

A Neural Network Approach to High Cost Patients Detection

Franklin Messias Barbosa(✉) and Renato Porfirio Ishii(✉)

Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil
franklin.barbosa@aluno.ufms.br, renato.ishii@ufms.br

Abstract. The growing aging of the world's population along with several environmental, social and economic factors, end up posing major challenges for public health. One challenge is the detection and treatment of high cost patients, i.e., a small percentage of patients associated with majority of expenditures in healthcare. The early detection of patients who may become high cost in the future can be used to better target interventions focusing on preventing their transition or, in the case of those who are already in such condition, to allow appropriate approaches, rather than generic ones. In both cases, the detection of such patients can be beneficial, reducing avoidable costs and improving patients' condition. In order to make such detection, this work has focused on using deep learning techniques, specifically, Neural Networks, along with a dataset composed of survey answers applied by the United States government, called Medical Expenditure Panel Survey (MEPS) and attributes gathered from the literature. For the purposes of this work, 11 years of the MEPS dataset were considered, including the years from 2006 to 2016. The models created have shown results ranging between 83% to 90% on metrics such as accuracy, precision, recall, specificity and f1-score. This work also aimed to make the creation and testing of such networks easier, by providing the tools developed during its evolution on GitHub.

Keywords: Machine learning · Deep learning · Big data · Classification · Smart cities

1 Introduction

Public health is an area of great importance both globally and nationally, which also makes it important to efficiently manage the resources available in this area. One of the challenges in this regard, which can be observed in both public and private contexts, are high cost patients, i.e. a small portion of users who represent a large share of the total expenditure. Such a phenomenon can be observed not only in Brazil [10], but also in several other countries, such as Canada [14] and USA [11].

In a report published in 2013, the Kaiser Family Foundation, an American non-profit organization that works on public health issues in the USA, points out

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 527–540, 2021.
https://doi.org/10.1007/978-3-030-86970-0_37

that in Medicaid [8], public health plan in the United States, 5% of the highest cost beneficiaries at the time accounted for 54% of total expenses [11].

Thus, a possible approach would be to identify such patients while they still have average costs, that is, before they become high cost, as pointed out by Chechulin [6]. Such identification may be useful as a guide to targeted interventions, in order to prevent the transition of the patient, preventing the condition complications already present and preventing the emergence of new problems, when possible, thus improving not only the efficiency of the agents responsible for healthcare, but also preventing avoidable expenses.

Similarly, the identification of patients who are already high cost is also beneficial, as stated by Blumenthal [5], high cost and high need patients are usually people that even receiving substantial services, have critical health needs that are not attended to. This portion of the population often receives inefficient care, such as unnecessary hospitalizations. According to Blumenthal [5], by giving high priority to the care of these individuals, the resources can be used where they are likely to produce better results, with lower costs.

1.1 Problem Definition

Given the current economic scenario, in addition to factors such as population growth and aging, adequate resource management is becoming increasingly important, especially in critical areas such as health. In order to assist decision making and, consequently, optimize efforts and resources, adequate data processing technologies are essential.

In this scenario, the use of machine learning algorithms provides an improvement in the detection of high cost patients compared to the traditional methods used previously. A possible alternative is to use Deep Learning approaches to deal with this problem. As this technique ends up discovering previously unknown interrelationships in the data sets in which they are applied, our hypothesis is that its use may present an improvement in the results found in previous works.

1.2 Objectives

The objectives of this work include the implementation of deep learning models for the high cost patient prediction problem, as well as to compare the results with those obtained in other previous studies, the development of the necessary tools for creating such models in order to simplify future projects in this area. And, as a final objective, to validate the set of attributes defined by Shenas [15] in a larger and more current dataset.

In order to achieve these objectives, different datasets were initially considered and after analyzing their contents, availability, ease of access and amount of data, the MEPS dataset was chosen for this work, more specifically, 11 years were considered, including the years of 2006 to 2016.

This paper is organized as follows: Sect. 2 presents brief theoretical reference, Sect. 2.1 expands on how the MEPS dataset is structured and what information

can be obtained from it. Section 2.2 explains the preprocessing steps taken to build the final dataset. Section 2.3 presents an explanation on how the preprocessing steps were implemented. Finally, on Sect. 2.4, the metrics used to evaluate the constructed model are presented and explained. Section 3 summarises the results obtained from the constructed models, and Sect. 4 the results and contributions that were derived from this work.

