GROUP 4 ASSIGNMENT- FINAL REPORT

Code

1. OVERVIEW OF THE PROBLEM

Traumatic injuries can produce both acute and more chronic consequences that lead to permanent disabilities, increase long-term mortality and reduced life expectancy. Patient attributes, injury characteristics and treatment interventions are the key factors in determining short-term survival and clinical outcomes of the patient.

The aim of this project is to apply machine learning algorithms to predict clinical outcomes for patients with severe bleeding from trauma, in order to inform clinical decision making in the hospital setting. In particular, we seek to build a classification model that will identify whether a patient is likely to make full recovery, require care and rehabilitation or at risk of dying. Intention of the project is identify patients who will become dependent after their injury and will require extensive care and patients who may not survive in the course of treatment. This will assist hospitals to organize appropriate rehabilitation and counselling resources for the patient and their kin in a timely manner. The project will compare the results of several classification approaches including Random Forest, K-Nearest Neighbour, Support Vector Machines, Bagging and Linear Models(Multinomial and Ordinal) to find the optimal model to predict results.

2. DATASET DESCRIPTION

This project presents the analysis of the CRASH-2 (Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage) study data. The CRASH-2 consists of 20,207 adult trauma patients with, or at risk of significant bleeding who were generally within 8 hours of injury (1).

Code

Code

Code

3. INITIAL DATA ANALYSIS AND VISUALISATION OF THE DATA

Initial data Analysis revealed that the CRASH-2 dataset had 10% NA values with at-least one NA in every row. The trauma patients had been classified into six different categories - “no symptoms”, “minor symptoms”, “some restriction in lifestyle but independent”, “dependent, but not requiring constant attention”, “fully dependent, requiring attention day and night” and “dead”. Our study, however, is concerned with identifying if a trauma patient after treatment dies, has a full recovery or recovers but is still dependent on others. In order to achieve this, we have reclassified the six categories into 3 main categories - ‘dead’, ‘independent’ and ‘dependent’.

