CSCI 795 Machine Learning Project Report

SURVIVAL PREDICTION OF LUNG-CANCER PATIENTS AFTER ONE YEAR OF THORACIC SURGERY

Health Risk Prediction

Manal Zneit
Sabina Bhuiyan

Team G5
Fall 2020

Contents

GITHUB LINK: https://github.com/mZneit/Survival-prediction-of-lung-cancer-patients

VIDEO: https://youtu.be/tQFzEOe4ZHY

Project Description

Problem Description/Goals

The field of Machine learning has recently infiltrated the medical domain through a wide array of applications. The growth of information and data on one hand and the advancement in medicine on the other hand, gave rise to significant breakthroughs in medical diagnosis, prognosis, pharmaceutical industry, and other areas… Given the advantages of the qualitative and quantitative tools provided by ML techniques in the medical field, however, there are certain considerations related to medical ethics that address the impacts of ML in making healthcare decisions. Equally concerning is the public trust in ML technology, where difficulties faced by AI and ML systems pose serious challenges when it comes to avoiding patients' harms and assuring their safety and privacy. Such harms have the potential to undermine public trust in these technologies that will in turn adversely impact the emergence and acceptance of AI/ML systems in healthcare and medicine. These harms can take different forms: physical, medical, racial biases, genetic discrimination, psychological, economic, and social inequalities.

This project addresses one form of ML applications in medicine. The aim is to apply ML technology in risk prediction through the build of ML models. These models will predict the risk of mortality of lung-cancer patients one year after thoracic surgery. As a matter of fact, lung-cancer is the leading cause of death worldwide according to the World Health Organization (WHO). The main reason associated with lung-cancer is smoking and exposure to second-hand smoke leading to more than 80% of the total lung-cancer deaths globally.

Team Member Roles

Manal Zneit – implemented the ML algorithms, built, implemented, and evaluated predictive models.

Sabina Bhuiyan – designed the visualization tool for comparison and analysis of the models, ranked risk factors and made the video. Sabina also participated in designing some predictive models as shown in her report.

State-of-the-art/Related Work

Risk prediction has been the subject of many research studies. Though not limited to the subject matter of this project, building ML models to predict the risk of mortality is common in the examination of postoperative complications as well as assessing the risk factors of certain diseases. In [2], a dataset of heart failure examination records was used to implement ML models and predict the risk of patients' survival. The methodology also performed feature selection and concluded that two features were sufficient to perform accurate predictions instead of using the entire dataset. Some research papers applied data mining techniques on the thoracic dataset to extract the most important variables in the incidence of lung-cancer. The results indicate that large tumor size was the leading factor in the mortality of lung-cancer patients. Moreover, the challenges of the dataset under study is the imbalanced class distribution of the target vector. To overcome the imbalanced data problem, [1] proposed a boosted SVM model for survival prediction using an oracle-based approach to extract rules to solve the imbalanced data problem. In [5,6], a comparative analysis of several ML models is conducted on different types of cancer. A general discussion of the performance was

provided. In [5] Random Forest was the best classifier, in [6] a comprehensive analysis was provided that depends on the cancer type and the associated dataset.

Approach

Several ML algorithms are implemented to predict the post-operative risk of mortality among lung-cancer patients. The problem is a classification problem and different classifiers were implemented and evaluated using multiple performance metrics. Seven classifiers were analyzed, and their performance measures were compared - KNN, Naïve Bayes, SVM, ANN, Logistic Regression, Decision Trees, and Random Forest. Ultimately, an ensemble learning algorithm (a voting algorithm) was utilized near the end of the project. It is a supervised learning task and the classifiers vary in their implementation (instance vs model based, parametric vs non-parametric, generative vs discriminative models). The approach is to implement a variety of predictors that are independent and diverse as possible, and then combine the predictions using a majority voting ensemble to generate a better predictor.

