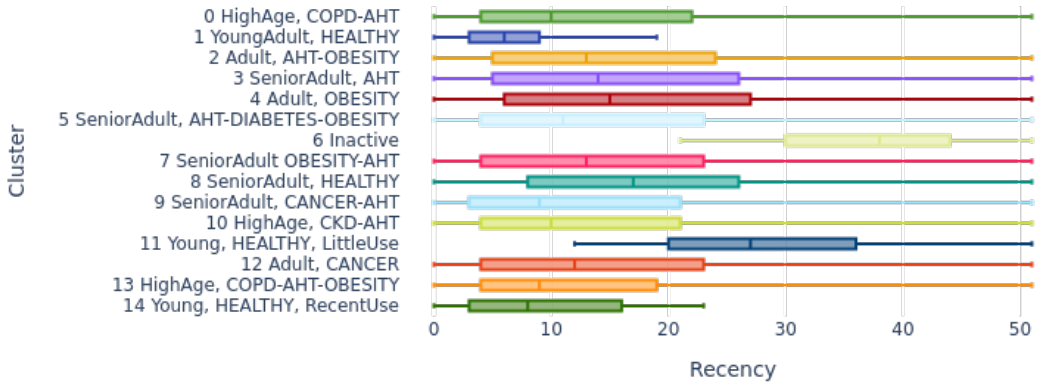

Figure 9. Distribution by cluster of last consultation time (Recency).

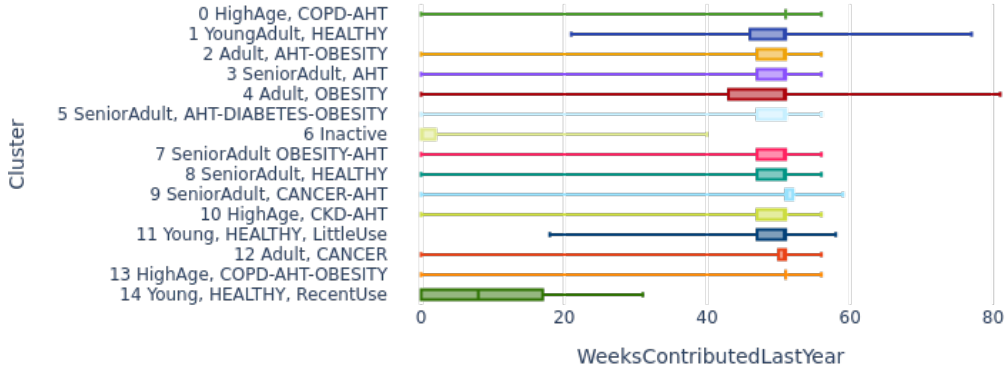

Figure 10. Distribution by cluster of weeks contributed since last year.

4.2.5. Distribution by Cluster of Continuous Contributed Weeks

This shows the number of weeks that the users have been affiliated since their first date of affiliation, as shown in Figure 11. It can be noticed that cluster 8 aggregates old healthy users that have been affiliated for a prolonged period.

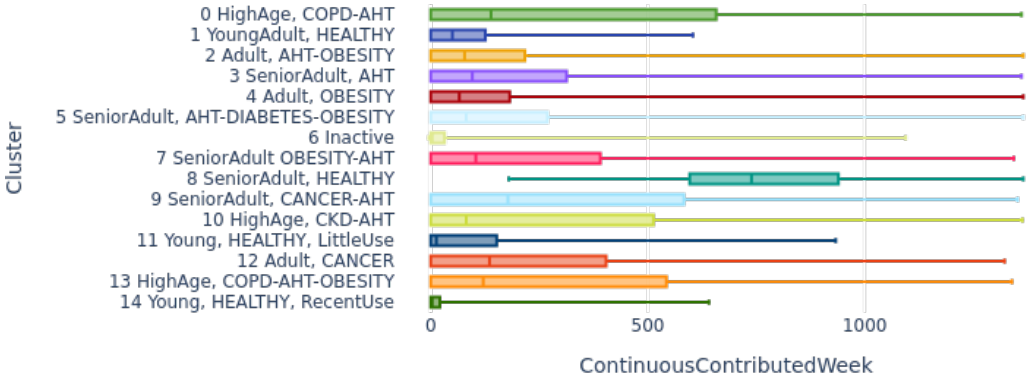

Figure 11. Distribution by cluster of continuous contributed weeks.

