1. Introduction

The fast advancing of analysis allows us to investigate problems that present more and more complexities. One of the many fields of application of such techniques is Healthcare. Optimizing the allocation of monetary resources is a major issue for public institutions. In the US, only 5% of the population qualifies as ‘high cost’ patients yet they account for 50% of the total annual healthcare spending[1]. In this paper, we try to predict whether a given patient will end up being a ‘high cost patient’ for the public healthcare system deploying machine learning models.

2. Dataset and Descriptive Analysis

2.1. Data Overlook

The Medical Expenditure Panel Survey, which began in 1996, is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers across the United States. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. We chose to use data from surveys performed during 2003. The dataset is composed by a total of 2000 rows and 26 attributes, both numerical and categorical. Going into detail, we propose a comprehensive list of all attributes and their respective description:

AGE: Age of the patient

ANYLIMIT: 1 if the patient suffers from any disability, 0 otherwise

COLLEGE: 1 if the patient graduated from college, 0 otherwise

HIGHSCH: 1 if the patient possesses a high school diploma, 0 otherwise

GENDER: 1 if the patient is a woman, 0 otherwise

MNHPOOR: self-assessment about mental health. 1 = weak, 0 = strong

INSURE: 1 if the patient was covered by an insurance policy in 1996, 0 otherwise.

USC: 1 if the patient is not satisfied with the standard treatment received, 0 otherwise.

UNEMPLOY: 1 if unemployed, 0 otherwise.

MANAGEDCARE: 1 if the patient has a family doctor.

FAMSIZE: number of family components.

COUNTOP: number of ambulatorial performances requested during the previous last year.

EXPENDOP: total amount of money spent for ambulatorial visits during the previous year.

RACE: patient’s race (Asian, Black, Native, White, other).

RACE1: like RACE but discretized as 1 if Asian, 2 if Black, 3 if Native, 4 if White 0 if other.

REGION: region where the patient lives (WEST, NORTHEAST, MIDWEST, SOUTH).

REGION1: like REGION but discretized as 0 if WEST, 1 if NORTHEAST, 2 if MIDWEST, 3 if SOUTH.

EDUC: level of education: lower(LHIGHSC), higher (HIGHSCH), college (COLLEGE).

EDUC1: like EDUC but 0 if lower, 1 if higher, 2 if college level.

MARISTAT: marital state (NEVMAR, MARRIED, WIDOWED and DIVSEP).

MARISTAT1: like MARISTAT but discretized as 0 if NEVMAR, 1 if MARRIED, 2 if WIDOWED and 3 if DIVSEP.

INCOME: annual income, categorized as: POOR, NPOOR, LINCOME, MINCOME HINCOME.

INCOME1: like INCOME but discretized as 0,1,2 and 3.

PHSTAT: self-assessment of physical health conditions (EXCE, VGOO, GOOD, FAIR e POOR).

PHSTAT1: like PHSTAT but discretized as 0, 1, 2, 3 and 4.

INDUSCLASS: type of industry the patient works in (FINANCE, LEISURE, MANUFACT, MILITARY, MINCONST, NATRESOURCE, OTHERSERV, PROFSERV, PUBADMIN, SALES, TRANSINFO).

2.2. Preliminary Descriptive Statistics

This Medical Expenditure dataset is the result of a cross-sectional survey. We conducted a brief descriptive analysis in order to better understand the data structure and gain more insight about the patterns within the data.

We discovered that even though the age spans from 18 to 65, the dataset is comprised mostly of married and employed white men in their forties, who gave a mostly positive self-assessment of their own mental and physical health and whose family is relatively big. Most of the subjects did not account for any ambulatorial visit during the previous year. For further statistics on the sample see Fig.1in the Appendix.

3. Pre – Processing

3.1. Data cleaning and Feature Engineering

Even though the dataset did not present critical issues concerning data quality, we had to perform some basic data cleaning and feature engineering operations in order to feed our algorithms with the correct input data.

First, we noticed how the missing values from the attribute “INDUSCLASS” were still considered as string type data by the software, so we converted them into proper missing value type data as recognized by Knime.

After that, we proceeded into the binning of the attribute “EXPENDOP” (an originally continuous attribute that indicates the total amount of money spent for ambulatorial visits during the previous year). We adopted the Sample Quantiles binning method set up for 4 quantiles (0.25, 0.50, 0.75, 1) and named the bins respectively 1, 2, 3 and 4. Then, we created a new binary feature called “EXPENDOP_top_quartile” whose domain is [0, 1] where the value 1 marks the patients whose expenses related to the ambulatory visits classify in the top quartile of the distribution and the value 0 marks all the others. In this way, we can easily identify all the high cost patients and model the prediction task as a binary classification problem.

4. Machine Learning Models

4.1. Evaluated Algorithms

Given the characteristics of our processed dataset, we decided to implement all the models studied during the course, as they are all potentially suitable for our prediction task. Going deeper into detail, we will implement the following algorithms:

- **J48**: is an implementation of the decision tree model “C4.5”. Through J48 it is possible to classify any type of data (either nominal or numerical). As for any “decision tree model” the dataset is divided, split into knots until a leaf node, whose frequency distribution is used to make the classification (classification is based on the most frequent class), is reached. We decided two as the minimum numbers of instances for leaf.