There are several sampling techniques that handle the imbalanced dataset problem. Clinical datasets may have missing values due to incomplete questionnaires or missing records at random or not at random; the dataset may also be unstructured due to the prevalence of noisy data. Missing values usually introduce an element of bias resulting in invalid results. Several sampling techniques can reduce the impact of missing data by either replacing missing values without introducing bias (imputation), or by eliminating variables with missing values. The dataset utilized in this project demonstrates an imbalance in the class distribution of the target vector where the positive instances are under-sampled, and the majority class (negative instances) introduced a bias when

training the ML algorithms (learning from one class). Rebalancing the dataset can be done by either oversampling the minority class or under sampling the majority class. Another approach is by penalizing the errors and inaccurate predictions of the minority class. However, a widely used technique known as SMOTE (Synthetic Minority Oversampling TEchnique) synthesizes data from the existing samples in the minority class in order to solve the imbalanced data problem. The technique uses randomization to generate data that lie close to the original data in the multidimensional feature space. A comparative analysis is also performed on the ML algorithms before and after the application of SMOTE on the imbalanced dataset.

Dataset

Source:

Creators: Marek Lubicz (1), Konrad Pawelczyk (2), Adam Rzechonek (2), Jerzy Kolodziej (2)

(1) Wroclaw University of Technology, wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland

(2) Wroclaw Medical University, wybrzeze L. Pasteura 1, 50-367 Wroclaw, Poland

Generated: November, 2013

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007 - 2011.

Number of instances: 470

16 features and 1 target vector (a binary-valued vector representing patient died (T) or survived (F)). A meaningless feature that represents the ID number of each patient was later dropped from the set of features. (10 binary attributes and 1 binary class label, 3 nominal and 3 numerical attributes.)

Missing attribute values: None.

Class distribution: 70 T (14.89% died), 400 F (85.11% survived).

Features	Description	Values (Nominal/Numeric)
DGN	Diagnosis-specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any	DGN3,DGN2,DGN4,DGN6,DGN5, DGN8,DGN1
PRE4	Forced vital capacity - FVC	Numeric
PRE5	Volume that has been exhaled at the end of the first second of forced expiration - FEV1	Numeric
PRE6	Performance status - Zubrod scale	PRZ2,PRZ1,PRZ0
PRE7	Pain before surgery	T, F
PRE8	Haemoptysis before surgery	T, F
PRE9	Dyspnoea before surgery	T, F
PRE10	Cough before surgery	T, F
PRE11	Weakness before surgery	T, F
PRE14	T in clinical TNM - size of the original tumor, from smallest to largest	OC11, OC14, OC12, OC13
PRE17	Type 2 DM - diabetes mellitus	T, F
PRE19	MI up to 6 months	T, F
PRE25	PAD - peripheral arterial diseases	T, F
PRE30	Smoking	T, F
PRE32	Asthma	T, F
AGE	Age at surgery	Numeric
Risk1Yr	1 year survival period - (T)rue value if died (T,F)	T, F

Experiments/Evaluation

The data is prepared through Exploratory Data Analysis (EDA) and the most correlated features are identified. Meaningless features such as the ID of the patient and redundant features were dropped leaving out the features that add the most significance to the classification task. The values of the categorical attributes were mapped to an integer representation using one hot encoding. The multinomial Naïve Bayes algorithm is applied together with the Gaussian NB for comparison and highlighting the multinomial NB since the attribute values are mostly categorical. KNN is trained and model selection was performed to tune the hyperparameters. The optimal model was trained and applied to the test data. Each of the models was trained and the optimal model was selected using k-fold cross-validation. The ROC curves of both the training and the test sets were generated as well as the AUC scores. Throughout the code, splitting the data into training and test sets takes into consideration the stratify parameter that balances the classes of the instances during the split. Moreover, the k-fold cross-validation splits the training set into stratified k-folds (80% training and 20% test data). These implementation details improved the performance of the classifiers (before and after the application of SMOTE) unlike when the stratify parameters were not chosen in the split of the data into training and test sets as well as in k-fold cross-validation.

Since this is a medical dataset, the models were optimized based on high recall and the aim was to generate models with high sensitivity and specificity. In addition to the ROC curve and the AUC scores, other evaluation metrics used are: confusion matrix, precision, recall, F1 score, and accuracy. The classification report was also generated

to visualize the performance metrics on both macro-average and weighted-average where the support representing the class distribution was the weight used to find the mean per label. The macro- and weighted- averages provide an indication of the effect that an imbalanced dataset can have on the performance measures of a classifier.