The model was evaluated with 4 and 16 memory cells, showing the reliability when first segmented by cluster, for all clusters except for clusters 1 and 3, where with 16, its results are better. As shown in the Table 9, it is preferable to use 4 memory cells.

After defining the clusters, and according to the cluster selection, we predicted the cost again using LSTM networks; this feature allows you to choose which cluster and over what period to project it. In this case, we chose cluster 3, resulting in the following projection as seen in Figure 12.

Table 9. Result for clusters with different RSMEs.

Cluster	Description	Number	RMSE (4)	RMSE (16)
0	HighAge, COPD-AHT	43.403	58,69	61,71
1	YoungAdult, HEALTHY	380.158	601,59	623,36
2	Adult, AHT-OBESITY	122.125	83,70	105,02
3	SeniorAdult, AHT	123.463	34,14	27,02
4	Adult, OBESITY	205.765	97,57	206,74
5	SeniorAdult, AHT-DIABETES-OBESITY	71.647	129,95	211,48
6	Inactive	154.907	274,06	418,10
7	SeniorAdult OBESITY-AHT	64.867	31,27	107,55
8	SeniorAdult, HEALTHY	71.372	89,20	98,81
9	SeniorAdult, CANCER-AHT	36.429	29,17	52,67
10	HighAge, CKD-AHT	51.153	85,02	114,07
11	Young, HEALTHY, LittleUse	411.973	463,20	445,10
12	Adult, CANCER	37.006	51,94	69,43
13	HighAge, COPD-AHT-OBESITY	33.504	15,15	25,98
14	Young, HEALTHY, RecentUse	11.3965	122,09	167,99

As such, patients were better modeled and performance was slightly increased, instead of working with the optimal values in performance provided by the elbow and silhouette methods (see Tables 9 and A1 for details of the performance of both approaches). It is also important to note that the allowance of 15 clusters, instead of 5, has also helped to identify two clusters of inactive patients (6) and 'Young and Healthy with Little Use' patients (cluster 11) whose predictability is not reliable ($R^2 < 0$) and could be biasing the models when using only five clusters.

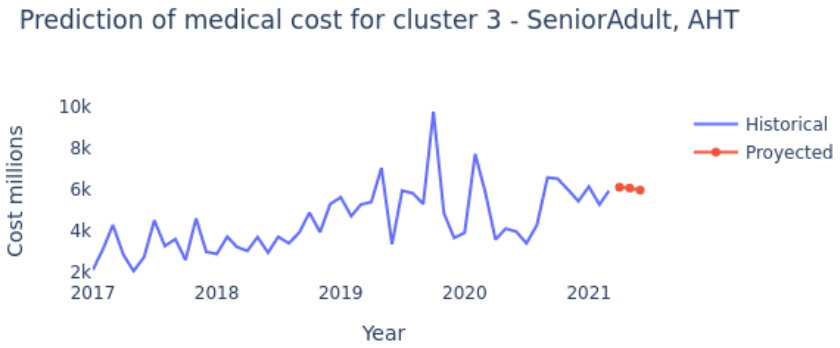

Figure 12. Cluster 3 medical cost projection.

We reviewed previous cost prediction model studies, namely a standard short-term memory model (LSTM) and a stacked LSTM model, to predict the monthly drug cost of more than 50,000 patients between 2011 and 2015. For the single-layer LSTM model, they obtained an RMSE value of 14.617 and an R^2 value of 0.8048. For the stacked LSTM model, the RMSE value was 13.693 and an R^2 value of 0.8159 [12]. Another works predicted the average weekly expenditure of patients on certain pain medications, with different models such as Arima, MLP, and LSTM selecting two medications among the 10 most prescribed pain medications in the US; the LSTM result yielded an RMSE value for medicine A of 143,69 and an R^2 value of 0.77 [22].