- **Random Forest:** a type of supervised machine learning algorithm based on learning multiple predictive models (decision trees) to form a single, more powerful prediction model: this algorithm combines many decision trees in a single model. Individually, the forecasts made by each decision-making tree may not be accurate, but once combined, the forecasts will be, on average, closer to the result. There is a direct relationship between the number of trees in the forest and the accurateness of the results it may achieve: the greater the number of trees, the more accurate the result. The advantages of the random forest model are the ability to classify or perform regressions and the possibility to avoid overfitting the model thanks to the high number of trees. We set up the model to generate ten decision trees.

- **Naïve Bayes:** a supervised learning algorithm used for troubleshooting binary (two-class) and multi-class classification problems. This algorithm exploits the concept of posterior probability, as explained by the Bayes' Theorem with strong independence assumptions between features, to classify all instances as either 0 or 1 based on their respective posterior probability.

- **Logistic regression:** a predictive analysis algorithm that uses a logistic function to model a binary dependent variable. The logistic regression hypothesis tends to limit the cost function between 0 and 1. The Sigmoid function allows to determine whether the class value is 0 or 1.

- **Multilayer perceptron:** is composed of a series of input and output nodes connected to each other to simulate a network of biological neurons. The model allows for the insertion of one or more levels of hidden layers: inside each hidden layer there are artificial neurons and each neuron, once activated, releases its output to every other neuron in the next layer. The nodes are not bi-oriented: the signal propagates from the input neuron to the output neuron without going back.

  We tuned the neural network after having conducted many experiments concerning the number of artificial neurons, while we kept the number of hidden layers fixed at 1 for computational reasons. We obtained the best results with 8 artificial neurons (see Fig.2 and Fig.3 in the Appendix for further data on the selection process).

5. Feature Selection

5.1. Elimination of Redundant Attributes

In order to achieve meaningful results, we must make sure that the attributes used for the classification are meaningful and that they are not correlated to each other. In Fig 4 we can see the correlation matrix for the processed dataset after having eliminated all the attributes that are clearly redundant.

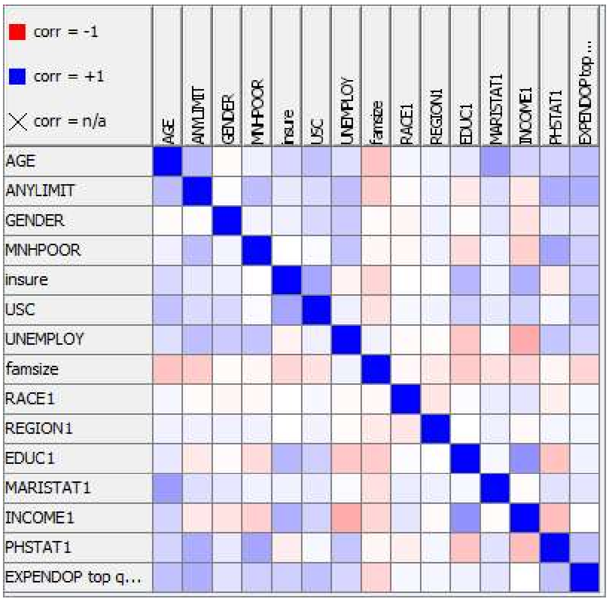


Figure 4 – Correlation Matrix for the processed dataset. (See Appendix for the numeric visualization Matrix, Fig 4.b)

5.2. Filters

As there are no signs of multicollinearity between our 15 variables, we can now investigate where the predictive power comes from by applying filters that automatically select the features that contribute the most to the prediction.

Specifically, we applied an AttributeSelectedClassifier node set up to perform the "Best First" search method on each model and evaluate the results with the "CfsSubsetEval" algorithm.

As a result, we reduced the number of attributes to just 6: "AGE", "ANYLIMIT", "MNHPOOR", "INSURE", "USC" e "PHSTAT1". Unsurprisingly, the filters selected the same attributes for all the above-mentioned classifiers.

In Fig 5 we can see the correlation matrix for the reduced dataset.

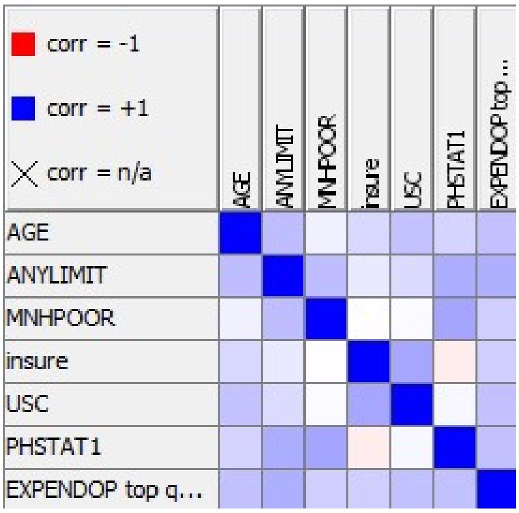


Figure 5 – Correlation Matrix after Feature Selection. (See Appendix for the numeric visualization Matrix, Fig 5.b)

