Prediction of Stroke Mortality Rate across Counties in the United States

Davis Agughalam Munachimso
School of Computing
National College of Ireland
Dublin, Ireland
agughalamdavis@yahoo.com

Adebola Abdullahi-Attah
School of computing
National College of Ireland
Dublin, Ireland
adebolaaattah@gmail.com

Babasoji Shobayo
School of computing
National College of Ireland
Dublin, Ireland
shobayosoji@gmail.com

***Abstract*—The application of machine learning for health related diagnosis is on the increase. Improvements in algorithms and computing power have been leveraged for quicker and arguably more accurate diagnosis and subsequent treatment increasing survival rates significantly. This research takes a look at leveraging machine learning algorithms, linear regression and decision trees to predict stroke mortality rate across counties in the United States based on selected factors heart disease mortality rate, population and teen birth rates and also understanding the relationship therein. The results were evaluated in terms of the mean squared error and linear regression model outperformed the decision tree model as its predicted values were closer to the actual values than the decision tree predicted values.**

***Index Terms*—stroke mortality rate, machine learning, heart disease mortality rate, population, teen birth rate.**

I. Introduction

Machine learning techniques have been widely adopted across many sectors to solve real world problems. In medical research, due to an alarming increase of diseases, physiological conditions and climate changes, data mining techniques have been adopted to gain insights for provision of solutions or palliatives to medical issues. Presently, one increasing cause of death is stroke which is an illness that can be triggered by many symptoms, causing the brain to experience a shortage in the supply of blood to the blood vessels. Due to the increasing rate of stress and other related diseases, the symptoms of stroke have also increased giving concern to the global health community on ways to curb the menace. Understanding some factors linked to stroke mortality rate is one of the areas where machine learning can be leveraged to gain insights. Therefore, this paper investigates the extent to which two machine learning algorithms, linear regression and Decision trees regression can be used to predict stroke mortality rate using selected predictors such as teen birth rate, heart disease mortality rate and population and also understanding the relationship between them. This research also compares the stroke mortality levels in males and females in the areas of focus.

II. Background

Machine learning algorithms can be broadly divided into supervised and unsupervised learning algorithms. The supervised learning algorithms are used when there is a clear target variable for either prediction or classification while the unsupervised algorithms are used when there is no clear target variable. They are mostly employed for inferential purposes or as precursors for the application of supervised learning algorithms. The algorithms used in this research, linear regression and decision tree algorithms are supervised machine learning algorithms and are used because there is a target variable stroke mortality rate to be predicted.

- **Multiple Linear Regression:** Linear regression models use a linear model that defines the relationship between the outcome variable and the independent variables[1]. In order to estimate the optimal relationship between the data points, it uses the ordinary least squares estimation technique to propose a line of best fit through the data. The equation for this line of best fit is the governing relationship between the predicted and the predictor variables. It is mathematically expressed as;

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \qquad (1)$$

  where y is the outcome variable, β_0 is the intercept, β_1, β_2, β_n are coefficients and X_1, X_2 and X_n are predictor variables.

- **Decision Tree:** Similar to a real life tree structure, the decision tree is a machine learning algorithm that can be used for both classification and regression. The decision tree can be conceptualized as a series of yes or no questions that lead towards the final goal of prediction from the data set[1]. The data is subdivided into homogenous subsets using the most notable splitting condition at each of the nodes. This splitting condition is based on different algorithms such as the gini impurity, chi square or information gain. At each node of the tree, a question is determined that when answered leads to the reduction of the gini impurity. The decision tree searches through the values of the features on which splits can be made to ensure lower gini impurity values. The Gini Impurity of a node is the probability that a sample chosen at random in a node would be labeled incorrectly if the distribution of samples in the node was used for labelling. The tree keeps dividing the nodes till

the final predictions are made [1].

