**COMPSCI 9637A**

**Introduction to Data Science I**

**Final Report**

**Due: December 4, 2020**

**Submitted By:**

Kolby Sarson

Sergio Zapata

**Abstract**

This project aims to find what clinical features may have the most significant effect on high blood pressure. Therefore, we analyzed a dataset of 299 patients with heart failure collected in 2015. To determine whether high blood pressure is present or not, we used methods such as logistic regression, decision tree, and random forest, among others. We also perform a logistic regression classifier given the best parameters from the grid search, using only the 3 top features and got similar accuracies and cross validation scores to the iteration with the full dataset. Our results have shown that serum creatinine(SC) smoking and age show promise in predicting high blood pressure in patients from medical records. This new approach has a big potential to healthcare providers, and can become a new supporting tool for health practitioners when predicting if a patient has, or will have, this condition. In fact, health practitioners aiming to predict if a patient’s likeliness to have high blood pressure may focus on serum creatinine, whether the patient smokes, and the patient’s age.

This report will begin with an Introduction, both to the problem space and the importance of our research, as well as our proposed approach to solving the problem. Next, in the Data section, it will outline the dataset being utilized throughout our research, including its structure, composition, and explanation of its features. After examining the data, in Methods, the design of our project is discussed and justified, both from machine learning and data science perspectives. Then we summarize the outcomes obtained throughout our research in the Results section. Finally, we end with a Discussion of these results and the conclusions we made. We also discuss the implications of our research outside of a research setting.

**Introduction**

The medical field is wide-spanning and integrally important to our betterment as a whole. With so many medical conditions known, any advancement in this field involving any of these conditions could prove overwhelmingly beneficial. High blood pressure is one such condition which can cause disastrous effects in the long-term, such as heart disease and circulatory complications. This is the condition that we decided to focus on in our research. Before beginning research we had to produce a research problem that specifies our objective. This question we constructed is; “What are the most important and least important clinical features (or risk factors) that may lead to high blood pressure?”

High blood pressure, also called hypertension, is blood pressure that is higher than normal. As such, blood pressure is measured using two numbers: The first number, called systolic blood pressure, measures the pressure in your arteries when your heart beats. The second number, called diastolic blood pressure, measures the pressure in your arteries when your heart rests between beats (i.e., If the measurement reads 120 systolic and 80 diastolic, you would say, “120 over 80,” or write, “120/80 mmHg”). Having blood pressure measured consistently above normal may result in a diagnosis of high blood pressure (or hypertension). Thus, the higher your blood pressure levels, the more risk you have for other health problems, such as heart disease, heart attack, and stroke [3].

Currently, many health systems across the world either use electronic medical records, or are transitioning from physical to electronic medical records to utilize clinical features in the tracking of a) patient’s symptoms, b) body conditions and, c) laboratory test values, among others. All these records can be used to perform biostatistics analysis, which aimed at highlighting patterns and correlations that might not be easy to detect by health practitioners. Machine learning in particular, can predict patients’ risk of several medical conditions from their medical records and can predict the most important clinical features that may lead to those.

Now back to our question, the approach we took to answer this question involved a combination of machine learning practices and data science analysis techniques. In simple terms, we would begin by reading in and taking a high-level look at the dataset chosen. We would then proceed to perform feature engineering in a range of different structures, trying different combinations of features, constructions, and encodings. Once we find a combination which we feel would be the best given the data, we then progress to investigating this altered version of the dataset at a high-level. Following this, we will apply a variety of different machine learning classifiers to the dataset using only the default parameters, then test for the best combination of model parameters to utilize them in a secondary application of the classifiers. We will use the accuracies we find in this step, alongside the definitions of the classifiers, to pick the best suited classifier for further investigation. Next, we will take the best suited classifier previously selected, fit it to the dataset, and perform a more in-depth investigation into the results, such as determining the best features, through the use of data science techniques. Finally, we will re-apply the logistic regression classifier on a dataset consisting of only the best features determined in the previous analysis.