6. Results and Validation

6.1. Overall Model Performance

The non - cross - validated models achieved satisfying error rates. The logistic regression shows an anomalous result that may be caused by the uneven distribution of results of the class variable. In Figure 6 we report the error rates for these models

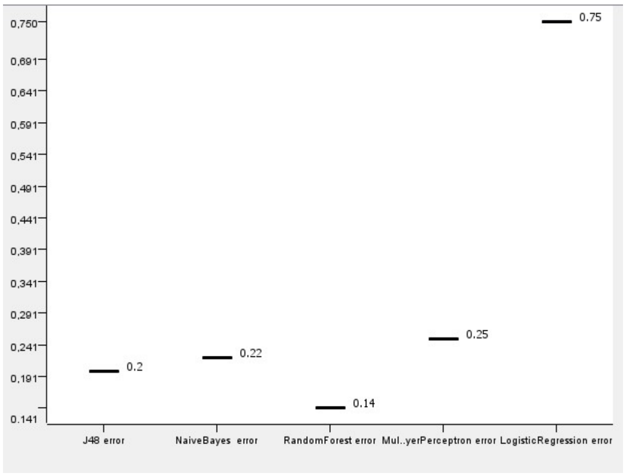


Figure 6 – Error rates for the non – Cross - Validated Models

However, in order to eliminate the bias that may be caused by the phenomenon of overfitting, we decided to perform a 10-fold Cross – Validation of each model, which produced the following error rates

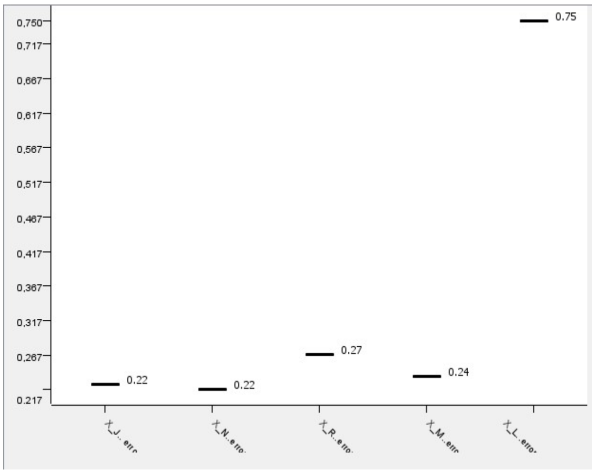


Figure 7 – Error rates of the 10-fold Cross - Validated Models

We can see how the performance of the Random Forest was biased by overfitting, while, on the other hand, the performance of the Multilayer Perceptron improves with Cross Validation and all the others remain basically the same.

6.2. Validation

When it comes to the validation and the interpretation of our results, we must remember to deal with the phenomenon of *class imbalance*, or the fact that within an attribute, one instance is far more frequent than the others. In our case, the instance 0 of our binary class attribute "EXPENDOP_top_quartile" accounts for 75% of all the occurrences.
Unfortunately, such imbalance may negatively affect the performance of the models.
In order to avoid such bias, we must compute a *confusion matrix* that allows us to calculate *Recall, Precision* and *F-Measure* for each model. Let us define these indices before illustrating the results:

		Inducer Prediction	
		-1	+1
Actual Class	-1	TN	FP
	+1	FN	TP

Figure 8 - Confusion Matrix Table

Recall: $\boldsymbol{R} = \frac{TP}{TP+FN}$. It indicates the capability of the model to correctly identify all the positive class instances

Precision: $\boldsymbol{P} = \frac{TP}{TP+FP}$. It indicates how well the model identifies the relevant instances

F-Measure: $\boldsymbol{F} = \frac{\boldsymbol{2*R*P}}{\boldsymbol{R+P}}$**.** It is configured as a harmonic mean of the two above–mentioned measures

We calculated these measures for the three best performing models in terms of error rates: The Naïve Bayes, the J48 Decision Tree and the Multilayer Perceptron. In Fig. 9,10 and 11 we can see the boxplots showing Recall, Precision and F – Measure for each model.