Below are the metrics we adopted for each model. These are: root mean square error (RMSE) [43,44]; mean absolute percentage error (MAPE) [45]; R^2; and adjusted R^2 [46]. The most common metric used for regression purposes is the root mean square error (RMSE) and it represents the square root of the average distance between the actual value and the predicted value. This indicates the absolute adjustment of the model to the data; how close are the observed data points to the model's predicted values. The RMSE measurement is an absolute mean of adjustment. As the square root of a variance, the RMSE can be interpreted

as a standard deviation of the unexplained variable, and it has the useful property of being in the same units as the response variable. Lower RMSE values indicate a better adjustment [47,48].

Mean absolute percent error (MAPE) measures the average percentage error. It is calculated as the average of the absolute percentage errors. MAPE is sensitive to scale and becomes meaningless for low volumes or data with zero demand periods. When aggregated or used with multiple products, the MAPE result is dominated by low volume or zero products [45].

R-squared and adjusted R-squared are often used for explanatory purposes and explain how well the selected independent variables explain the variability in their dependent variables. The coefficient of determination or R^2 is another measure used to assess the performance of a regression model. The metric helps us compare our current model to a constant baseline and tells us how much better our model is. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R^2 is a scale-free score which implies that regardless of whether the values are excessively large or excessively small, R^2 will always be less than or equal to 1 [22].

Adjusted R^2 represents the same meaning as R^2 but is an improvement on it. R^2 suffers from the problem that scores improve in increasing terms even though the model is not improving. The adjusted R^2 is always smaller than R^2 as it adjusts for increasing predictors and only shows an improvement if there is a real improvement [46].

In summary, when the LSTM network model is executed with the selected data, in this case, women in the diabetes cohort, the data are grouped into two variables, "ProvisionDate" and "InvoicedValue", which are those used in the network. The results are shown in Table 10.

Table 10. LSTM network model results.

Model	RMSE	MAPE	R^2	Adj. R^2
LSTM networks	89,03	36,25%	0.89	0.835

After segmenting patients and executing the LSTM network again for all clusters, we obtained the following results shown in Table 11.

Table 11. LSTM network model results after segmenting patients.

Cluster	Description	RMSE	MAPE	R^2	Adj. R^2
0	HighAge, COPD-AHT	58,69	28,25%	0.881	0.821
1	YoungAdult, HEALTHY	601,59	25,42%	0.925	0.888
2	Adult, AHT-OBESITY	83,70	15,80%	0.940	0.910
3	SeniorAdult, AHT	34,14	4,93%	0.996	0.993
4	Adult, OBESITY	97,57	17,42%	0.940	0.910
5	SeniorAdult, AHT-DIABETES-OBESITY	129,95	41,43%	0.818	0.727
6	Inactive	274,06	2405,8%	0.031	−0.453
7	SeniorAdult OBESITY-AHT	31,27	12,16%	0.941	0.912
8	SeniorAdult, HEALTHY	89,20	60,38%	0.753	0.629
9	SeniorAdult, CANCER-AHT	29,17	9,67%	0.994	0.991
10	HighAge, CKD-AHT	85,02	17,93%	0.878	0.818
11	Young, HEALTHY, LittleUse	463,20	341,29%	0.206	−0.191
12	Adult, CANCER	51,94	17,28%	0.959	0.939
13	HighAge, COPD-AHT-OBESITY	15,15	9,42%	0.971	0.957
14	Young, HEALTHY, RecentUse	122,09	21,37%	0.956	0.934

5. Discussion

The purpose of this paper was to show techniques for predicting the costs of patients. The first model is an approach to simulate costs considering the decrease or increase in a particular population of a certain cohort. With the projected cost for each cohort, in case it

decreases or increases, we can have an estimate of the costs that the company could save so that it can implement strategies such as investing in promotion and prevention plans for cohorts.