**Data**

The dataset we selected for this research, titled “heart\_failure\_clinical\_ records,” is a dataset collected with the original intent to look for relationships that could predict heart failure survival rates. The dataset consisted of 299 patients experiencing heart failure; 194 male and 105 female, all of which are older than 40 years and from Faisalabad, Pakistan. It featured a follow up time in days with a binary classification of a death event, either true or false representing survival. A variety of personal information and bodily measurements were recorded, including: Age, Anaemia, Creatinine Phosphokinase (CPK), Diabetes, Ejection Fraction (EF), High Blood Pressure (BP), Platelets, Serum Creatinine, Serum Sodium, Sex, and Smoking. These recordings were a combination of discrete and continuous, with some ranging from 0-1 and other exceeding 1. Normal ranges for the bodily measurements are as follows: EF between 30 and 45, Platelets between 212,500 and 303,500, Serum Sodium between 135 and 145, and Serum Creatinine less than or equal to 1.5 (Table 1).

Given that this dataset was created with the intent of predicting heart failure survival rates, we did not want to re-examine this problem. Instead we wanted to use this dataset to investigate other possibilities, namely other relationships involving high blood pressure. As such, the time feature and death event label could be removed entirely, and high blood pressure would act as our new label.

**Methods**

This section is divided into two subsections, *Machine Learning Classifiers* and *Data Science Techniques*. Each subsection will discuss the methods utilized for the associated portion of our research.

***Machine Learning Classifiers***

In our research, we wanted to test a variety of machine learning classifiers for baseline performance prior to diving into a data science focused analysis. The classifiers we chose to include were Gaussian Naive Bayes, Perceptron, Decision Tree, Logistic Regression, K Nearest Neighbours, and Random Forest. We applied a grid search on some common parameters for each in order to test combinations of parameters and determine a set of best parameters for each. The selected parameters are shown below.

*Gaussian Naive Bayes* - var\_smoothing: 1e-09

*Perceptron* - max\_iter: 500, penalty: L2, random\_state: 0

*Decision Tree* - criterion: entropy, splitter: test, random\_state: 0

*Logistic Regression* - max\_iter: 500, penalty: L2, random\_state: 0

*K Nearest Neighbours* - metric: chebyshev, n\_neighbours: 16, weight: uniform

*Random Forest* - criterion: entropy, max\_depth: 2, n\_estimators: 50, random\_state: 0

***Data Science Techniques***

As stated previously, the aim of this project is to find what clinical features may have the most significant effect on high blood pressure. As such, we applied logistic regression [8] to the dataset, and dropped some of the original clinical features that have no relevant use for this project (i.e., time, and DEATH\_EVENT). Then, we use feature ranking importance as the logistic regression model coefficient for each variable (Table2).

We also found the 95% confidence interval(CI) of coefficients [7] for each variable to reject the null hypothesis (i.e., non-variable has any effect on high blood pressure). However, we did not reject the null hypothesis as we found one feature that has a significant effect on high blood pressure (Table3, Graphs 1 & 2). In addition, we implemented odds [2] to find what other features might have an effect on high blood pressure. Thus, we found out that the feature we found by using CI was the most important feature by odds as well. As a result, we were also able to find two other important features with a significant effect on high blood pressure (Table 4 and Graph 3).

**Note:** All the data analysis and methods were implemented using Anaconda – python [4,6].

**Results**

In this project, cross validated (CV) accuracy score is used to determine how well each model performs and Maximum likelihood estimate (MLE) is used to estimate the parameter for each independent variable that influences the high blood pressure. Also, we applied a logistic regression model to this dataset, and feature ranking importance as the logistic regression model coefficient for each variable.