Discussion of results

The performance of the classifiers was compared based on the AUC performance metric. The best performance was achieved by SVM (94%) followed by ANN (87%) and Random Forest (87%). DT and Multinomial Naïve Bayes had the lowest AUC values (64% and 72%) and NB was the weakest classifier in terms of precision, recall, F1 score, and accuracy. The reason behind the weak performance of Naïve Bayes is the simplifying assumption that presupposes conditionally independent features given the class label. In this dataset, it is not reasonable to assume the independence of features since the complications that a patient may experience are not simply independent as NB assumes. NB algorithm assumes the case that a patient being a smoker will not give any further information about coughing or breathing difficulties before surgery, which is an inaccurate assumption to make in this task. NB is a task-dependent classifier, and it is not an accurate predictor for this task.

Decision trees are hierarchal models that perform a sequence of tests to achieve a decision at the leaf level. The division of the problem into branches is based on the “most important” attribute that makes the most significant difference in classification. For this task, DT had the lowest performance measure (64% AUC score for the test data) even though the algorithm had a far better AUC score for the training data (80%) and this is due to the sample size. DT is a non-parametric model that is prone to overfitting

when the sample size is too small. The thoracic dataset is small and it affected the split points in the tree as well as depth of the tree and the final decisions at the leaf level. DT depends on the size of the data in the classification problem and it overfits for small datasets.

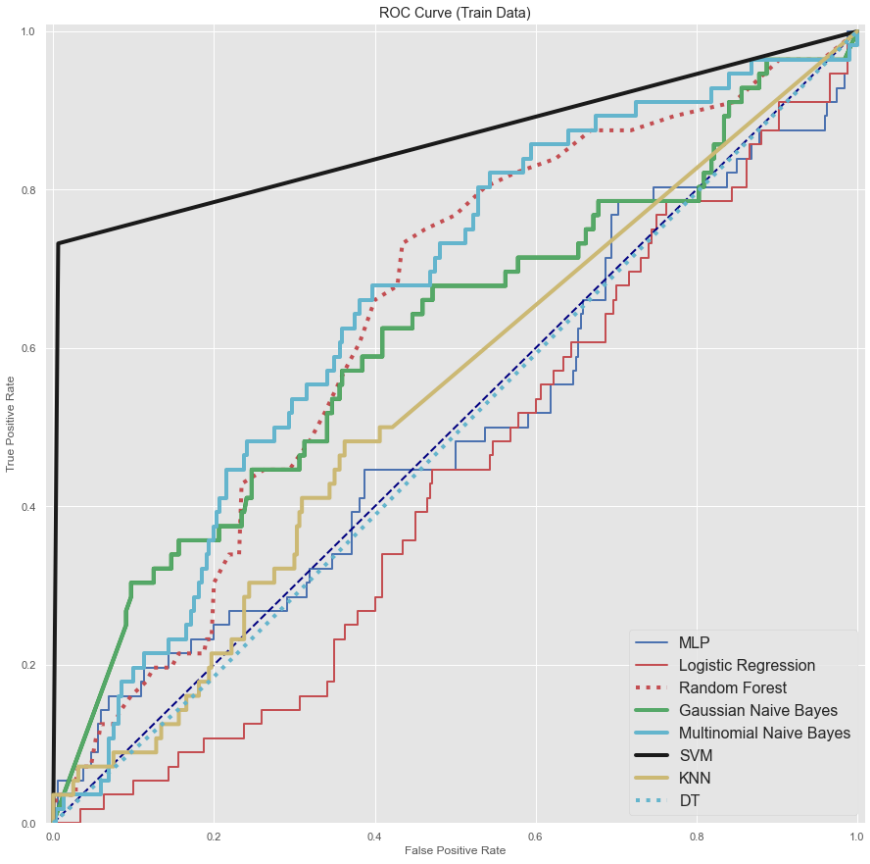
Similar to DT, KNN is a non-parametric model that depends on the data distribution. The AUC score is 82% but with better precision, recall, F1 score and TPR than DT. The tuned hyperparameters generated an optimal model with number of neighbors equals 3. It is an instance-based model that performs better on larger datasets to avoid overfitting (the AUC of the training set is 92%).