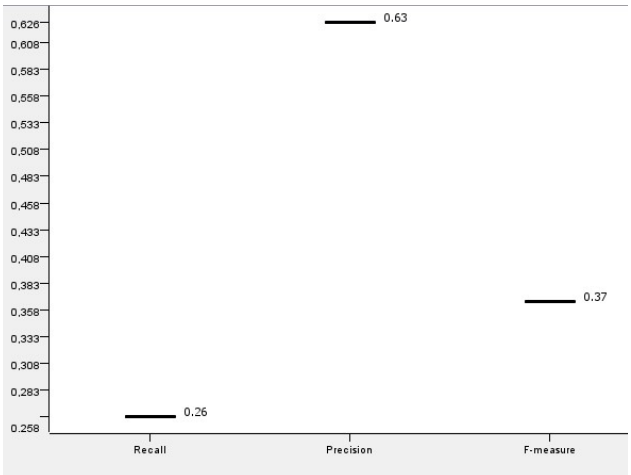


Figure 9 – Validation Measures for the J48 Decision Tree

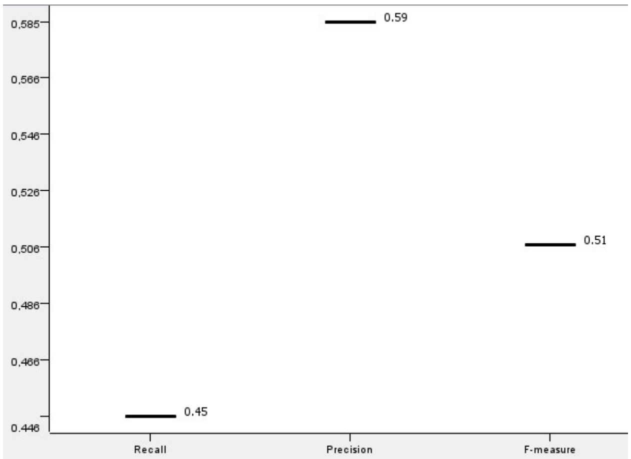


Figure 10 – Validation Measures for the Naive Bayes Classifier

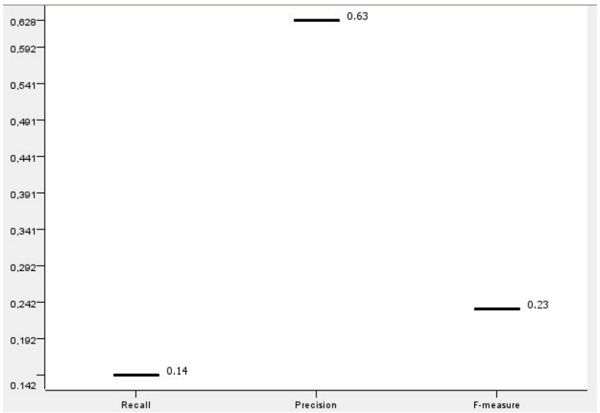


Figure 11 – Validation Measures for the Multilayer Perceptron

6.3. ROC Curve and Area Under the Curve

So far the Naive Bayes is the best model. However, in order to be fully certain of the solidity of our results, we calculated the *ROC Curve* for each model and the corresponding Area Under the Curve.

The *Receiver Operating Characteristics Curve* is a graphical plot that that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied.
The *ROC curve* is created by plotting the *True Positive Rate (TPR)* against the *False Positive rate (FPR)* at various threshold settings. Figure 12, 13 and 14 show the *ROC curves* obtained (See Appendix).

The *Area Under the Curve* is equal to the probability that the given classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In this perspective, a value of 0.5 means that the classifier performs as good as a random classifier, while a value of 1 indicates a theorically perfect classifier that will always rank positive instances higher than negative instances and thus will never make classification mistakes.

For the J48 Decision Tree we obtained an AUC of 0.63, the Naive Bayes reache a value of 0.794 and the Multilayer Perceptron 0.8.

6.4. Algorithm Choice

Given the obtained results, we consider the Naive Bayes to be the best model, as it shows the best error rate (0.22), a good value of Area Under the Curve (0.794), and the best trade off concerning Recall (0.45) and Precision (0.59) as demonstrated by the F-Measure (0.51).

On the other hand, the Multilayer Perceptron scored similarly, so we calculated the Confidence Interval for the variance of the difference of the mean error rates in order to check if the difference of error rates is actually statistically relevant or not.

We adopted the following formula for the computation of the lower and upper bound of the confidence interval:

$$(d - z_{1-\frac{\alpha}{2}} * \sigma_d^2 \; ; d + z_{1-\frac{\alpha}{2}} * \sigma_d^2 \;)$$

Where:

e_{mNB} = mean error rate for the Naïve Bayes Classifier
e_{mML} = mean error rate for the Multilayer Perceptron
d = $e_{mNB} - e_{mML}$

α = 0.05

The resulting interval contains the value 0, so we conclude that the difference between the mean error rates of the two models is not statisticaly relevant at the confidence level of 95%. In Fig 15 we can see the actual interval obtained

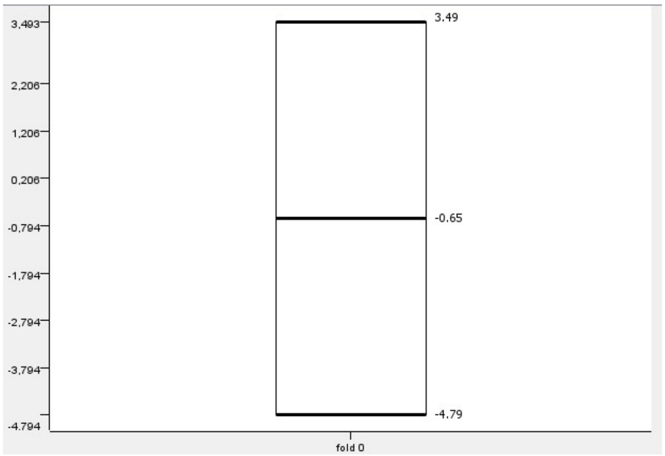


Figure 15 – Confidence Interval for the difference of the variance of the error rates of the Naive Bayes Classifier and the Multilayer Perceptron, the interval contains the value 0