As a result, from the confidence interval (CI) coefficients of each variable, we saw that only x10's (or SC’s) CI does not include 0. As such, we do not reject the null hypothesis that SC has no effect on high blood pressure. Indeed, SC may be the most important feature. From the odds we also saw that x1 (or age), x7 (or smoking), and x10 (or SC) are greater than 1, with scores 1.201354, 1.264496, and 1.462010, respectively. As units increase on those three variables, the response is more likely to be positive. We also saw that x10 (or SC) has the largest odds, which indicates that it is the most important feature. In addition, if one unit increases in SC, there will be 0.46 increase in high blood pressure. We also determined age and smoking to be important features, as every unit increased in those, will increase 0.2 and 0.26, respectively, in high blood pressure.

To verify further the predictive power of SC, age and smoking, we illustrated a scatterplot with all the variables on the x axis and the odds values on the y axis. This plot shows a clear distinction between those variables with high effect and the ones with the lower effect.

After all of our previous analyses of the data, we came to the conclusion that serum creatinine, age, and smoking are the most important features, in that order. Given this conclusion, we decided to create a dataset consisting of only these 3 features in addition to our label of high blood pressure. After creating this dataset we applied our logistic regression model to it in order to compare the accuracy and CV accuracy score to that obtained from the full dataset. In the model for the full dataset we obtained an accuracy score and CV accuracy score of 0.59 and 0.56, respectively. In the model for the 3 feature dataset we obtained an accuracy score and CV accuracy score of 0.66 and 0.59, respectively. We can see that both a simple accuracy score and the CV accuracy score performed better by 0.07 (7%) and 0.03 (3%), respectively.

**Discussion**

From the receiver operating characteristic (ROC) [5] curve for each classification method we applied, we state that these do not perform very well. This may be caused by imbalanced data, as there are approximately two thirds observations that do not have enough observations of high blood pressure. From the CI of the coefficients for each variable, we can see that only SC’s CI does not include 0. Thus, from all the clinical features of this data set, SC is the most important clinical feature. From the odds we can see that age, smoking, and SC have values greater than 1. As a consequence, every time that units increase on those three variables, the response is more likely to be positive. We can see that SC has the largest odds, which indicates that it may be the most important feature. In conclusion, from the logistic regression SC is the most important feature.

Our results have shown that serum creatinine (SC), smoking, and age show promise in predicting high blood pressure in patients from medical records. This new approach has a big potential to health providers, and can become a new supporting tool for health practitioners when predicting if a patient has or will have this condition. In fact, health practitioners aiming to predict a patient’s likeliness to have high blood pressure may focus on serum creatinine, whether the patient smokes, and the patient’s age.

In terms of the overall performance of our models, the results were not promising. As mentioned in our results section, we achieved an accuracy of 0.66 and a cross-validated accuracy of 0.59 when applying logistic regression on the best 3 features dataset. Even taking the accuracy score, a 66% accuracy is promising, but it is not sufficient as is, especially considering that our work is in the field of medicine. Because of this, we do not think that our classifier could be used in practice, at least not in its current state. The accuracy of our classifier would need to be improved before being acceptable for use in medical diagnosis.

To improve upon the work we have already done, there are two main approaches. The first would be to consider the models used. There are different ways in which we can consider improvements through the use of the models. We could consider changes to our current models, or as a better alternative, we could implement new models, such as convoluted neural networks. The second approach would be to consider the data itself. Being that this dataset was created for the intention of predicting heart failure survival, and not for predicting high blood pressure, the data might not be favourable for our uses in this research. That is why the second approach of considering the data is one which we recommend. We could either reconsider our feature engineering of the current data, which we feel would be ineffective, or we could consider gathering more data with more relevant features that could prove to be better for our classification of high blood pressure.

**References**

[1] Ahmad, Tanvir Haider, et al. “Survival Analysis of Heart Failure Patients: A Case Study.” *PLOS ONE*, Public Library of Science, 20 July 2017, journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001.