When we made the prediction with the initial values filtered by woman as gender and with diabetes using the LSTM networks, we observed that the RMSE metric shows that, on average, the mean prediction error corresponds to 89.03. In this case, MAPE indicates that, on average, the forecast is wrong by 36.25%. For R^2, 89% of the variations of the dependent variable are explained by the independent variables of our model. We see that the R^2 is high, indicating a high linear relationship between ProvisionDate and InvoicedValue. Finally, the adjusted R^2 value is 83% of the variability explained by the model, considering the number of independent variables, as shown in Table 10.

With the other approach, when we do the clustering first using the k-means technique with its fifteen groups and then run the LSTM network for each of the clusters as shown in Table 11, we obtain better results. For RMSE, for clusters 0, 2, 3, 7, 8, 9, 10, 12, and 13, they have a better average mean prediction error for each one. MAPE has a lower forecast error for clusters 0, 2, 3, 4, 7, 9, 10, 12, 13, and 14. The R^2 for all clusters indicates a high relationship between the variables InvoicedValue and Date. Finally, the adjusted R^2 value for all clusters has a higher percentage of variability explained by the model. The values for clusters 1, 5, 11, and 14 have a high percentage of adjusted R^2, which can be interpreted as good. However, it shows an RSME as an average prediction error that is high enough to project. Clusters that did not perform as well, e.g., young, HEALTHY, and LittleUse are users with little history, and therefore it is more complex to predict their behavior.

The main implication of our results is that combining the use of the clustering algorithms to identify patient groups with deep learning LSTM networks to predict future costs for these groups enables a more accurate prediction of the costs of patients for healthcare providers.

6. Conclusions

The results demonstrate the feasibility of segmenting the population by cluster (k-means), and finally the LSTM network to project the cost of each group. Having a tool that allows the organization to know the cost for the next month or up to three months allows it to better provision resources. We do not consider it appropriate to project beyond three months because the model may lose reliability. The results obtained show the validity of the initial approach—which remains probabilistic—based on care events, which can be improved with the incorporation of clinical variables.

This first phase estimates the probabilistic projection of costs grouped by population segments to address a second phase of the project, which aims to consolidate a patient-focused cost model based on their medical records in such a way that it allows us not only to predict the potential services and costs related to each patient, but also to identify the potential operational, clinical, and administrative strategies to improve the quality of life of patients, preventing the accelerated development of diseases and/or events that impair their health and consequently provide a better life expectancy and reduce future costs related to these potential events. This approach allows us to help health organizations to be ready for providing healthcare by optimizing costs, giving an accurate diagnosis of diseases, improving service quality by grouping patients, optimizing resources, and improving clinical results [49].

By having more variables for a person, such as demographic variables, the identification of a provisioning event, clinical, diagnostic, and risk variables, and the cost of all the services provided, either with their own infrastructure or third-party infrastructure, the results are more accurate. With all the patient's related variables over time and their cost, it is possible to predict the risks and costs of a person and thus be able to implement survival models.

The goal is to have projected monthly costs, which can be used to assess a chronic patient or a recurring patient and their cost pattern, and model through clusters in cohorts

to provide preventive care, allowing the health system to reduce costs and significantly improve the quality of life of patients.

Author Contributions: The authors contributed equally to this work. Conceptualization, D.R.S.S., J.C.R., J.M.-R., V.G.-D. and E.R.N.-V.; Methodology, D.R.S.S., J.C.R., J.M.-R.,V.G.-D. and E.R.N.-V. Software, D.R.S.S., J.C.R. and J.M.-R.; Validation, D.R.S.S., J.C.R. and J.M.-R.; Formal Analysis, D.R.S.S., J.C.R. and J.M.-R.; Investigation, D.R.S.S., J.C.R. and J.M.-R.; Resources, D.R.S.S., J.C.R. and J.M.-R.; Data Curation, D.R.S.S., J.C.R. and J.M.-R.; Writing—Original Draft Preparation, D.R.S.S., J.C.R.. and J.M.-R.; Writing—Review and Editing, D.R.S.S., J.C.R., J.M.-R., V.G.-D. and E.R.N.-V.; Visualization, D.R.S.S., J.C.R. and J.M.-R.; Supervision, V.G.-D. and E.R.N.-V.; Project Administration, V.G.-D. and E.R.N.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable. Data is deidentified.