Code

Code

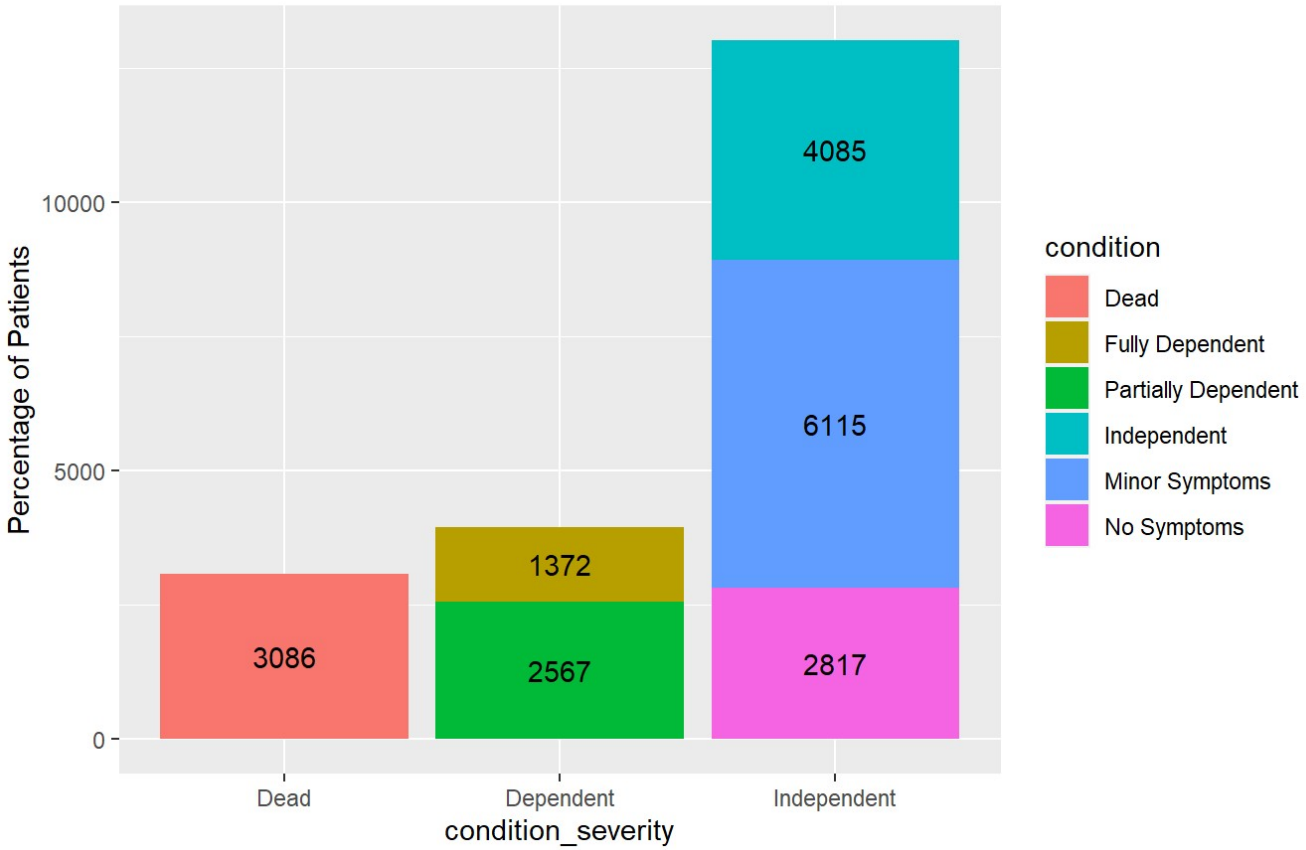
Code

Code

Code

Code

FIGURE 1
Patients Across Outcome Variables



Code

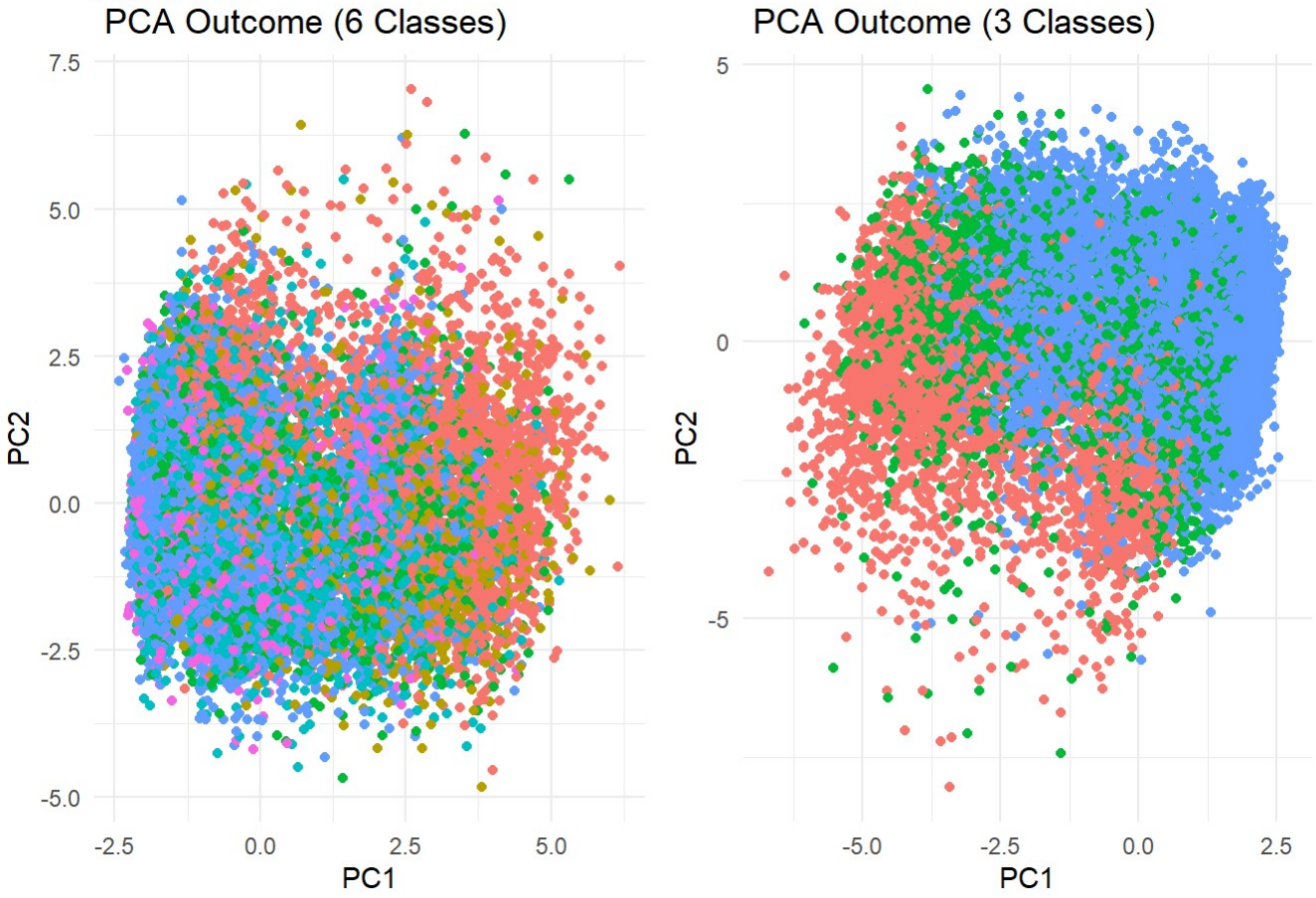
Code

Code

PRINCIPAL COMPONENT ANALYSIS

Code

FIGURE 2

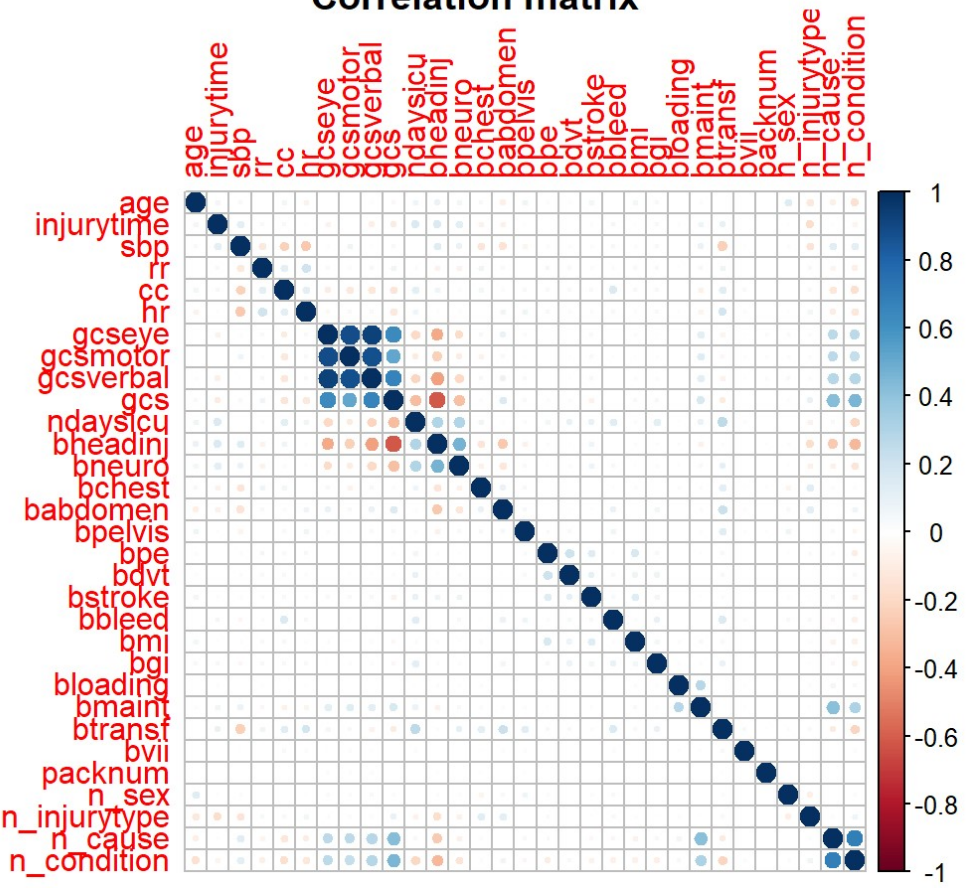


CORRELATION

From the correlation plot, it can be seen that there is a high correlation between gcs and gcseye. This is because gcs (glasgow coma score) and gcseye(glasgow coma score eye opening) have high correlation since gcseye is a subpart of the gsc test. The plot also revealed that there is good correlation between bneuro(Neuro surgery done) and bheadinj(severe head injury) which is expected as patients with severe head injuries are highly likely to have neuro surgeries performed.

Code

FIGURE 3
Correlation matrix



Code
Code
Code
Code
Code
Code
Code
Code

After IDA, data cleaning and re-classification of the condition parameter; the CRASH-2 dataset was split into 3 subsets- Training data (10864 rows), Validation Data (3621 rows) and Testing Data (3619 rows). The three datasets have 32 columns. The data across the 3 classification parameters was imbalanced. Because balancing the dataset before fitiing a model biases the model and throws out potentially useful data, we focused on using a **balanced accuracy** metric to evaluate the best predictive algorithm.

Code
Code

4. FEATURE ENGNEERING

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. Three methods- Step-wise selection, MARS(Multivariate Adaptive Regression Splines), and Random Forest have been used to perform feature selection.

4.1 STEP-WISE AIC

Bi-Directional Step-wise variable selection was run with Multinomial Logistic Regression to identify features that improve model performance based on AIC. AIc was selected to penalize for model complexity to avoid over-fitting. 7 features were identified to be removed from the model - sex, injury time, bvii, platelets, heart rate, gcs verbal and DVT.