III. Related work

The application of machine learning algorithms in the health sector for early detection of diseases, understanding relationship between some symptoms and diseases or correct diagnosis has increased in recent times with many researchers looking in this direction. Kim et al. [2] in their research to detect acute hemorrhage shock using rats applied machine learning algorithms logistic regression (LR), Artificial neural networks (ANN), random forest (RF) and support vector machine (SVM). They carried out tests on thirty six anesthetized rats and grouped them in 3 according to the volume of blood loss.They took measurements such as heart rates, blood lactate concentration amongst others and these served as predictors in the model. The SVM model performed better than the other models in terms of sensitivity and area under the curve with values of 1.0 and 0.972 respectively. As insightful as the results were, some of the limitations faced include the fact that due to the use of anesthetics, the recordings from the subjects might not have been the correct value under normal circumstances because of the impact on the cardiovascular system. They also recommended the use of isuflorane anesthesia as this provided more reliable results. Ma et al. [3] employed machine learning techniques to predict the risk of acute renal failure in patients. Three data mining algorithms, logistic regression, C5.0 decision tree and artificial neural networks were used and compared to each other. The C5.0 decision tree algorithm performed better than the other two algorithms in predicting patient death and dialysis. Their results suggested that the decision tree model can be leveraged for real life prognosis. Rafael et al. [4] used machine learning algorithms to identify clinically relevant features in hypertensive patients. They used penalized regression techniques to find the most important features for prediction of cardiovascular illnesses in hypertensive patients. The models used were the least absolute shrinkage and selection operator (LASSO), elastic net (EN) and the cox proportional hazards regression. They collected laboratory data from 1470 patients in a clinic and fit the models using 10 fold cross validation on the data. The cox model was able to identify 5 features while the other two models identified only 3. In terms of evaluation metrics, the cox model also outperformed the other algorithms, though further statistical tests showed that there were no statistically significant differences between the model predictions. Their research provided models that can be used by clinicians for quick detection of patients at cardiovascular illness risk. Jabbar et al. [5] explored the use of alternating decision trees for early diagnosis of heart diseases to produce a timely and cost effective decision system.They applied the algorithm on various heart disease patient records obtained from different hospitals using principal component analysis (PCA) to select the most essential features for heart disease diagnosis. The proposed model had a high accuracy of 91.66 percent and proved efficient in heart disease diagnosis. Predicting stroke using SVM was the focus of the research carried out by Jeena and Kumar [6]. They relied on factors such as age, sex and walking symptoms amongst others for this prediction. They tested different kernels for the SVM classifier and the linear kernel recorded the highest accuracy of 91 percent. It also outperformed the other kernels in terms of sensitivity, specificity, precision and F1-score. Their work demonstrated the predictive power of a machine learning model for health diseases even though it was only dealt with a small set of input parameters. Garcia-Terriza and Rosello [7] compared different machine learning approaches to model stroke subtype classification and risk prediction.They proposed combining machine learning techniques with non invasive monitoring systems for quick diagnosis and even extending to prediction of future risk such as eventual death. 7 different algorithms, decision tree, K-nearest neighbours (KNN), logistic regression, naive bayes, neural networks, random forest and SVM were used and compared to each other based on classification evaluation metrics. The random forest model outperformed the other models with an accuracy of 93 percent for diagnosis and 97 percent for future risk. This further elaborated the application of machine learning for real life diagnosis of health challenges. Their experiment was successful but only considered medical features that are particular to the patient and did not include environmental features that may be of impact. Heo et al. [8] researched the development of a machine learning based model for predicting outcomes in patients with acute stroke. 3 machine learning models, deep neural networks, random forest and logistic regression were used on data for 2,604 patients. These models were compared to each other based on their accuracy and area under the curve (AUC) values. They were also compared to the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score. The area under the curve for the deep neural network was significantly higher than that of the ASTRAL score but the AUC for the random forest and logistic regression were not significantly higher. Using only the 6 variables used for the ASTRAL score the machine learning models did not perform significantly better. Rather than using health related features, Kansadub et al.[9] researched stroke prediction based on demographic data such as province, level of education and occupation. 3 classification algorithms, decision tree, naive bayes and neural networks are used and evaluated based on accuracy and area under the curve. The decision tree algorithm had the most accuracy for classification while the naive bayes algorithm had the highest area under the curve. However, this research was limited due to high data imbalance and the undersampling technique used to balance the data leading to possible loss of valuable information from the data. To further buttress the applicability of machine learning in disease prediction, Kohli and Arora [10], applied classification algorithms logistic regression, decision trees, random forest, support vector machine and Adaptive boosting for prediction of heart disease, breast cancer and diabetes. For prediction of breast cancer, the adaboost algorithm outperformed the other algorithms 95.29 percent. The support vector machine with