Data Availability Statement: The source code, training data, and all other supplementary resources are available online at https://github.com/sandovaldanny/Prediction_Health_Cost (accessed on 8 February 2022). To set up the workspace and repeat the experiments, follow the instructions in the corresponding ReadMe file.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

As indicated by the elbow and silhouette methods, the result of running with 5 clusters is shown, highlighting a slight increase in performance.

Table A1. Result of running with 5 clusters.

Cluster	R^2	Adj. R^2
0	0.91	0.87
1	0.95	0.91
2	0.92	0.82
3	0.98	0.92
4	0.97	0.96

References

1. Yang, C.; Delcher, C.; Shenkman, E.; Ranka, S. Machine Learning Approaches for Predicting High Utilizers in Health Care. In Proceedings of the International Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 26–28 April 2017; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2017; Volume 10209 LNCS, pp. 382–395.
2. Current Health Expenditure (CHE) as Percentage of Gross Domestic Product (GDP) (%). Available online: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/current-health-expenditure-(che)-as-percentage-of-gross-domestic-product-(gdp)-(-) (accessed on 19 January 2022).
3. Morid, M.A.; Sheng, O.R.L.; Kawamoto, K.; Ault, T.; Dorius, J.; Abdelrahman, S. Healthcare Cost Prediction: Leveraging Fine-Grain Temporal Patterns. *J. Biomed. Inform.* **2019**, *91*, 103113. [CrossRef] [PubMed]
4. Sushmita, S.; Newman, S.; Marquardt, J.; Ram, P.; Prasad, V.; de Cock, M.; Teredesai, A. Population Cost Prediction on Public Healthcare Datasets. In Proceedings of the 5th International Conference on Digital Health 2015, Florence, Italy, 18–20 May 2015; ACM International Conference Proceeding Series. Association for Computing Machinery: New York, NY, USA, 2015; Volume 2015, pp. 87–94.
5. Ministerio de Salud y Protección Social $31.8 Billones Para La Salud En 2020. Available online: https://www.minsalud.gov.co/Paginas/31-8-billones-para-la-salud-en-2020.aspx (accessed on 4 January 2022).
6. El Presupuesto de La Nación de 2021 Destinará $75 Billones Para Deuda, 6.7% Del PIB. Available online: https://www.larepublica.co/economia/presupuesto-de-la-nacion-de-2021-destinara-75-billones-para-deuda-67-del-pib-3038167 (accessed on 5 January 2022).
7. About Keralty—Keralty. Available online: https://www.keralty.com/en/web/guest/about-keralty (accessed on 3 May 2021).
8. Giedion, U.; Díaz, B.Y.; Alfonso, E.A.; Savedoff, W.D. The Impact of Subsidized Health Insurance on Access, Utilization and Health Status in Colombia. *Utilization and Health Status in Colombia (May 2007). iHEA 2007 6th World Congress: Explorations in*