Code

4.2 MARS(MULTIVARIATE ADAPTIVE REGRESSION SPLINES)

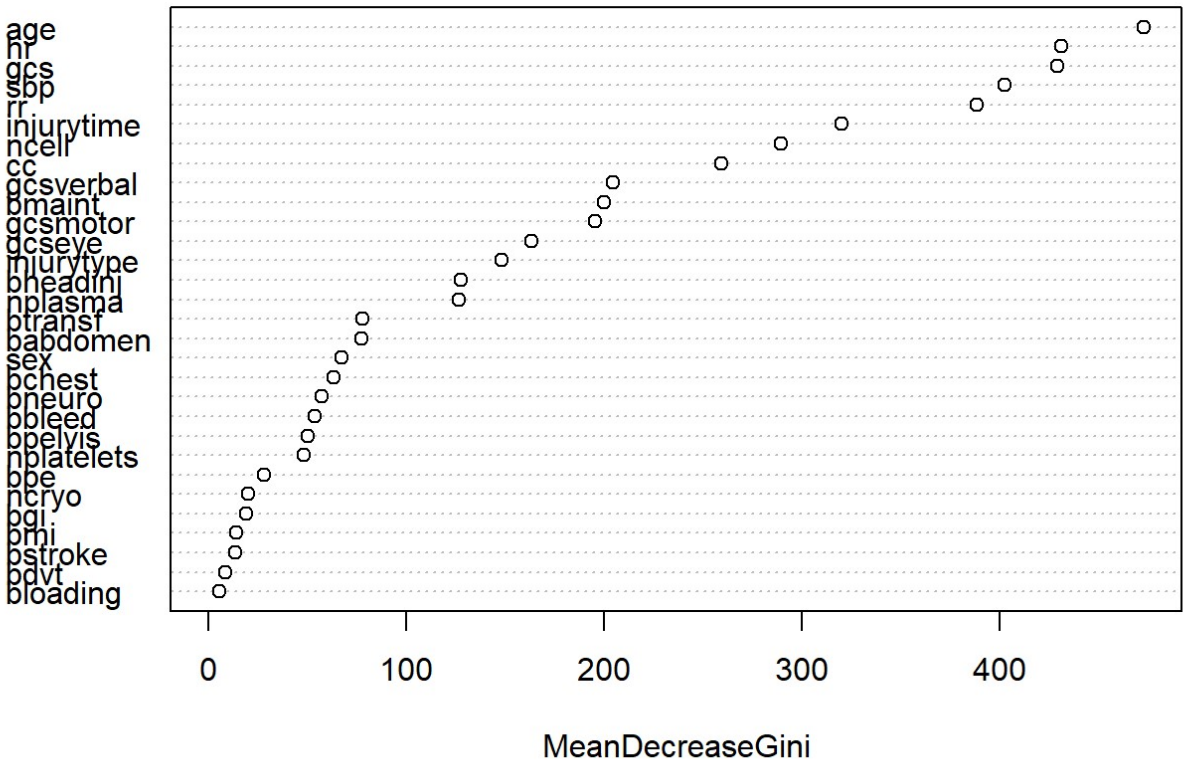
MARS is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables. The top features identified from this method were - gcs, bmaint, ncell, injurytypepenetrating, age, babdomen, rr, bpe, bheadinj, sbp, bpelvis, bstroke and nplatelets.

Code

4.3 RANDOM FOREST

Code

FIGURE 4
Random Forest Feature Engineering



Each feature selection model generated slightly different results. Ultimately, the feature engineering model used, was based on the classification algorithm being trained because each classification algorithm performed better on varying attributes.

5. CLASSIFICATION ALGORITHMS USED

5.1 LINEAR MODELS

Various Linear Models were examined to identify the performance of these models on the classification task. Multinomial Linear Regression was used as a base model. Features identified by both stepwise AIC

and random forest were tried. Stepwise AIC was shown to have better performance as expected. We have also examined ordinal linear regression as our outcome variable has an inherent order to its classes. However, the ordinal model did not show good results. This could be due to the variance within the dependent class which would not satisfy some of the assumptions of ordinal model. The best results were selected from multinomial regression with stepwiseAIC and provided in the model comparison.