a linear kernel was found to have the highest accuracy for diabetes classification having an accuracy of 85.71 percent. The logistic regression had outperformed the other algorithms for the case of diagnosing heart disease with an accuracy value of 87.10 percent. The results further emphasize the need to test more than one algorithm for a particular purpose as there is no one model fits all when using machine learning especially for delicate sectors like the health sector. The central theme of application of machine learning algorithms for health related prediction or classification is evident in the reviewed works with most of them focusing directly on the patient and the patient's features or symptoms. Looking at the mortality rates of the disease in an area with many individuals rather than focusing on one individual can generate broader insights towards potential causes and possible solutions. That is the gap this research attempts to fill by investigating stroke mortality rate and its relationship with other social indicators such as teen birth rate, heart disease mortality rate and population.

IV. Methodology

To facilitate a systematic approach to the research, the knowledge discovery in databases (KDD) process was followed. The process is enumerated as follows.

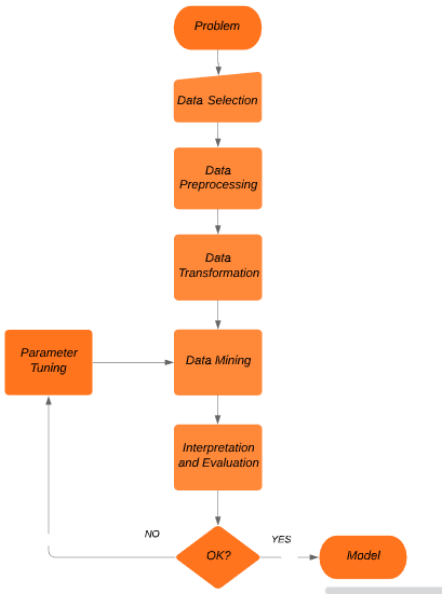

Fig. 1: KDD process

A. Data Selection

As the aim of this research is to predict stroke mortality rate across counties in the US thus the datasets selected had to be related to this. The first dataset contained data for the stroke mortality rate across the counties in 59,095 rows and 23 columns, the second data set contained drug poisoning rates across the counties in 53,388 rows and 8 columns but most importantly, had information about the population of the counties and hence it was chosen. The third dataset contained information about the teen birth rates across the counties in 40,782 and 9 rows while the fourth data set had information about the heart disease mortality rates across these counties in 59,905 rows and 23 columns. These were likely factors affecting stroke mortality rate and was leveraged for the prediction. The four datasets were gotten from the data.gov repository of the United States in JSON format and programmatically stored in Mongo dB database as scrapped from the web. Mongo dB as a NoSQL database was preferred for initial data storage as it could readily accept data in JSON format without much transformation. Another reason for storing the data in Mongo dB was its ability to enable selection of only required data as the datasets had data for years that were not of interest.