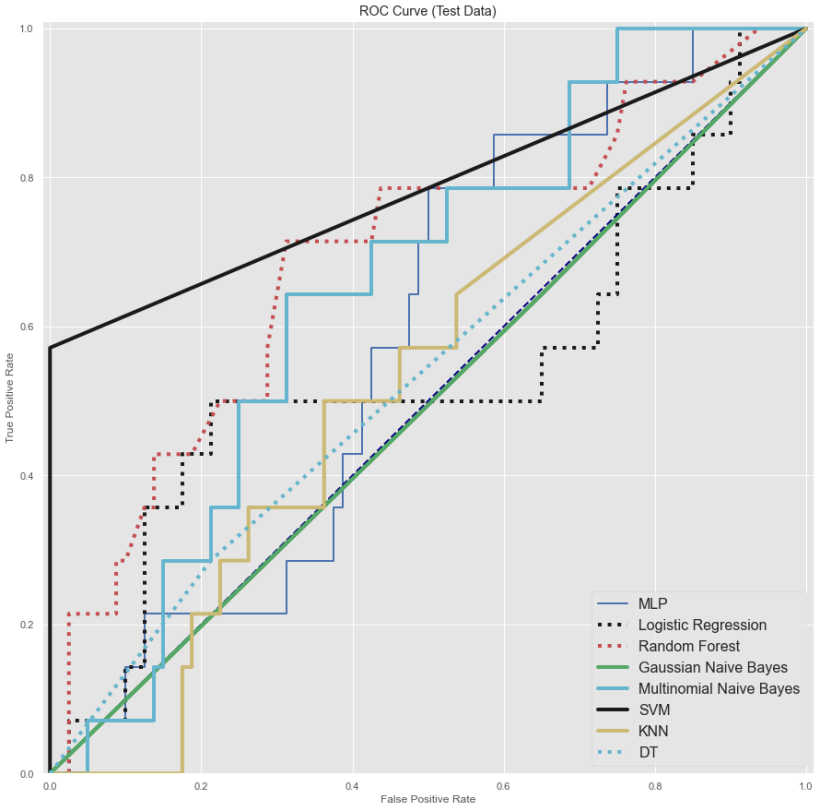
Random Forest, an extension to DT by ensemble techniques that use multiple DTs, had a relatively high performance compared to other classifiers. It minimized the overfitting examined by single trees and obtained the decision of multiple individual trees trained on random subsets of the training set. Logistic regression demonstrated a similar performance to Random Forest yet with a bit lower measures.

SVM and ANN had the best performance (94% for SVM and 87% for ANN). ANN performed well on all metrics. At the end of the project, an ensemble classifier was implemented to generate a majority vote among the algorithms. This classifier performed well on test data and it was a technique that added diversity in classification to the overall predictive task. The performance of all the classifiers was far better after applying the balancing technique SMOTE. The following graphs show this improvement for both training and test data.

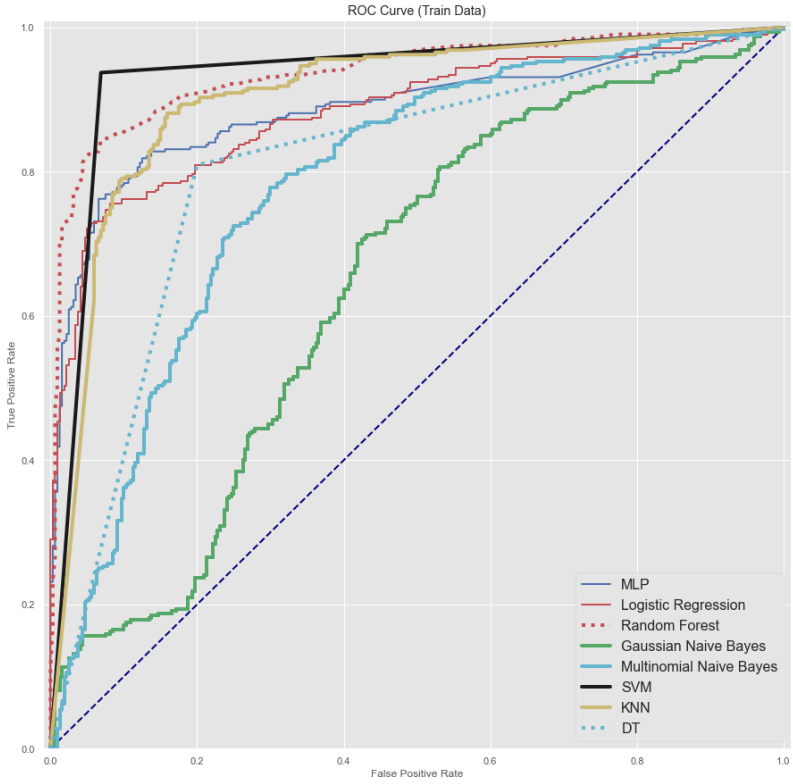
Model	Accuracy	F1 Score	Precision	TPR	AUC
Random Forest	0.89	0.89	0.89	0.89	0.87
SVM	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
ANN	0.91	0.91	0.91	0.91	0.87
DT	0.81	0.81	0.81	0.81	0.64
KNN	0.86	0.86	0.87	0.86	0.82
Logistic Regression	0.86	0.86	0.86	0.86	0.83
Multinomial Naïve Bayes	0.72	0.71	0.75	0.72	0.72