Code

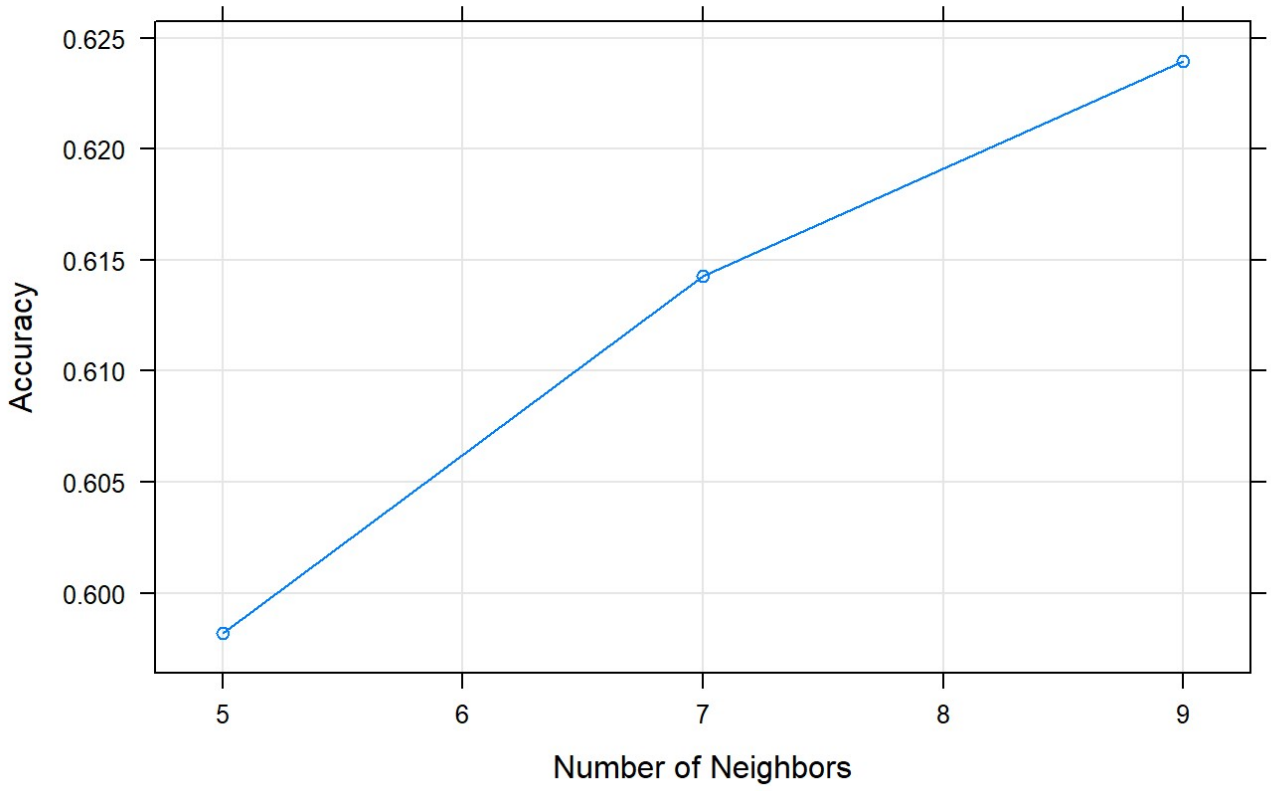
5.2 K-NEAREST NEIGHBORS

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. k-NN is a type of lazy learning, where the function is only approximated locally and is sensitive to the local structure of the data. 25 features selected by step forward method were used to implement k-NN. The categorical values were converted to numerical values and the data was normalised to further improve the balanced accuracy. Optimal value of k was determined by using the train function from caret class to train the algorithm on a number of k values.

Code

Code

FIGURE 5
OPTIMAL VALUE FOR k-NEAREST NEIGHBORS



Based on this plot, k=9 was chosen to classify the condition of patients.

Code

Code

5.3 RANDOM FOREST

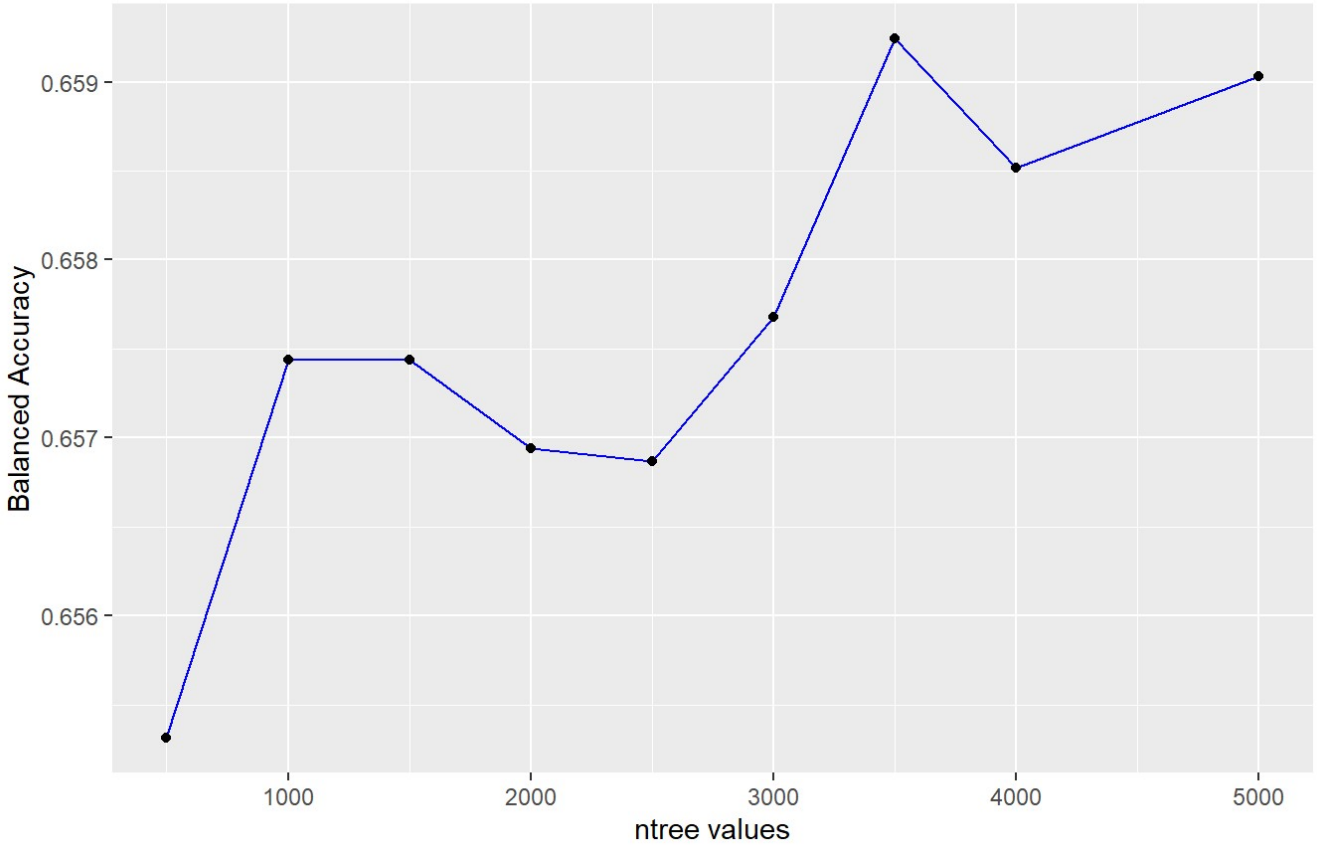
Random Forest is a robust machine learning algorithm that uses a large number of small decision trees, called estimators, to produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction. Performance of Random Forest was evaluated on