B. Data Preprocessing

The part of the datasets that were of interest was programmatically retrieved from the Mongo database for preprocessing. The unwanted columns that had mostly meta data were dropped and the parts of interest were renamed. Some of the data sets had the word 'county' in the county columns while others did not. To get all datasets on an even standing, the word was stripped from all the data entries. The stroke mortality dataset also had state abbreviations instead of full state names, another dataset containing the states of the US and their abbreviations was merged together with the stroke mortality rate dataset. Further observation of the dataset showed that some counties in different states had the same name. This would pose a problem to merging the dataset. To resolve this, the county column was merged with the state column to produce a column called 'StateCounty' for all the datasets. After individual preprocessing, the four datasets were merged into one on the 'StateCounty' column. A heat map was also plotted to check for missing values in the combined dataset and as shown below, there were no missing values. This was the final dataset for analysis and was programmatically stored in PostgreSQL database for easy retrieval for analysis.

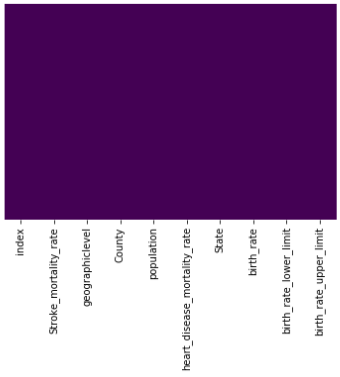

Fig. 2: Heat map checking for missing values

C. Data Exploration and Transformation

After preprocessing and subsequent storing in the database, the data was retrieved for further exploration and transformation to prepare for analysis. Correlation of

contributing predictors is an important part of a regression analysis and this was checked using a correlation heat map. The plot showed a 0.43 correlation between teen birth rate and stroke mortality rate, 0.54 correlation between heart disease mortality rate and stroke mortality rate, and -0.086 between stroke mortality rate and population. This meant that the population had a very little negative relationship with the stroke mortality rate. The independent variables were also reasonably correlated with each other but not to an extent of causing multicollinearity. A combined scatter plot were also

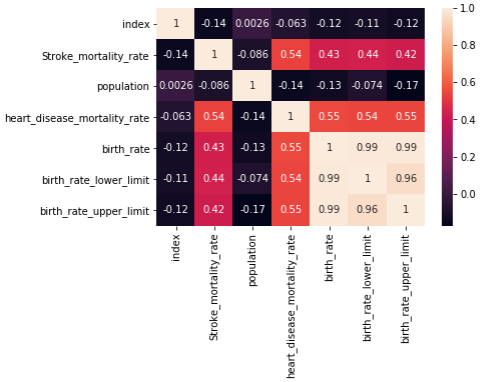

Fig. 3: Correlation heat map

used to check for the relationship between the independent variables and the dependent variable.

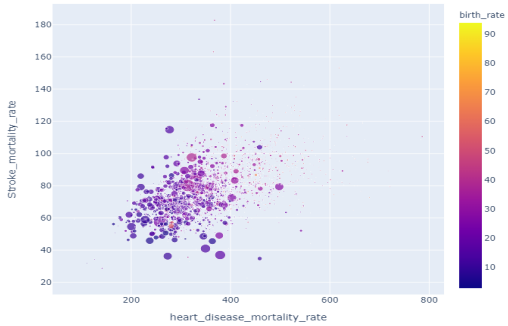

Fig. 4: Combined scatter plot

Fig 4 shows the relationship between stroke mortality rate, birth rate, population and heart disease mortality rate. The size of the bubble represents the population and the color represents the birth rate. It can be seen that there is some linear relationship between the stroke mortality rate, birth rate and population. As stroke mortality rate increases, the heart disease mortality rate increases as well. The colour also increases as the legend shows with increase in stroke mortality rate. The size of the bubble, representing population decreases as the stroke mortality rate increases. This relationship is further investigated with individual scatter plots as shown in Figs 4, 5 and 6.