2 Materials and Methods

2.1 Dataset

The Medical Expenditure Panel Survey, or MEPS, "is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. MEPS is the most complete data source on the cost and the use of health care and health insurance coverage" [2]. The results derived from these surveys are periodically released as public domain data, and aim to present information on which health services are used more frequently, how frequently they are used, what are the associated costs, how they are paid for. In addition, to several other information that may be important for research in this area, including, mainly, information at the individual level. Concerning patient privacy, according to AHRQ (Agency for Healthcare Research and Quality), all information is collected with assurances of confidentiality, and are to be used for research purposes only, with unauthorized disclosure of survey data being subject to substantial fines [3].

The MEPS dataset is divided by years, and further divided by panels, with each year, a new panel is started, that panel will be concluded in the following year, meaning that a full panel contains data for each individual during a two years period. These data include information on several topics defined by the AHRQ as follows [4]:

- Unique person identifiers and survey administration variables, which contain information related to conducting the interview, the household and family composition and person level status code;
- Geographic Variables, which indicate the census region for the individual;
- Demographic Variables, which include information about the demographic characteristics of each person, such as age, sex, race, ethnicity, marital status, educational attainment and military service;
- Income and Tax Filing Variables provide income and tax-related information, such as poverty status and number of dependents;
- Person-Level Condition Variables, which refers to the individual's physical and mental health status, including several disease diagnostics, such as angina, bronchitis, high cholesterol, among others. In addition, it includes information on how old the individual was when the condition was diagnosed, if the condition persists or if it was recently treated.
  Information on the diagnostics of rare diseases, more specifically, for types of cancer that appear on the list of rare diseases of the National Institutes of

Health, such as cancer of the esophagus, pancreas, larynx, leukemia, among others are presented in the same way, in order to protect the anonymity of the interviewees;
- Health Status Variables include information about the individual's physical conditioning, addressing issues such as the difficulty of lifting 10 pounds, climbing 10 steps, walking 3 blocks, walking 1 mile, standing up for 20 min, and other activities that are common in everyday life, such as crouching, or picking up elevated objects, in addition to other physical limitations, such as low vision or hearing. It also indicates if there are any social or cognitive limitations, as well as behavioral factors. Finally, information about the clinical history is included in this class of variables, indicating the time since certain procedures were performed, including mammography, colonoscopy and hysterectomy, in addition to others;
- Disability Days Variables, which contains questions about time lost from work or school because of a physical illness or injury, or a mental or emotional problem or days lost treating other people with such conditions;
- Access to Care Variables carry information on basic access to healthcare, what is the provider and if they find any difficulty when needing such services;
- Employment Variables consist of person-level indicators such as employment status, and hourly wage and occupation area;
- Health Insurance Variables indicate which health insurance the interviewee has, when applicable, and information on the history of plans used;
- Finally, utilization, expenditure, and source of payment variables contain information such as the number of times that each health service was used, such as outpatient visits, dental services, prescription drugs, emergency care, among others. The total cost of the procedures associated with the interviewee and how they were paid for. For the problem of high cost patients, these variables can be used to define the classes of each individual in the training set.

2.2 Data Preprocessing

The data preprocessing phase consists of a set of techniques and activities on the chosen dataset. This process is intended to prevent a series of problems that may arise in the future due to the state in which the data can be received at first, in addition to ensuring that a greater amount of relevant information can be extracted from the original dataset, contributing to the quality of the proposed models.