[2] Bland, J. Martin, and Douglas G. Altman. "The odds ratio." *Bmj* 320, no. 7247 (2000): 1468.

[3] Center for Disease Control and Prevention, High Blood Pressure. <https://www.cdc.gov/bloodpressure/about.htm> Accessed Nov 1, 2020.

[4] Hackeling, Gavin. *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.

[5] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology* 143, no. 1 (1982): 29-36.

[6] Horstmann, Cay S., and Rance D. Necaise. *Python for everyone*. Wiley Publishing, 2015.

[7] Hosmer, David W., and Stanley Lemeshow. "Confidence interval estimation of interaction." *Epidemiology* (1992): 452-456.

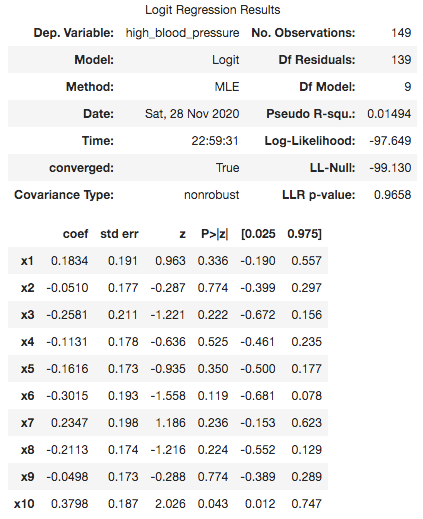
[8] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

**Appendix**

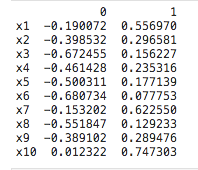
**Table 1**

|  |  |  |  |
| --- | --- | --- | --- |
| Meanings, measurement units, and intervals of each feature of the dataset | | | |
| **Feature** | **Explanation** | **Measurement** | **Range** |
| Age | Age of the patient | Years | [40, ..., 95] |
| Anaemia | Decrease of red blood cells or hemoglobin | Boolean | 0, 1 |
| Creatinine phosphokinase | Level of the CPK enzyme in the blood | mcg/L | [23, ..., 7861] |
| Diabetes | If the patient has diabetes | Boolean | 0, 1 |
| Ejection fraction | Percentage of blood leaving  the heart at each contraction | Percentage | [14, ..., 80] |
| Sex | Woman or man | Binary | 0, 1 |
| Platelets | Platelets in the blood | kiloplatelets/mL | [25.01, ..., 850.00] |
| Serum creatinine | Level of creatinine in the blood | mg/dL | [0.50, ..., 9.40] |
| Serum sodium | Level of sodium in the blood | mEq/L | [114, ..., 148] |
| Smoking | If the patient smokes | Boolean | 0, 1 |
| (target) High blood pressure | If a patient has hypertension | Boolean | 0, 1 |
| mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre | | | |

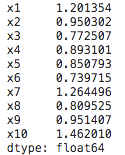
**Table2**



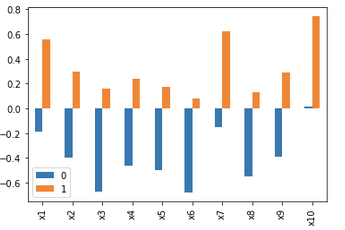
**Table 3**

****

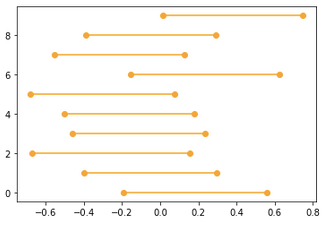
**Table 4**

****

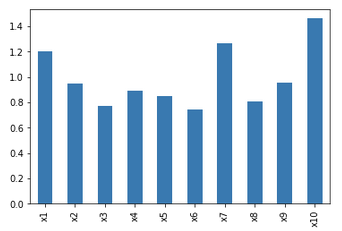
**Graph 1**

****

**Graph 2**

****

**Graph 3**

****