The scatter plots confirm a linear relationship between the

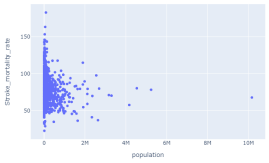

Fig. 5: Stroke mortality rate vs population scatter plot

Fig. 6: Stroke mortality rate vs birth rate plot

Fig. 7: Stroke mortality rate vs heart disease mortality rate scatter plot

variables. Another important exploration was checking the distribution of the dependent variable stroke mortality rate. Fig 7 and 8 show that the data was left skewed and had quite a few outliers. A log transformation was used on to try to correct the skewed distribution and the resulting distribution is shown in fig 9 and 10.The log transformation corrected the distribution to an extent but there were still outliers as shown in the box plot. The outliers were further investigated during analysis to avoid losing important information by removing them.The analysis was carried out with and without the outliers to guage whether they were erroneous high leverage points or actual data points within the distribution

The data was further explored to gain more information. Fig 11 shows the stroke mortality rate values for the top 10 states and it was seen the Mississippi had the highest value which was very close to Alabama. Fig 12 showed the top 10

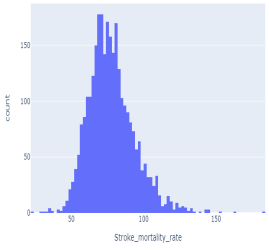

Fig. 8: Distribution of stroke mortality rate across the counties

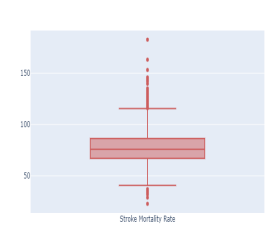

Fig. 9: Boxplot showing distribution of stroke mortality rate

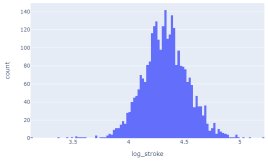

Fig. 10: Distribution of log transformed stroke mortality rate

Fig. 11: Boxplot showing distribution of log transformed stroke mortality rate

counties in terms of high mortality rates and Angelina county was the highest. The mortality rates for males and females across the counties were also compared. The males had a higher average mortality rate to the females with a value of 77.95 per 100,000 of the population while the females had an average of 75.88 per 100,000 of the population. it was also seen that Brooks county had the highest birth rate and Los angeles was the most populated county.

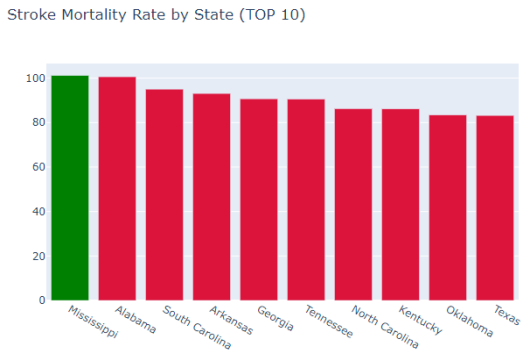

Fig. 12: Top 10 states with high stroke mortality rate.

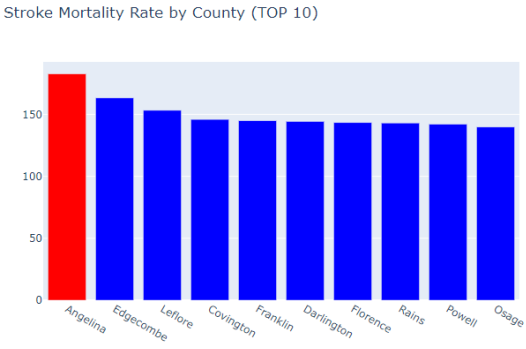

Fig. 13: Top 10 counties with high stroke mortality rate.

D. Regression

The data mining phase of this research employed multiple linear regression and decision tree regression to predict the stroke mortality rate across the counties by the population, heart disease mortality rate and birth rate using python Scikit learn library in python. Python is a flexible programming language with many libraries tailored towards handling data such as pandas and numpy. The Scikit learn library is also a very important library for data mining as it comes with most of the algorithms out of the box. It also includes evaluation metrics to check the effectiveness of the algorithms after use. Due to the need for transformation of stroke mortality rate and the presence of outliers, six different regression models were built to test the effect each of the actions. Multiple linear regression and decision trees were used to build models for the data without any transformation, the data with log transformation but no removal of outliers and the data with log transformation and outliers removed. Where transformations were done, the results were reversed back to the original values.