"Validation Dataset" using different values of "ntree". The best ovaerall balanced accuracy was obtained for ntree = 3500. The plot below shows the overall balanced accuracy for different values of ntree.

Code

Code

FIGURE 6
Random Forest ntree Performance



Code

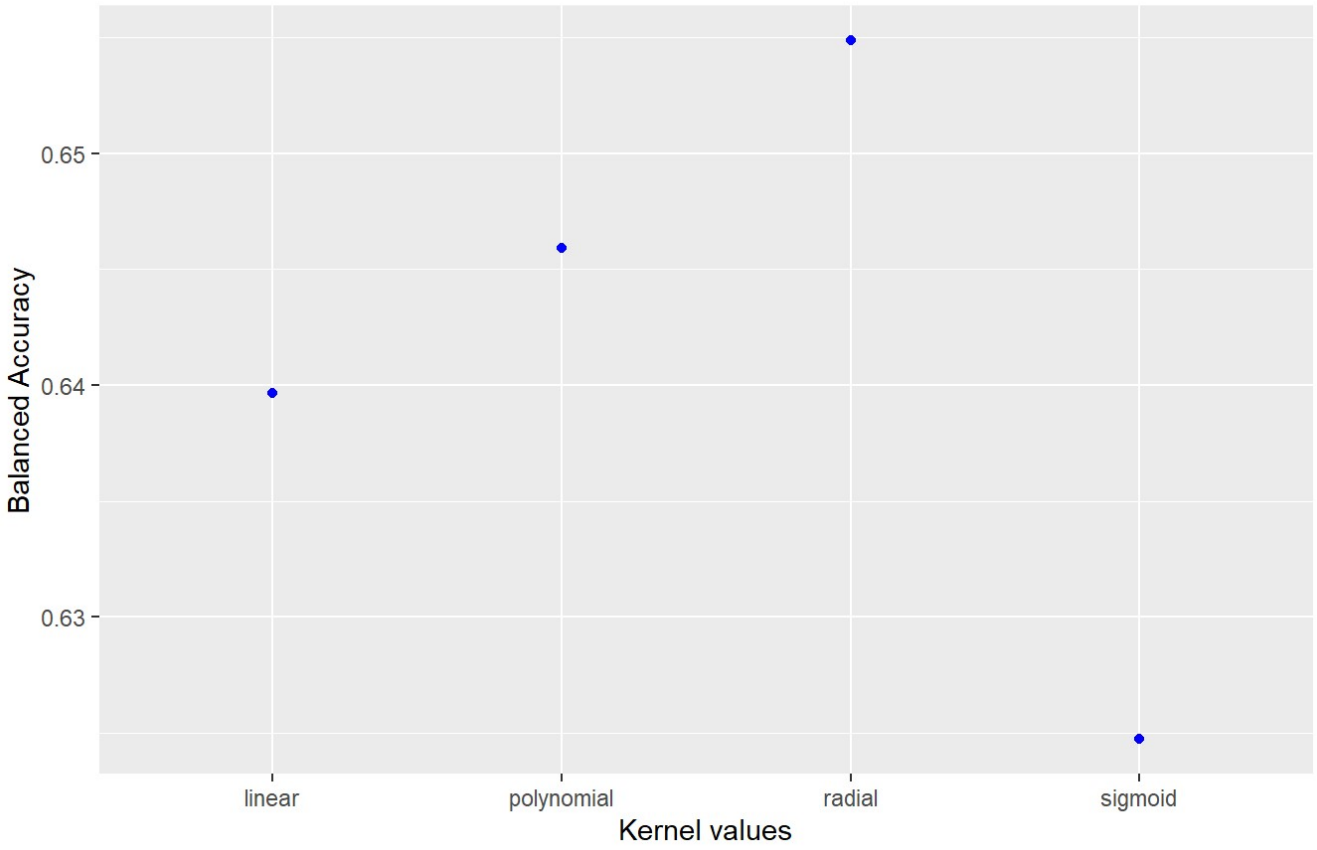
5.4 SUPPORT VECTOR MACHINES

Support Vector Machines is a powerful Machine learning algorithm which aims to find a hyperplane or a decision boundary in an N-dimensional space that distinctly classifies the data into accurate classes. Implementation of SVM on the given dataset was tried using different available kernels, and it was observed that the overall balanced accuracy was best for the "Radial Basis" when tested against "Validation Dataset". The plot shows performance of SVM using different kernels. SVM was trained using 10 fold cross validation which was achieved by specifying value of input parameter "cross" as 10.