For the purposes of this work, the processes described by Shenas [15] were followed, in a work that defined the minimal set of attributes on the MEPS dataset. More specifically, only individuals older than 17 were considered in his work, those that have non negative record for their expenditure (values for neither years are missing) and those with a positive person-level weight, since records with a zero person-level weight do not represent any part of the population and should not be considered in studies conducted on the US household population [15]. In total, 39 attributes were chosen, divided in 5 categories:

1. Demographics
   - Age, Sex, Race, HIDEG (highest degree of education), Region, Marry and POVCAT (family income as percent of poverty line);
2. Health Status
   - RTHLTH (perceived health status), MNHLTH (perceived mental health status), ANYLIM (any limitation on daily activities), and BMINDX (body mass index);
3. Preventive Care
   - Check (routine checkup), BPCHEK (blood pressure check), CHOLCK (cholesterol check), NOFAT (if a doctor advised a lower cholesterol diet), EXRCIS (if a doctor has advised more exercise), ASPRIN (frequent aspirin use), BOWEL (sigmoidoscopy or colonoscopy), STOOL (blood stool test), and DENTCK (frequency of dental checkup);
4. Priority Conditions
   - HIBPDX (high blood pressure), CHDDX (coronary heart disease), ANGIDX (angina), MIDX (myocardial infarction), OHRTDX (other heart disease), STRKDX (stroke), ASTHDX (asthma), EMPHDX (emphysema), CANCERDX (cancer), DIABDX (diabetes), ARTHDX (arthritis), CHOLDX (high cholesterol), PC (presence of any priority condition), and PCCOUNT (number of conditions present);
5. Visits Counts
   - OBTOT (office based visits), OPTOT (outpatient visits), ERTOT (emergency room visits), IPDIS (inpatient hospital stays), and RXTOT (number of prescription medicines).

These attributes were also preprocessed according to Shenas' [15] results, mainly, several reductions in the number of possible categories were proposed by them, these reductions include joining similar categories, time periods or normalizing values, in order to reduce the number of possible values for each attribute, making their interpretation easier [15], those changes are detailed below.

"Age" was divided in 4 categories, ages in the 18–49 range were replaced by 1, 50–65 by 2 and those older than 65 by 3. In addition, all records with an age value lesser than 18 are removed, since many of the other attributes are not applicable in these cases. "Race", originally a categorical variable with 6 possible values was reduced to 3 categories, with values of 1, 2 or 3 for whites, blacks and others, respectively. "HIDEG" was changed from a categorical variable that represented the Highest Degree of Education (high school, bachelor's, master's, doctorate, etc.) to a binary flag, with 1 for having a higher education, and 0 for lacking it.

The Marry variable was reduced from its original 10 possible values to a 0–2 range, with 0 being "never married", 1 for "married" and 2 for "widowed", "separated" or "divorced", this was done by excluding the category for underage people, since all minors were excluded from the final dataset, by joining the widowed, separated and divorced classes. The Marry attribute also had classes for people who married, widowed or separated in the current round, those were also converted to the new value.

Finally, POVCAT, represents the family's income as percent of poverty line. Originally, this attribute had 5 possible values, ranging from 1–5 and representing "poor/negative", "near poor", "low income", "middle income" and "high income", respectively. In the final dataset, these values were reduced to a 1–3 range, with 1 representing "poor/negative/near poor", 2 representing "low/middle" and 3 "high income". The remaining attributes from the demographics category were left unchanged.

From the Health Status category, aside from BMINDX, the body mass indicator, all attributes are left unchanged. BMINDX originally contained a continuous value and was changed to an ordinal attribute on the range of 1–4, representing "underweight", "normal weight", "overweight" and "obese". RTHLTH and MNHLTH represent the individual's self perceived physical and mental health status, in a 1–5 range, representing "excellent", "very good", "good", "fair" and "poor" health. Finally, ANYLIM is a binary attribute indicating if the individual has any limitations in activities of daily living, such as eating or dressing or limitations in instrumental activities of daily living, such as housework or shopping. The preventive care category contains attributes that indicate the time since the individual's last routine checkup (Check), blood pressure test (BPCHEK), cholesterol check (CHOLCK), if a doctor has advised the person to eat fewer high cholesterol foods (NOFAT), if a doctor has advised the individual to exercise more (EXRCIS), if the person takes aspirin frequently (ASPRIN), if the person ever had a sigmoidoscopy or colonoscopy (BOWEL), if a blood stool test was ever necessary (STOOL) and frequency of dental check-up (DENTCK). All attributes on this category are either binary, in which case were left unchanged, or follow a similar pattern for values, ranging from 1–6, where each value represents the number of years since said event or exam occurred, exceptions being 5, which represents 5 or more years and 6, representing never. These values were reduced to a 0–2 range, where 0 represents never, 1 representing said event occurred within the past year and 2 for 2 years or more. Additionally, DENTCK was changed to a 0–2 range, where 0 represents never, 1 being twice a year or more and 2 less than twice a year.