V. Results

A. Linear Regression without removal of outliers and transformation of stroke mortality rate

The first of the models was a linear regression model with all the outliers and also without transforming the stroke mortality rate variable. This model had a mean squared error of 175.39, root mean squared error (RMSE) of 13.24, and an R squared value of 0.35 meaning that only 35 percent of the variance in the dependent variable was explained by the predictors. To further test the suitability of the model, it was used to predict values for the test data and a visualization of the result for the first 10 data points is shown in the figure below and it can be seen that distance between the actual and predicted values were significant

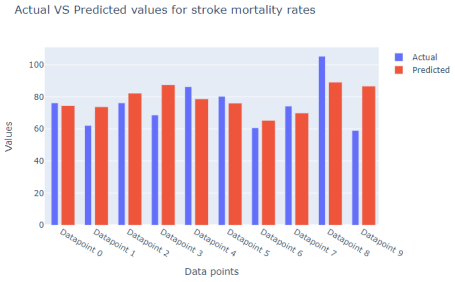

Fig. 14: Linear regression without transformation and removal of outliers prediction

B. Decision tree Regression without removal of outliers and transformation of stroke mortality rate

A decision tree was built with the same conditions and the model had a mean squared error (MSE) of 343.84 and RMSE of 18.54. Further prediction was done with model on the test data and the results are shown below.The decision tree model performed worse than the linear regression model as the predicted values were farther from the actual values.

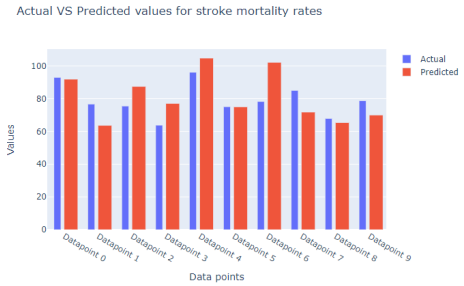

Fig. 15: Decision tree regression without transformation and removal of outliers prediction

C. Linear Regression with transformation of stroke mortality rate and without removal of outliers

The third model built was a linear regression model with outliers but the stroke mortality rate was log transformed to improve normality of distribution. As expected, this model performed better than the previous models. The evaluation metrics were reversed to their actual state due to the earlier log transformation. The model had an MSE of 1.01, RMSE of 1.00 and R squared of 0.29. This goes to show that this model had predicted values closer to the actual values than the other models built previously.

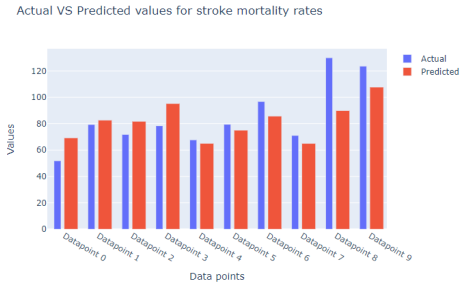

Fig. 16: Linear regression with transformation and without removal of outliers prediction

D. Decision tree Regression with transformation of stroke mortality rate and without removal of outliers

To compare with the linear model, a decision tree model was built under the same conditions and it performed slightly worse with a mean squared error of 1.02. Its predictions were farther than the actual values as shown below for the first 10 data points.

E. Linear Regression with transformation of stroke mortality rate and removal of outliers

For the final part of the analysis, a linear regression was built with the stroke mortality rate log transformed and outliers removed. This model produced an MSE of 1.01 and R squared of 0.32. The prediction was much closer to the actual values as shown in the bar chart below.