Health Economics Paper. 2007, p. 199. Available online: https://www.researchgate.net/publication/228233420_The_Impact_of_Subsidized_Health_Insurance_on_Access_Utilization_and_Health_Status_in_Colombia (accessed on 4 February 2022).
9. Plan Obligatorio de Salud. Available online: https://www.minsalud.gov.co/proteccionsocial/Paginas/pos.aspx (accessed on 8 January 2022).
10. Paho—Health in the Americas—Colombia. Available online: https://www.paho.org/salud-en-las-americas-2017/?p=2342 (accessed on 3 May 2021).
11. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
12. Kaushik, S.; Choudhury, A.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. Using LSTMs for Predicting Patient's Expenditure on Medications. In Proceedings of the 2017 International Conference on Machine Learning and Data Science (MLDS 2017), Noida, India, 14–15 December 2017; pp. 120–127. [CrossRef]
13. Graves, A. Generating Sequences with Recurrent Neural Networks. *arXiv* **2013**, arXiv:1308.0850.
14. Tu, L.; Lv, Y.; Zhang, Y.; Cao, X. Logistics Service Provider Selection Decision Making for Healthcare Industry Based on a Novel Weighted Density-Based Hierarchical Clustering. *Adv. Eng. Inform.* **2021**, *48*, 101301. [CrossRef]
15. Zhang, Z.; Murtagh, F.; van Poucke, S.; Lin, S.; Lan, P. Hierarchical Cluster Analysis in Clinical Research with Heterogeneous Study Population: Highlighting Its Visualization with R. *Ann. Transl. Med.* **2017**, *5*, 75. [CrossRef] [PubMed]
16. Abbi, R.; El-Darzi, E.; Vasilakis, C.; Millard, P. A Gaussian Mixture Model Approach to Grouping Patients According to Their Hospital Length of Stay. In Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems, Jyvaskyla, Finland, 17–19 June 2008; pp. 524–529. [CrossRef]
17. Santos, A.M.; de Carvalho Filho, A.O.; Silva, A.C.; de Paiva, A.C.; Nunes, R.A.; Gattass, M. Automatic Detection of Small Lung Nodules in 3D CT Data Using Gaussian Mixture Models, Tsallis Entropy and SVM. *Eng. Appl. Artif. Intell.* **2014**, *36*, 27–39. [CrossRef]
18. 2.3. Clustering—Scikit-Learn 1.0.2 Documentation. Available online: https://scikit-learn.org/stable/modules/clustering.html (accessed on 24 January 2022).
19. Implementing a K-Means Clustering Algorithm from Scratch | by Zack Murray | the Startup | Medium. Available online: https://medium.com/swlh/implementing-a-k-means-clustering-algorithm-from-scratch-214a417b7fee (accessed on 8 January 2022).
20. K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks | by Imad Dabbura | towards Data Science. Available online: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (accessed on 8 January 2022).
21. Fontalvo-Herrera, T.; Delahoz-Dominguez, E.; Fontalvo, O. Methodology of Classification, Forecast and Prediction of Healthcare Providers Accredited in High Quality in Colombia. *Int. J. Product. Qual. Manag.* **2021**, *33*, 1–20. [CrossRef]
22. Kaushik, S.; Choudhury, A.; Sheron, P.K.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Front. Big Data* **2020**, *3*, 4. [CrossRef]
23. Kabir, S.B.; Shuvo, S.S.; Ahmed, H.U. Use of Machine Learning for Long Term Planning and Cost Minimization in Healthcare Management. *medRxiv* **2021**. [CrossRef]
24. Scheuer, C.; Boot, E.; Carse, N.; Clardy, A.; Gallagher, J.; Heck, S.; Marron, S.; Martinez-Alvarez, L.; Masarykova, D.; Mcmillan, P.; et al. Predicting Utilization of Healthcare Services from Individual Disease Trajectories Using RNNs with Multi-Headed Attention. *Proc. Mach. Learn. Res.* **2020**, *116*, 93–111. [CrossRef]
25. Elbattah, M.; Molloy, O. Data-Driven Patient Segmentation Using K-Means Clustering: The Case of Hip Fracture Care in Ireland. *ACM Int. Conf. Proc. Ser.* **2017**, 1–8. [CrossRef]
26. Nedyalkova, M.; Madurga, S.; Simeonov, V. Combinatorial K-Means Clustering as a Machine Learning Tool Applied to Diabetes Mellitus Type 2. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1919. [CrossRef] [PubMed]
27. Salud—SONDA. Available online: https://www.sonda.com/industrias/salud/ (accessed on 4 January 2022).
28. Kotsiantis, S.B.; Kanellopoulos, D.; Pintelas, P.E. Data Preprocessing for Supervised Leaning. *Int. J. Comput. Inf. Eng.* **2007**, *1*, 4104–4109. [CrossRef]
29. Keras: The Python Deep Learning API. Available online: https://keras.io/ (accessed on 1 February 2022).
30. Keras | TensorFlow Core. Available online: https://www.tensorflow.org/guide/keras?hl=es-419 (accessed on 1 February 2022).
31. Pedregosa FABIANPEDREGOSA, F.; Michel, V.; Grisel OLIVIERGRISEL, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
32. Welcome to Python.org. Available online: https://www.python.org/ (accessed on 19 January 2022).
33. Streamlit • The Fastest Way to Build and Share Data Apps. Available online: https://streamlit.io/ (accessed on 8 January 2022).
34. Google Introducción a AI Platform | AI Platform | Google Cloud. Available online: https://cloud.google.com/ai-platform/docs/technical-overview?hl=es-419 (accessed on 4 January 2022).
35. Shiranthika, C.; Shyalika, C.; Premakumara, N.; Samani, H.; Yang, C.-Y.; Chiu, H.-L. Human Activity Recognition Using CNN & LSTM. Available online: https://www.researchgate.net/publication/348658435_Human_Activity_Recognition_Using_CNN_LSTM (accessed on 17 January 2022).
36. Illustration of an LSTM Memory Cell. | Download Scientific Diagram. Available online: https://www.researchgate.net/figure/Illustration-of-an-LSTM-memory-cell-7_fig1_348658435 (accessed on 19 January 2022).