The priority conditions category contains information on diagnostics of several health conditions for each individual, such as high blood pressure (HIBPDX), coronary heart disease (CHDDX), angina or angina pectoris (ANGIDX), heart attack, or myocardial infarction (MIDX), other heart diseases (OHRTDX), stroke (STRKDX), asthma (ASTHDX), emphysema (EMPHDX), cancer (CANCERDX), diabetes (DIABDX), arthritis (ARTHDX) and high cholesterol (CHOLDX). Since all attributes are binary, all were left unchanged. The PC and PCCOUNT attributes are not present in the original MEPS dataset, being derived from the previous attributes, PC is a binary attribute that indicates if the individual has any of the listed priority conditions and PCCOUNT accepts values in a 0–4 range, where 0–3 is the number of conditions and 4 represents 4 or more conditions.

Finally, the visits counts category indicate the number of visits the individual has had to several medical providers, such as office based (OBTOT),

hospital outpatient (OPTOT), emergency room (ERTOT), inpatient hospital stays (IPDIS) and the number of prescription medicines (RXTOT). All attributes in this category were normalized using minmax scaling, which changes every value to a 0–1 range using the equation described in 1, where x' is the new value for the attribute, $x_{\min}$ is the minimal value found for that attribute, and similarly, $x_{\max}$ is its maximum value.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

2.3 Preprocessing Implementation

As one of this work's objectives, the whole preprocessing and data preparation steps defined by Shenas [15] and explained on the previous section were automated for future studies using the MEPS dataset, this process was divided in several steps, detailed below, allowing easier understanding, accessibility and eventual modification of the source code. The only prerequisites for using this implementation are a working Python 3 environment and the required MEPS data files, in csv format.

1. The first step consists on applying the conditions mentioned before to the original full year dataset. This step excludes records with a personal level weight of 0, those with an age lesser than 18 and those with a missing value for total expenditure. Since at this point we have the full dataset, adding or removing restrictions can easily be done by editing this file;
2. The second step is used only for the 2006 and 2007 full year files. Prior to 2008, the cancer diagnostic attribute was not present in the full year file, being instead found in a separate medical conditions file. This step searches the medical condition file and appends the cancer diagnostic attribute to the main file. Files from 2008 onwards can skip this step;
3. The third step creates the PC and PCCOUNT attributes, by counting how many priority conditions are present on each record. Priority conditions can be easily changed on this step, by simply adding or removing their names from the list;
4. The fourth step retrieves the second year expenditure for each individual in the main file. Records without a second year expenditure value are excluded;
5. The fifth step creates a HIGHCOST attribute for each individual, at this point the user can choose whether to consider the first or second year expenses to determine the high cost population, in addition to the percentage of the dataset that will be considered high cost. In other words, the user can choose to consider the top 5% as the high cost population considering their second year expenses, in this scenario, 5% of the dataset, those with the highest values for their total expenditure on year 2, will have the HIGHCOST attribute set to true, while the remaining will have it as false;
6. The sixth step applies changes to the values of the records, applying minmax normalization, or changing value intervals, according to the specifications

listed above. Over the years, several fields in the MEPS dataset have had their names, or even values, changed. HIDEG, for instance, the field representing the highest degree of education an individual has achieved, is not present in the 2014 and 2013 full year files, instead, we use a different field, called EDUYRDG (years of education or highest degree) which provides similar information. In this step we take such changes into account, ensuring that each of the 39 variables described above are present;