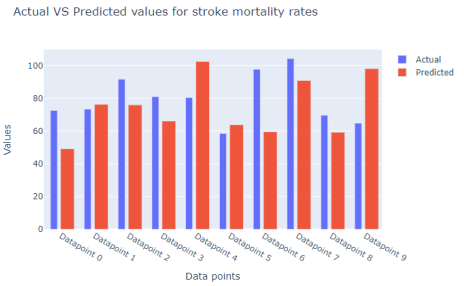

Fig. 17: Decision tree regression with transformation and without removal of outliers prediction

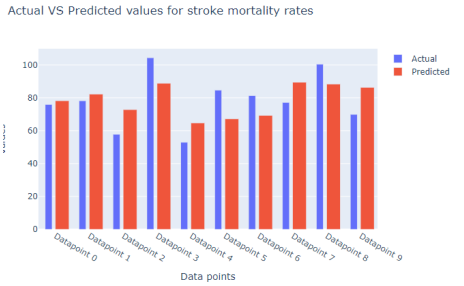

Fig. 18: Linear regression with transformation and removal of outliers prediction

F. Decision tree Regression with transformation of stroke mortality rate and removal of outliers

For direct comparison, a decision tree model was also built. This produced an MSE of 1.02 which shows that the model performed slightly worse than the linear model under the same conditions.

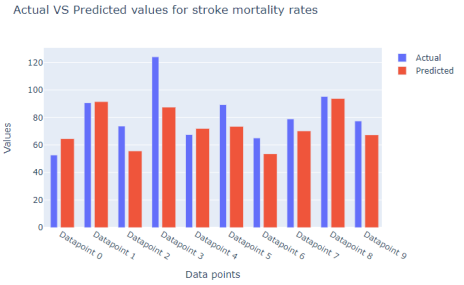

Fig. 19: Decision tree regression with transformation and removal of outliers prediction

From the results of each analysis, it can be seen that the

TABLE I: Results

Condition	Model	MSE	RMSE
without transformation and outlier removal	linear regression	175.39	13.24
	Decision tree	343.84	18.54
with transformation and without outlier removal	linear regression	1.01	1.00
	Decision tree	1.02	1.01
with transformation and outlier removal	linear regression	1.01	1.00
	Decision tree	1.02	1.01

linear regression model outperformed the decision tree model

in predicting the stroke mortality rates across all 3 conditions tested for. It is also seen that the log transformation drastically reduced the errors in the model. The outliers had no impact on the predictions as the MSE scores with and without them were approximately the same for both the linear regression and decision trees. This shows that even though they are outliers, they are not erroneous high leverage points and as such are part of the data. From the best linear regression model, the relationship between the stroke mortality rate, heart disease mortality rate, population and teen birth rate can be expressed as;

$$y = 1.69 + 1.88e^{-09}X_1 + 4.24e^{-04}X_2 + 1.33e^{-03}X_3 \quad (2)$$

where y = stroke mortality rate, X_1 = population, X_2 = heart disease mortality rate, X_3 = teen birth rate.

VI. Conclusion

In this research, machine learning was applied to predict stroke mortality rates in different counties in the US. Two machine learning techniques linear regression and decision tree regression were applied to determine which of the algorithms was a best fit for the purpose. It was found that the linear regression model was a better performing model in terms of the MSE. The impact of outliers on the predictions were also assessed and it was seen that though they were outliers, they were not high leverage points and as such, are an integral part of the data. The relationship between stroke mortality rate, population, heart disease mortality rate and the teen birth rates in these counties were also explored. It was seen that these factors also play a part in high stroke mortality rates as they are good predictors except for population which did not have as much impact as one would have thought. It further goes to show that finding a way to reduce these conditions in those counties, would also have an impact on reducing the stroke mortality rate. The negative and very minimal relationship between the stroke mortality rate and the population was an interesting find because one would think that in a highly populated place, the stroke mortality rates would be higher because there are more individuals but this wasn't the case. This could be due to other factors such as better healthcare systems or better individual enlightenment. Those are beyond the scope of this research but would be worth investigating further. The R squared of the model showed that the selected predictors only accounted for 30 percent of the variance in the dependent variable. This suggests that more predictors could have been added for better results. Investigating the impact of economic indices such as unemployment rate and average level of income in addition to the selected predictors on stroke mortality rate to expand the scope would have been ideal. For future work, different machine learning algorithms can also be applied to check their performance against the ones used in this research.