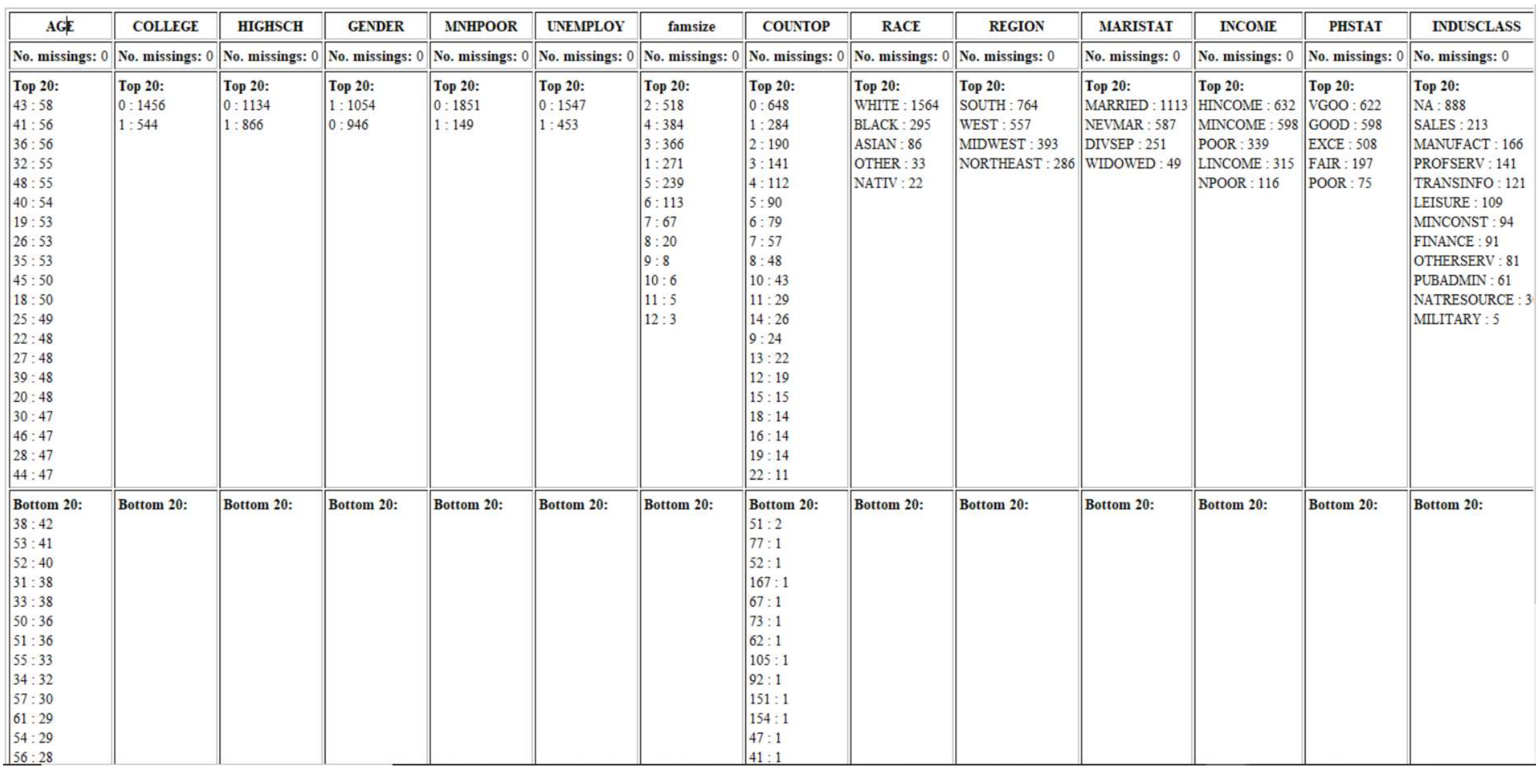
7. Conclusions

In this paper we tried to build a prediction model able to classify previously unseen records and thus identify the high cost patients. The best results were obtained with a 10 fold Cross Validated *Naive Bayes* Classifier, which scored an error rate of 0.22, an F-measure of 0.51 and a value of 0.794 of *Area Under the Curve*. It would be very interesting to see how these numbers changed if we introduced diagnostic attributes to the dataset. In particular, research[2] shows that one of the main cost drivers for public healthcare systems is comorbidity, but, unfortunately, we were not able to find any medical data that could enrich our original dataset.