7. The seventh step deals with missing or invalid values in each of the relevant fields in the dataset. By default, MEPS uses certain values for questions the individual has failed to provide an answer for, for instance, if they do not know the answer to a question, the value is set as -8, if they refused to answer, the value is set as -7, and so on for various other reasons. For our purposes, such differences are not relevant, as such, for every field where such notation is used, every value under 0 is replaced by -1, this is done so that all invalid values are equal, preventing invalid assumptions to be made based on the original MEPS value notation. Furthermore, this also allows for an easier identification of invalid values, in case the user wishes to exclude them from the final dataset or impute them by using other methods;
8. The eighth step can be further divided into 4 parts, in short, it allows the user to add fields that were not included in the original 39, and, depending on the format the data is in, either normalize it, reduce its range (for the variables that refer to time since last event) and remove negative values as mentioned above. This allows the user to easily change the fields in the final dataset, while applying the necessary data preprocessing. At the end of this step all fields not used are removed from the dataset;
9. The ninth step handles the final dataset, in it, we divide the dataset into high cost and low cost, based on the variable created previously, in order to keep the dataset balanced, we count the number of records in the high cost dataset, and randomly pick the same number of records in the low cost dataset. Since the low cost and high cost datasets are not deleted in the process, this can be repeated several times, each time creating a new low cost dataset with randomized records.

2.4 Model Evaluation

Assuming any binary classifier that takes an example and return as a hypothesis one between two possible classes, “A” and “B”, for example, any prediction returned by that classifier will fit into one of four possible scenarios: the model predicted class correctly predicted “A” for the example; the model predicted class “A”, but the example belonged to class “B”; the model returned “B” as a prediction and the example was class “A”; or, finally, the model returned “B” and the example really belonged to the “B” class. Viewing the classes “A” and “B” as positive and negative, the cases can be represented with matrix 1, where the first line represents the prediction returned by the classifier, and the first column represents the true class of the example. This table is called the confusion matrix [9].

Table 1. Example of a confusion matrix

	A	B
A	True Positive	False Negative
B	False Positive	True Negative

After constructing a model, at the end of the testing process, the matrix can be filled, according to the results found, and from it some metrics are calculated that can be useful when assessing the quality of the constructed model. The first metric is accuracy, which represents the rate at which the model was able to correctly predict the class of any given example, and is defined in 2.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Number\ of\ Examples} \tag{2}$$

Accuracy presents a valid metric for the analysis of models, however, it is not reliable in certain cases. More specifically, it does not represent well the predicting capacity of a model whose dataset is unbalanced [16]. In other words, if most of the examples in the test set are from the same class, we can have a model that classifies all entries as being in the majority class. As most examples will be predicted correctly, the accuracy of this model will be high, even if it is not actually making any kind of prediction itself. Thus, there are other metrics, similar to accuracy, that are less sensitive to this type of scenario and can be used in conjunction with it to better represent the capacity of the evaluated model.

One of these metrics is the recall, also referred to as sensitivity, which measures the hit rate of the positive class. In other words, of all the examples that belong to the positive class, the recall indicates how many were classified correctly. It can be calculated using Eq. 3. A model that does not produce false negatives has a value of 1. A third metric that can be considered is precision, defined by Eq. 4, which indicates the proportion of predictions for the positive class that were correct. A model that does not produce false positives has a precision of 1. Finally, the rate of false positives is similar to recall, and is defined by Eq. 5 [9,13].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{4}$$

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative} \tag{5}$$

From these three metrics, a fourth can be derived, the f1 score, which is defined as the harmonic mean between precision and the recall and is calculated according to Eq. 6. It gives equal weight to both measures, so the f1 score of a

model is high if both recall and accuracy are high. On the other hand, if either one decreases, the final f1 score will also be lower. This metric is important for the evaluation of models where both metrics have similar importance, that is, in problems where it is not advantageous to increase either of them at the expense of the other [13].

$$f1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

3 Results

The resulting dataset was used to train several neural networks using Keras [7], 11 years (2006–2016) of data from the MEPS dataset were used, resulting in a dataset with 125,457 total records. We have considered the top 5% of the population as high cost, as such, the dataset can be divided into high cost, containing 6,280 records, and low cost, containing 119,177. Since the classes are clearly unbalanced, we follow the steps detailed in the previous section, keeping the 6,280 original high cost individuals and randomly selecting the same quantity of low cost records, keeping the dataset balanced. During training, the resulting dataset is divided randomly between training and validation sets, where the training set contains 80% of the records, while the validation set contains the remaining 20%.

Each model was trained separately 100 times, each using different low cost records, the results for each network were logged, and an average result for each metric was calculated using the results found in each iteration. This process was done twice, once using the same year expenditure as a parameter to define high cost individuals, and then using the second year's expenses.