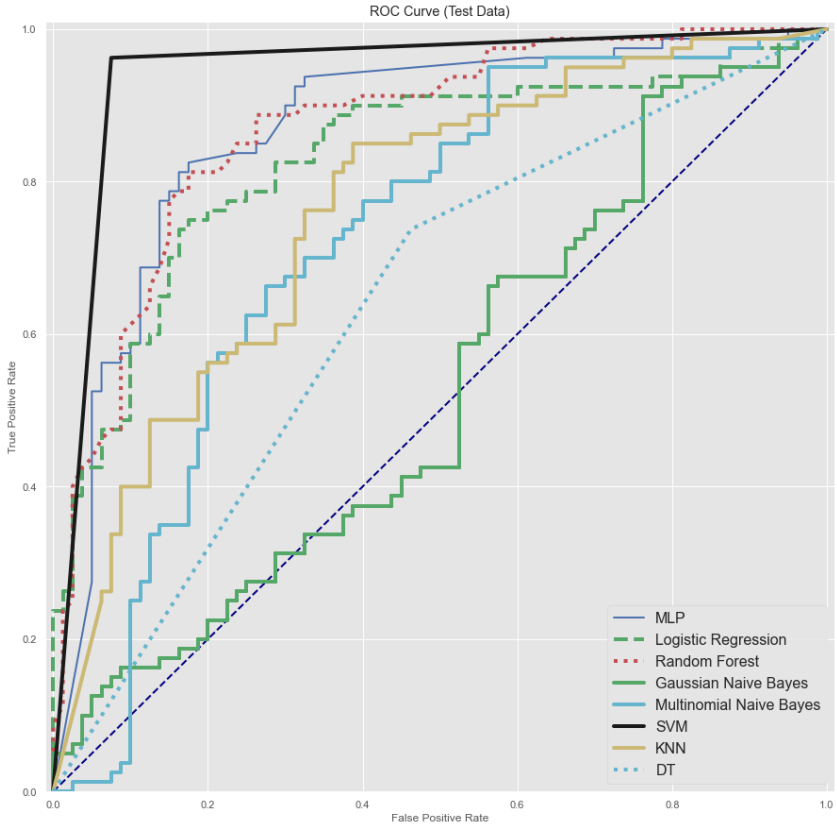
Before applying the balancing technique SMOTE:





After applying SMOTE:





Lessons Learned about ML topics due to project

ML topics demonstrate a variety of learning algorithms that can be utilized in different tasks. Starting with the type of learning, whether supervised, unsupervised, or reinforcement learning, the goals of the problem should be clearly specified beforehand in order to determine the type of the learning task. Then the data should be thoroughly examined to better understand the problem at hand. Datasets can have an effect on the learning process, whether small or large datasets; with independent or dependent attributes. Missing values should also be taken into consideration since they can impact the fairness of the learning algorithm and add an element of bias. Imbalanced data add a challenge and should be handled accordingly. Moreover, a good understanding of each learning algorithm helps in choosing the exact algorithm to implement. At the end,

a comparative analysis between the performance of algorithms can help decide and choose the best classifier for the task at hand.

Challenges faced and goals that were not achieved

The most challenging aspect of this problem was the imbalanced dataset as well as the small size of the input samples. A larger dataset would have provided more reliable results. This will be also an indication that the problem will generalize well on new test points. The other challenge is the accuracy of the results. It is the problem of every learning task to have classifiers with high values of accuracy as well as the other performance metrics (precision, recall, F1 score, AUC score).

References

[1] Zieba, M., Tomczak, J., Lubicz, J., & Swiatek, J., Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, vol. 14 (2013), 99-108. DOI: 10.1016/j.asoc.2013.07.016

[2] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak,* 20, 16, (2020). DOI: 10.1186/s12911-020-1023-5

[3] Desuky, A. S., Bakrawy, L. M. E. Improved Prediction of Post-operative Life Expectancy after Thoracic Surgery. *Advances in Systems Science and Applications*, 16(2), 70-80, (2016).