References

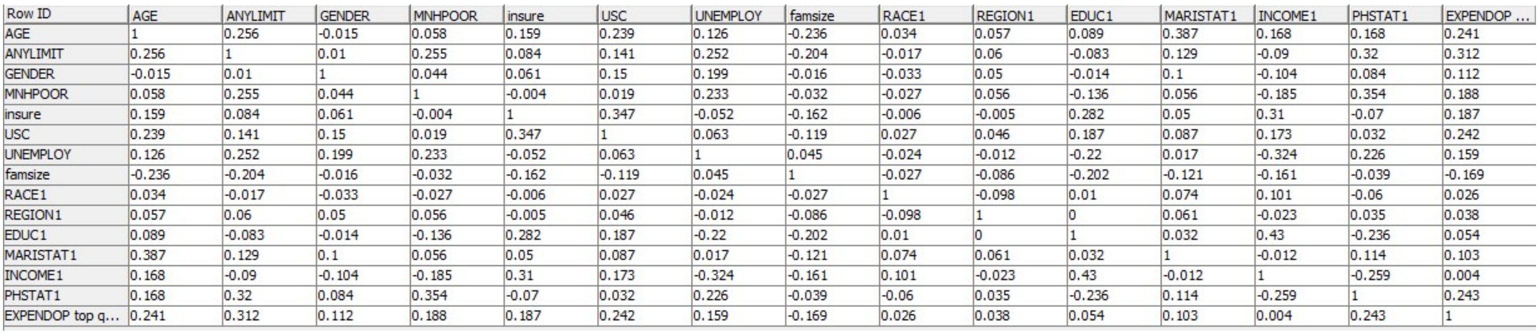
1. Blumental, D. New engla nd journal. 1–3 (2016).
2. Fleishman, J. A. & Cohen, J. W. Using Information on Clinical Conditions to Predict High-Cost Patients. 532–552 (2009). doi:10.1111/j.1475-6773.2009.01080.x

Appendix



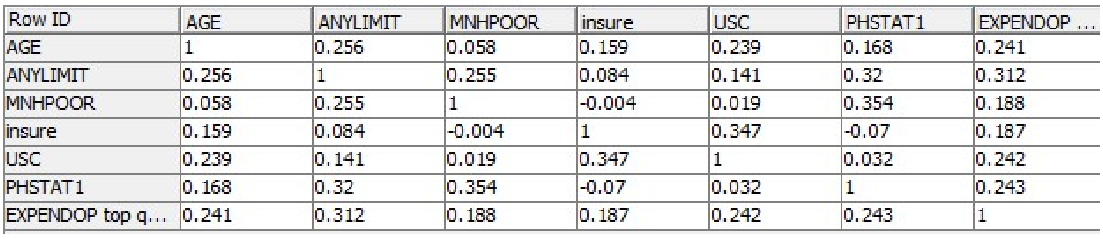
AGE	COLLEGE	HIGHSCH	GENDER	MNHPOOR	UNEMPLOY	famsize	COUNTOP	RACE	REGION	MARISTAT	INCOME	PHSTAT	INDUSCLASS
No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0
Top 20: 43 : 58 41 : 56 36 : 56 32 : 55 48 : 55 40 : 54 19 : 53 26 : 53 35 : 53 45 : 50 18 : 50 25 : 49 22 : 48 27 : 48 39 : 48 20 : 48 30 : 47 46 : 47 28 : 47 44 : 47	**Top 20:** 0 : 1456 1 : 544	**Top 20:** 0 : 1134 1 : 866	**Top 20:** 1 : 1054 0 : 946	**Top 20:** 0 : 1851 1 : 149	**Top 20:** 0 : 1547 1 : 453	**Top 20:** 2 : 518 4 : 384 3 : 366 1 : 271 5 : 239 6 : 113 7 : 67 8 : 20 9 : 8 10 : 6 11 : 5 12 : 3	**Top 20:** 0 : 648 1 : 284 2 : 190 3 : 141 4 : 112 5 : 90 6 : 79 7 : 57 8 : 48 10 : 43 11 : 29 14 : 26 9 : 24 13 : 22 12 : 19 15 : 15 18 : 14 16 : 14 19 : 14 22 : 11	**Top 20:** WHITE : 1564 BLACK : 295 ASIAN : 86 OTHER : 33 NATIV : 22	**Top 20:** SOUTH : 764 WEST : 557 MIDWEST : 393 NORTHEAST : 286	**Top 20:** MARRIED : 1113 NEVMAR : 587 DIVSEP : 251 WIDOWED : 49	**Top 20:** HINCOME : 632 MINCOME : 598 POOR : 339 LINCOME : 315 NPOOR : 116	**Top 20:** VGOO : 622 GOOD : 598 EXCE : 508 FAIR : 197 POOR : 75	**Top 20:** NA : 888 SALES : 213 MANUFACT : 166 PROFSERV : 141 TRANSINFO : 121 LEISURE : 109 MINCONST : 94 FINANCE : 91 OTHERSERV : 81 PUBADMIN : 61 NATRESOURCE : 3 MILITARY : 5
Bottom 20: 38 : 42 53 : 41 52 : 40 31 : 38 33 : 38 50 : 36 51 : 36 55 : 33 34 : 32 57 : 30 61 : 29 54 : 29 56 : 28	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:** 51 : 2 77 : 1 52 : 1 167 : 1 67 : 1 73 : 1 62 : 1 105 : 1 92 : 1 151 : 1 154 : 1 47 : 1 41 : 1	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**