37. Choosing the Right Hyperparameters for a Simple LSTM Using Keras | by Karsten Eckhardt | towards Data Science. Available online: https://towardsdatascience.com/choosing-the-right-hyperparameters-for-a-simple-lstm-using-keras-f8e9ed76f046 (accessed on 19 January 2022).
38. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2014.
39. Metrics. Available online: https://keras.io/api/metrics/ (accessed on 15 January 2022).
40. Nielsen, A. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*; O'Reilly Media: Sebastopol, CA, USA, 2019; p. 480.
41. K-Means Clustering from Scratch in Python | by Pavan Kalyan Urandur | Machine Learning Algorithms from Scratch | Medium. Available online: https://medium.com/machine-learning-algorithms-from-scratch/k-means-clustering-from-scratch-in-python-1675d38eee42 (accessed on 1 March 2022).
42. Umargono, E.; Suseno, J.E.; Vincensius Gunawan, S.K. K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. In Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019), Yogyakarta, Indonesia, 25 November 2019. [CrossRef]
43. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning-Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009; p. 282.
44. Willmott, C.J.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
45. Forecast KPI: RMSE, MAE, MAPE & Bias | towards Data Science. Available online: https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d (accessed on 4 March 2022).
46. Why Not MSE or RMSE A Good Enough Metrics for Regression? All about R^2 and Adjusted R^2 | by Neha Kushwaha | Analytics Vidhya | Medium. Available online: https://medium.com/analytics-vidhya/why-not-mse-or-rmse-a-good-metrics-for-regression-all-about-r%C2%B2-and-adjusted-r%C2%B2-4f370ebbbe27 (accessed on 2 March 2022).
47. How Do You Check the Quality of Your Regression Model in Python? | by Tirthajyoti Sarkar | towards Data Science. Available online: https://towardsdatascience.com/how-do-you-check-the-quality-of-your-regression-model-in-python-fa61759ff685 (accessed on 8 January 2022).
48. What Does RMSE Really Mean? | by James Moody | towards Data Science. Available online: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e (accessed on 8 January 2022).
49. Muniasamy, A.; Tabassam, S.; Hussain, M.A.; Sultana, H.; Muniasamy, V.; Bhatnagar, R. Deep Learning for Predictive Analytics in Healthcare. In Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications, Jaipur, India, 13–15 February 2020; Springer: Cham, Switzerland, 2020; Volume 921, pp. 32–42.