The models created were all fully connected, sequential networks. Different networks created had a different number of layers, number of nodes in each layer, different activation functions, number of dropout layers and dropout probability.

The model used to generate the results shown on Tables 2 and 3 contained 3 hidden layers, with 32, 64 and 32 nodes respectively, with a dropout layer between layers 1 and 2, and another, between layers 2 and 3, the dropout rate for both of those layers was set to 0.5. A final dropout layer was added between the third and output layers, with a dropout rate of 0.1. The activation function used on the layers was the Rectified Linear Unit (relu), for the hidden layers, and the hyperbolic tangent function (tanh) for the output layer. When compiling the model, the RMSprop algorithm was used as the optimizer, the loss function used was binary cross entropy and the metric accuracy. Finally, the model was trained for 100 epochs, with a batch size of 32.

For the same year prediction, the metrics for the best performing model are detailed on Table 2. For models trained to predict high cost individuals of the following year, the best results were detailed on Table 3.

All results are similar to the ones obtained by neural networks constructed by Shenas [15], shown on Table 4 except for accuracy, which was slightly higher. This corroborates the minimal set of attributes defined in that work, in addition to

Table 2. Metrics for our model when using the top 5% expenditures as target for the same year

	Value	Standard deviation
Sensitivity	90%	0.018
Specificity	83%	0.02
Accuracy	86%	0.007
Precision	84%	0.01
F1 score	87%	0.006

Table 3. Metrics for our model when using the top 5% expenditures as target for the following year

	Value	Standard deviation
Sensitivity	91%	0.02
Specificity	84%	0.03
Accuracy	88%	0.009
Precision	86%	0.02
F1 score	88%	0.008

suggesting that similar or better results can be achieved with a greater quantity or more recent data.

Table 4. Metrics presented by Shenas [15] when using the top 5% expenditures as target using decision trees (C5.0 and CHAID) and Neural Networks (NN)

	C 5.0	CHAID	NN
Sensitivity	56%	**90**%	87%
Specificity	**96**%	86%	86%
Accuracy	**96**%	86%	86%

As a second point of comparison we can point the results presented by Meehan [12], shown on Table 5, who used traditional machine learning algorithms to generate models for the same problem.

A final point of comparison can be made by observing the confusion matrices generated by both our work and the ones reported by Shenas, shown on Tables 6, 7 and 8. On all three Tables, we compare the total number of predictions for each case, similar as what can be observed on Table 1. On the top headers, "Predicted HC" means the model predicted an individual as a high cost patient, while "Predicted LC" means the model predicted the individual as not being high cost, similarly, the headers on the left show the real class for the

Table 5. Results presented by Meehan [12]

	Logistic regression	Naive Bayes	J48
Sensitivity	74%	**74.10**%	67.20%
Specificity	77.10%	76.90%	**82.20**%
Accuracy	**75.55**%	75.49%	74.73%
F-Measure	**75.50**%	**75.50**%	74.60%

predictions, with "Actual HC" meaning the individual was a high cost patient, and "Actual LC" the opposite. This way, it is possible to grasp how many wrong classifications each model has made for each of the possible error types (classifying a high cost as a non high cost or classifying a non high cost as high cost), as well as how many predictions were correctly made.

Table 6. Confusion matrix presented by Shenas [15] when using the top 5% expenditures as target (Neural Network)

	Predicted HC	Predicted LC
Actual HC	405	56
Actual LC	891	6884

Table 7. Confusion matrix presented by Shenas [15] when using the top 5% expenditures as target (CHAID)

	Predicted HC	Predicted LC
Actual HC	460	52
Actual LC	1262	7790

Table 8. Confusion matrix generated by one of our tests, when using the top 5% expenditures as target

	Predicted HC	Predicted LC
Actual HC	1160	229
Actual LC	100	1023

The matrices show that, although the previously mentioned metrics were similar, the proportion of cases where the model returned a high cost prediction for a low cost patient was higher on Shenas' model. For the neural network model this puts precision at around 31%, with the CHAID model at 26%, while the average precision for the models generated in this work is over 80%.

4 Conclusion and Contributions

The main conclusions drawn from this work include the validation of the previously defined set of attributes using more data from a wider period and the results obtained in the high cost patient prediction problem using neural networks, both conclusions indicate that neural networks are a valid approach to this problem, especially if more data is available in the future.