Fig 1. Descriptive Statistics for the original Dataset



Row ID	AGE	ANYLIMIT	GENDER	MNHPOOR	insure	USC	UNEMPLOY	famsize	RACE1	REGION1	EDUC1	MARISTAT1	INCOME1	PHSTAT1	EXPENDOP ...
AGE	1	0.256	-0.015	0.058	0.159	0.239	0.126	-0.236	0.034	0.057	0.089	0.387	0.168	0.168	0.241
ANYLIMIT	0.256	1	0.01	0.255	0.084	0.141	0.252	-0.204	-0.017	0.06	-0.083	0.129	-0.09	0.32	0.312
GENDER	-0.015	0.01	1	0.044	0.061	0.15	0.199	-0.016	-0.033	0.05	-0.014	0.1	-0.104	0.084	0.112
MNHPOOR	0.058	0.255	0.044	1	-0.004	0.019	0.233	-0.032	-0.027	0.056	-0.136	0.056	-0.185	0.354	0.188
insure	0.159	0.084	0.061	-0.004	1	0.347	-0.052	-0.162	-0.006	-0.005	0.282	0.05	0.31	-0.07	0.187
USC	0.239	0.141	0.15	0.019	0.347	1	0.063	-0.119	0.027	0.046	0.187	0.087	0.173	0.032	0.242
UNEMPLOY	0.126	0.252	0.199	0.233	-0.052	0.063	1	0.045	-0.024	-0.012	-0.22	0.017	-0.324	0.226	0.159
famsize	-0.236	-0.204	-0.016	-0.032	-0.162	-0.119	0.045	1	-0.027	-0.086	-0.202	-0.121	-0.161	-0.039	-0.169
RACE1	0.034	-0.017	-0.033	-0.027	-0.006	0.027	-0.024	-0.027	1	-0.098	0.01	0.074	0.101	-0.06	0.026
REGION1	0.057	0.06	0.05	0.056	-0.005	0.046	-0.012	-0.086	-0.098	1	0	0.061	-0.023	0.035	0.038
EDUC1	0.089	-0.083	-0.014	-0.136	0.282	0.187	-0.22	-0.202	0.01	0	1	0.032	0.43	-0.236	0.054
MARISTAT1	0.387	0.129	0.1	0.056	0.05	0.087	0.017	-0.121	0.074	0.061	0.032	1	-0.012	0.114	0.103
INCOME1	0.168	-0.09	-0.104	-0.185	0.31	0.173	-0.324	-0.161	0.101	-0.023	0.43	-0.012	1	-0.259	0.004
PHSTAT1	0.168	0.32	0.084	0.354	-0.07	0.032	0.226	-0.039	-0.06	0.035	-0.236	0.114	-0.259	1	0.243
EXPENDOP top q...	0.241	0.312	0.112	0.188	0.187	0.242	0.159	-0.169	0.026	0.038	0.054	0.103	0.004	0.243	1

Fig 4.b – Numeric Correlation Matrix for the processed dataset



Row ID	AGE	ANYLIMIT	MNHPOOR	insure	USC	PHSTAT1	EXPENDOP ...
AGE	1	0.256	0.058	0.159	0.239	0.168	0.241
ANYLIMIT	0.256	1	0.255	0.084	0.141	0.32	0.312
MNHPOOR	0.058	0.255	1	-0.004	0.019	0.354	0.188
insure	0.159	0.084	-0.004	1	0.347	-0.07	0.187
USC	0.239	0.141	0.019	0.347	1	0.032	0.242
PHSTAT1	0.168	0.32	0.354	-0.07	0.032	1	0.243
EXPENDOP top q...	0.241	0.312	0.188	0.187	0.242	0.243	1

Fig 5.b – Numeric Correlation Matrix for the dataset after Feature Selection

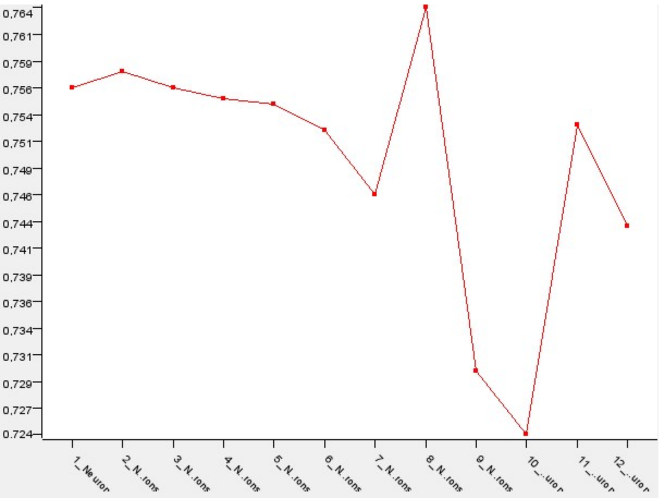


Fig 2. MultiLayer Perceptron Study – A comparison of the accuracy achieved as the number of artificial neurons in the hidden layer increases