Code

Code

FIGURE 7
SVM kernel Performance



Code

5.5 BAGGING

Bagging(bootstrap aggreagation) is used as one of the mode for this classification task. This models is trained using the features identified by the stepwise feature selection and also using all the features. But the bagging with and without stepwise features selection is giving very similar results. So we have decided to use all the features instead of doing any feature selection. The results were as follows

Code

Code

6. CLASSIFICATION PERFORMANCE ALGORITHM

The table below lists performance all the classification models implemented as a part of this project. It specifies performance on different metrics for each class. It also includes overall balanced accuracy for all the models. From the table it can be observed that almost all the models, except for KNN, performed similarly or equally when compared on the basis overall balanced accuracy. Performance of Bagging has been the best, and is marginally better than Linear Model, SVM and Random Forest.

Code

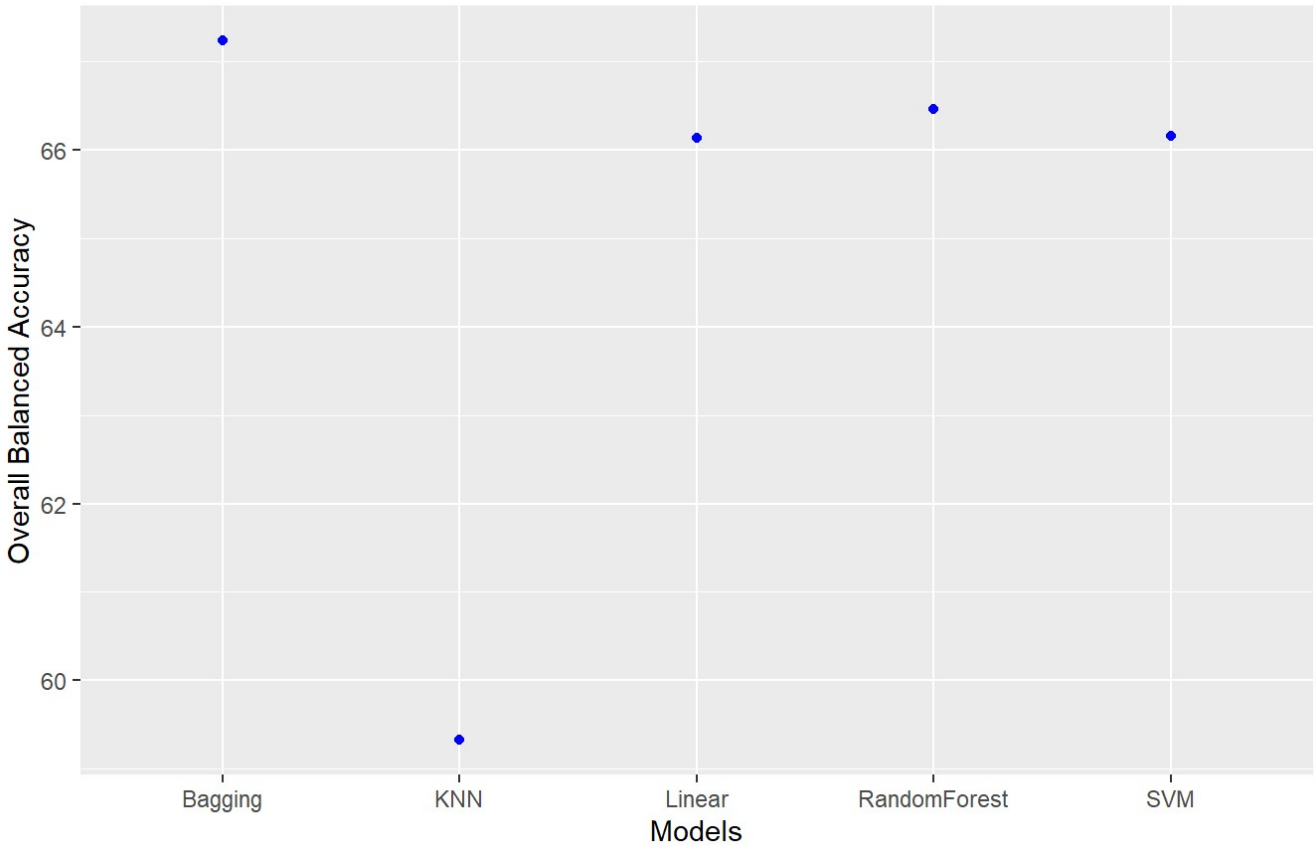
FIGURE 8 : Performane of Models Table

Metric	KNN	Linear	SVM	RandomForest	Bagging
Sensitivity for Class Dead	31.215	56.157	58.769	55.970	52.985