With proper data collection from multiple sources, both geographical, as in data collected from several countries, as well as technological, with the use of sensors and other technologies, the scope of this work could certainly be expanded to other countries and scenarios. Since the minimal set of attributes defined by Shenas [15] has shown good results even when expanding the time period from three to eleven years, these attributes are a strong starting point for this expansion, which could motivate relevant healthcare agents to start their own datasets, allowing for richer dataset to be analyzed, which, in turn could lead to both better predictions for high cost individuals, along with the associated benefits mentioned previously, as well as an increase in resource efficiency.

Our contributions include the creation of the necessary tools to preprocess MEPS dataset files, tested in 11 years of data, from 2006 to 2016. These tools allow the user to create models without the need to handle the raw data and year to year differences in the MEPS dataset, such as attributes or value changes, and are available on github [1]. As well as the previously mentioned conclusions, which could aid in future research concerning the high cost patient prediction problem.

Acknowledgments. The InterSCity project is sponsored by the INCT of the Future Internet for Smart Cities funded by CNPq, proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and FAPESP, proc. 2014/50937-1 and 2015/24485-9. This paper is based upon work supported by Fundect, Programa de Educação Tutorial (PET/MEC/FNDE), and Federal University of Mato Grosso do Sul (UFMS).

References

1. Github repository. https://github.com/franklin-ll/Highcost-MEPS. Accessed 6 Apr 2021
2. AHRQ: Agency for healthcare research and quality, about MEPS. https://www.meps.ahrq.gov/mepsweb/. Accessed 6 Apr 2021
3. AHRQ: Agency for healthcare research and quality, confidentiality of MEPS data. https://meps.ahrq.gov/communication/participants/confidentialitymcp.shtml. Accessed 28 June 2021
4. AHRQ: Agency for healthcare research and quality, MEPS data file types. https://bit.ly/3AaBXow. Accessed 6 Apr 2021
5. Blumenthal, D., Anderson, G., Burke, S., Fulmer, T., Jha, A.K., Long, P.: Tailoring complex-care management, coordination, and integration for high-need, high-cost patients a vital direction for health and health care. Documentos/National Academy of Medicine, Washington, DC, vol. 9, pp. 1–11 (2016). https://bit.ly/2UkH5FR

6. Chechulin, Y., Nazerian, A., Rais, S., Malikov, K.: Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). Healthc. Policy **9**, 68–81 (2014). www.longwoods.com/content/23710
7. Chollet, F., et al.: Keras (2015). https://keras.io/. Accessed 6 Apr 2021
8. CMCS: The center for medicaid and CHIP services. www.medicaid.gov/about-us/organization/index.html
9. Google: Machine learning crash course. https://developers.google.com/machine-learning. Accessed 6 Apr 2021
10. Kanamura, A.H., Viana, A.L.D.: Gastos elevados em plano privado de saúde: com quem e em quê. Documentos/Scielo **41**, 814–820 (2007). https://www.scielosp.org/article/rsp/2007.v41n5/814-820/
11. Kanamura, A.H., Viana, A.L.D.: The Kaiser commission on medicaid and the uninsured. Documentos/Kaiser Family Foundation **1**, 27 (2013). https://kaiserfamilyfoundation.files.wordpress.com/2010/06/7334-05.pdf
12. Meehan, J., Chou, C.A., Khasawneh, M.T.: Predictive modeling and analysis of high-cost patients. In: Industrial and Systems Engineering Research Conference (2015)
13. Powers, D.M.W.: Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. School of Informatics and Engineering, Flinders University of South Australia (2007). https://bit.ly/3w3XcoK
14. Rashidi, B., Kobewka, D.M., Campbell, D.J.T., Forster, A.J., Ronksley, P.E.: Clinical factors contributing to high cost hospitalizations in a Canadian tertiary care centre. BMC Health Serv. Res. **1**, 27 (2017). https://bmchealthservres.biomedcentral.com/track/pdf/10.1186/s12913-017-2746-6
15. Shenas, S.A.I.: Predicting high-cost patients in general population using data mining techniques (2012). https://doi.org/10.20381/ruor-6153
16. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002. https://www.sciencedirect.com/science/article/pii/S0306457309000259