Appendix

TEAM WORK DISTRIBUTION

All 3 team members participated equally in the project work load remotely and in the library. The distribution of work was done on the capabilities and skills of the member. Each team member retrieved and cleaned a data set before they were all merged into one for further exploration and analysis. After that, all 3 team members partook in the report writing part of the project by writing different sections. All team members were aware of each other's work and made contributions where necessary leading to successful completion of the project. A group chat, regular meetups in the study room were set up during the course of the project. Each team member contributed 33.33 percent to the project.

References

[1] B. Lantz, *Machine learning with R*, 3rd ed. Packtz, Birmingham UK, 2019.

[2] K. Kim, J. Y. Choi, T. K. Yoo, S. K. Kim, K. Chung, and D. W. Kim, "Mortality prediction of rats in acute hermorrhagic shock using machine learning techniques", *International Federation for Medical and Biological Engineering*, vol. 51, pp. 1059-1067, June 2013.doi:10.1007/s11517-013-1091-0

[3] C. Ma, C. Chao, and B. Cheng, "Predicting patients at risk of acute renal failure in intensive care units by using artificial intelligence tools", *International Journal of Organizational Innovation*, vol. 5, no. 2, pp. 232-247, Nov. 2012. [Online]. Available:https://ezproxy.ncirl.ie/login?url=https://search.proquest.com/docview/1095773461? accountid=103381 [Accessed on: December 12, 2019].

[4] G. C. Rafael, B. P. Oscar, M. J. Inmaculada, S. R. Christina, and G. E. Rebecca, "Identification of clinically relevant features in hypertensive patients using penalized regression: a case study of cardiovascular events," *Medical and Biological Engineering and Computing*, vol. 57, no. 9, pp. 2011-2026, Sep. 2019. doi:10.1007/s11517-019-02007-9

[5] M. A. Jabbar, B.L Deekshatulu, and P. Chandra, "Alternating decision trees for early diagnosis of heart disease," *in 2014, Int. Conf. on circuit, communication, control and computing, Bangalore, India, Nov. 21-22, 2014*, pp. 322-328.[Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=7057816 [Accessed on Nov 10,2019].

[6] R. Jeena, and S.Kumar, "stroke prediction using SVM," *in 2016, Int. Conf. on control instrumentation, communication and computational Technologies, Kumaracoli, India, December 16-17,2016*, pp 600-602.[Online]. Available: IEEE Xplore, https://www.ieee.org/ [Accessed on Nov 20,2019].

[7] L. Temza, J.L. Risco-Martin, J.Ajala, G.R Rosello, and J.M. Camarasaltas,"Comparison of different machine learning approaches to model stroke subtype classification and risk prediction," *in 2019, Spring Simulation Conf. , Tucson, USA, April 29- May 2,2019*. doi:10.23919/SpringSim.2019.8732846

[8] J.N. Heo, J.G Yoon, H.Park, and Y.D Kim, "machine learning based model for prediction of outcomes in Acute Stroke,"*AHA journals*, vol. 50, pp 1263-1265, May 2019. doi: 10.1161/STROKEAHA.118.024293.

[9] T. Kansadub, S. Thammaboosadee, S. Kiatissin, and C. Jalayondeja, "Stroke risk prediction model based on demographic data," *in 2015, 8th Biomedical Engineering Int. Conf., Pattaya, Thailand, Nov. 25-27, 2015*. doi:10.1109/BMEiCON.2015.7399556

[10] P. S. Kohli, and S. Arora, "Application of machine learning in disease prediction," *in 2018, 4th Int. Conf. on Computing Communication and Automation, Greater Noida, India, Dec. 14-15, 2018*. doi:10.1109/CCAA.2018.8777449