Metric	KNN	Linear	SVM	RandomForest	Bagging
Specificity for Class Dead	95.493	94.392	94.716	95.300	94.522
Balanced Accuracy for Class Dead	63.354	75.274	76.743	75.635	73.753
Sensitivity for Class Dependent	10.249	11.757	8.990	12.863	24.066
Specificity for Class Dependent	93.856	96.446	97.481	96.480	90.476
Balanced Accuracy for Class Dependent	52.053	54.101	53.236	54.672	57.271
Sensitivity for Class InDependent	93.310	94.960	95.850	95.600	89.460
Specificity for Class InDependent	31.820	43.130	41.140	42.570	51.950
Balanced Accuracy for Class InDependent	62.570	69.050	68.500	69.090	70.700
Overall Balanced Accuracy	59.325	66.141	66.159	66.465	67.241

Code

FIGURE 9
Performance of Models Plot



7. CONCLUSION

This study was undertaken with the objective to classify trauma patients into 3 main catedories-'Dead', 'Dependent' and 'Independent'. This classification would help the doctors all over the world to make informed clinical decision making when treating the patients. Five classification algorithms were implemented whose balanced accuracy varied from 59.26% to 66.65%.

Attempts were also made to increase the accuracy in predicting the patients who were likely to be 'Dead' or

'Dependent' in the future. These prediction could be used to organise appropriate rehabilitation resources for the patient and their families in a timely manner.

One of the reasons for the comparitively lower accuracy can be attributed to the lack of previous medical history for the trauma patients. In a real world setting, a patient's previous health conditions do impact the probability of them surviving any trauma, which the CRASH - 2 dataset did not consider.

Another reason is due to the class imbalance especially in dependent class that are harder to separate from dead and independent. SMOTE was used to handle for this but showed minor improvements and we avoid it to avoid overfitting on the noise in the imbalanced classes.

Due to the lower accuracy, the current evaluated models may not be useful for decision making in hospital settings. Further attempts will be undertaken in the future to improve this accuracy by working on the data imbalance and implementing some more classification algorithms.

8. AUTHOR'S CONTRIBUTIONS

Aakanksha Jain(312153430)

Worked on the plan and exploratory data analysis. From this work the missing value analysis and correlation analysis was used in the final report. Also assisted in removing redundant features from the original data frame. The rest of the work was not used as other member had better graphical representations of the EDA. Additionally, assisted in editing the final EDA report with the rest of the team. Assisted in editing the final presentation with the rest of the team. Worked on writing the final report, specifically on the overview of the problem, dataset description, correlation, and feature engineering.

Aditya Deshpande(500262382)

I must mention that it was a collective effort from everyone. In fact, after a long time I had an opportunity to work with some responsible, sensible and mature people. I would like to thank everyone for their efforts. My contribution involved initial data analysis and data manipulation, specifically in terms of filtering and organizing the data, including splitting the data into Training, Test and Validation dataset. The intention was to get the data in the format so that it can be directly used to train the models. Also, I managed to implement Random Forest and SVM models for classification.

Amit Desai(500589441)

For the final group project we worked collectively and collaborated together. For the initial EDA we all put our efforts. For my part I did some descriptive statistics using R code to calculate statistical information like number of male/female patients, type of injuries among our sample dataset. For feature engineering I suggested and applied Random forest algorithm to derive variable importance and also tried Logistics Regression. For the final part I had taken the responsibility of Presentation and created our group presentation. I also recorded the video for the final presentation.

Anirudh Bhat(490492189)

I started working on the project a bit late so it took some time to get updated on others contribution upto that point. I analysed the dataset and the pre-processing had already been done. Since there were 5 models in total and each of them was assigned to one team member, I did not work on the model. However I gave advice and made necessary changes to each of the models to perfect them and give the required results. Happy with the team as well as my individual efforts.

Geogy Sabu Jose(490556492)

Worked on EDA for re-classifying the outcome variable, stepwiseAIC for feature selection and Linear Models (Ordinal and Multinomial) for classification.

Jhanvi Gupta(500249305)

Worked on the Introduction and Overview part of exploratory data analysis. Additionally, worked on feature extraction where I implemented MARS(Multiple Adaptive Regression Splines). For the final report, I was responsible for compiling the work by all my team members, structuring the report(which involved aligning it, adding captions to images, creating the basic template for the report and writing about our dataset description, IDA, Models implemented and conclusion).

Rana Hamad Khan (500215676)

I worked on the first stage of the project which was EDA and I did descriptive statistics using R to calculate statistical information like whats type of injuries among male and female and their age distribution and also how many days spent in ICU by different types of injuries. All these information shown in histogram with different color distribution. And we divided different ML models among team members and I was working on KNN. My primary task was to train the model on the best feature which impacts the model and find the optimal value of K using different techniques. Other than that I tested my model accuracy on the validation data set and presented the balance accuracy of each class, sensitivity, specificity and overall balance accuracy of my model on the validation date set.

Shabari Gadewar(490607392)

I did some of the initial data analysis, to be specific, find correlation among different parameters and created correlation heatmap. Also did some feature engineering to find the most important parameters. for classification, I used Bagging for modeling.