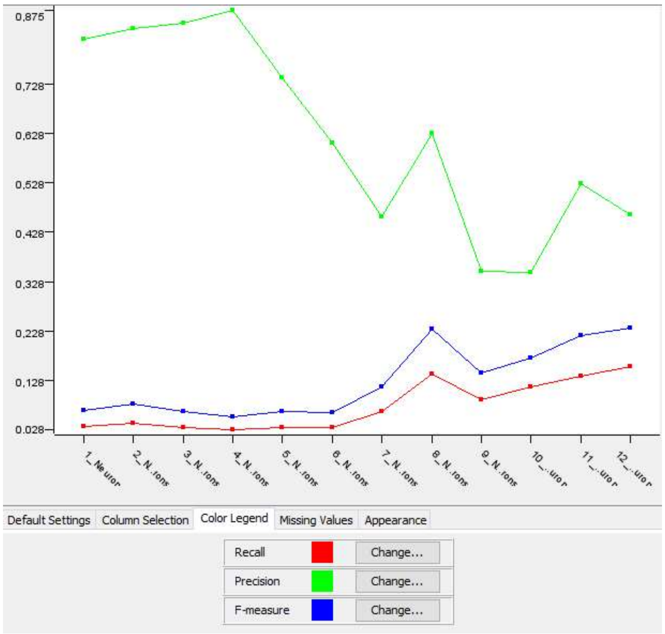


Fig 3. MultiLayer Perceptron Study – Comparison of the level of Recall, Precision and F-Measure achieved as number the neurons in the hidden layer increases

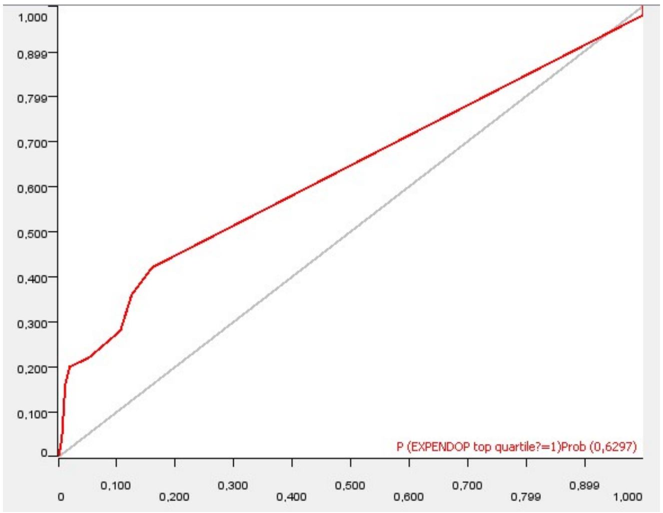


Fig 12. ROC curve obtained from the J48 Decision Tree

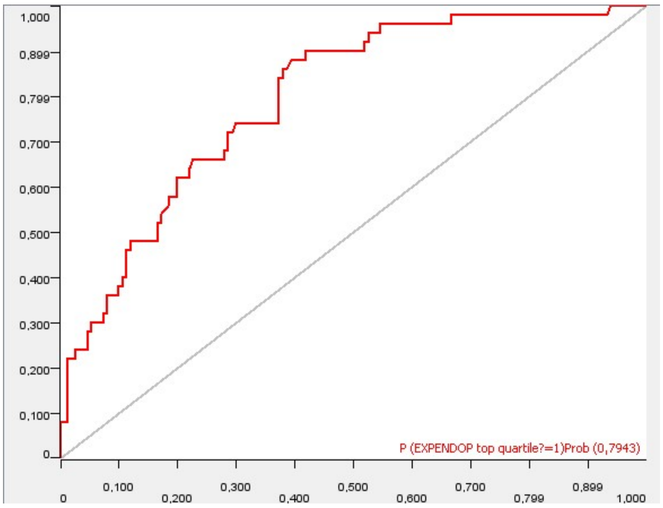


Fig 13. ROC curve obtained from the Naive Bayes Classifier

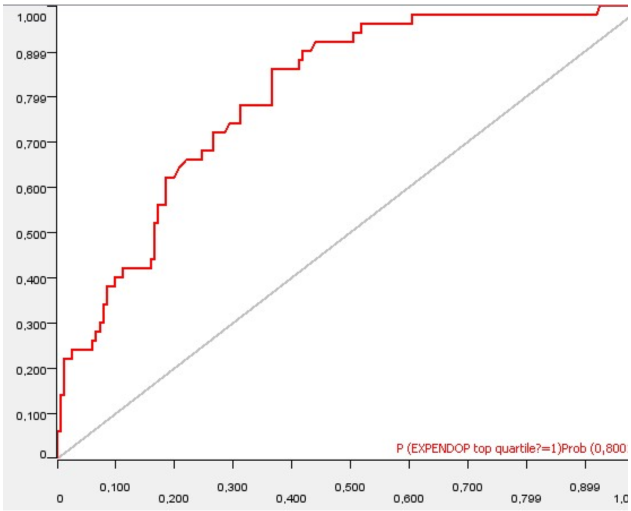


Fig 14. ROC curve obtained from the MultiLayer Perceptron