[4] Nachev, A. and Reapy, T. Predictive models for post-operative life expectancy after thoracic surgery. *Mathematical and Software Engineering*, 1(1), 1-5, (2015).

[5] V. Sindhu, S. A. S. Prabha, S. Veni , and M. Hemalatha, “Thoracic surgery analysis using data mining techniques” , *International Journal of Computer Technology & Applications* , vol. 5, pp 578-586, May, 2014.

[6] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadisa, “Machine learning applications in cancer prognosis and prediction”, *Computational and Structural Biotechnology Journal*, vol 13, pp 8-17, 2015.

Methods:

To determine which risk factors contributed the most to death one year after surgery, we decided to implement several classifiers on the data: KNN, Multinomial NB, Gaussian NB, SVM, and Random Forest. First, we performed SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset since the majority class was much larger, as seen in Figure 1.

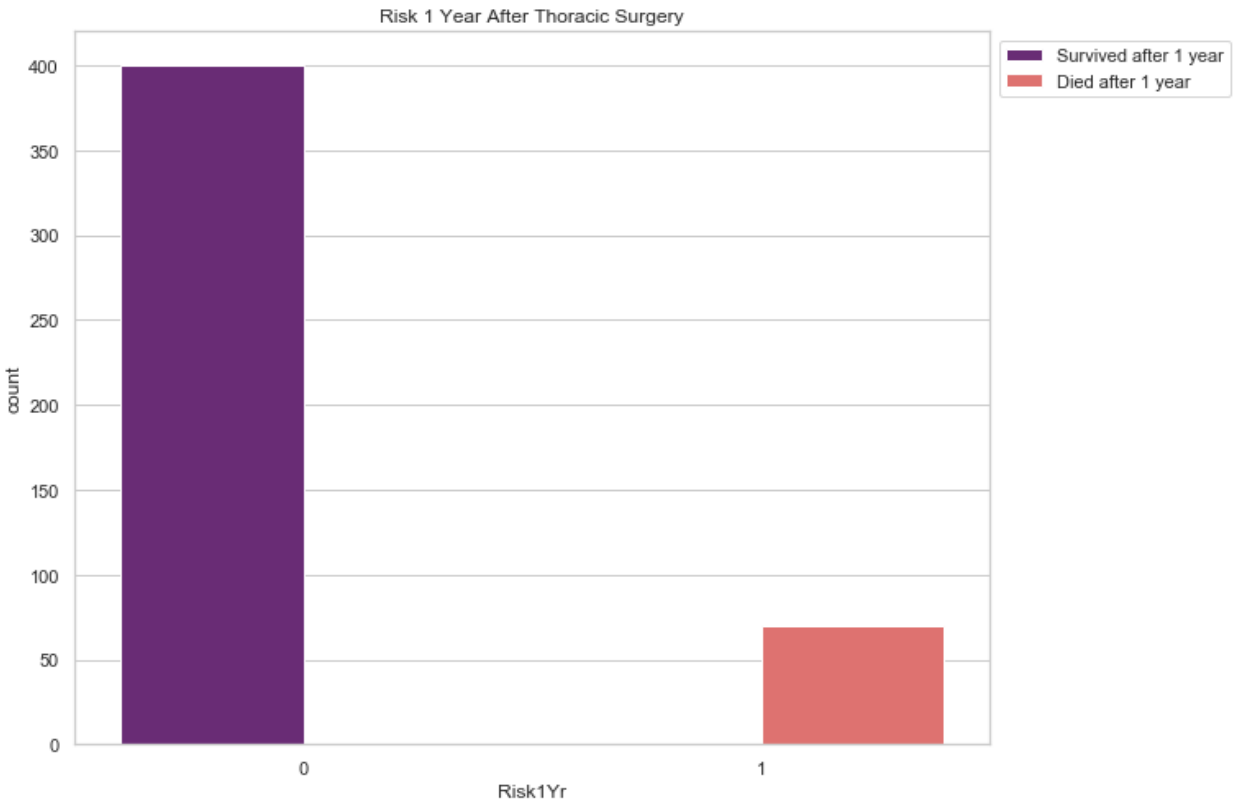


Figure 1: Class Distribution of Target Feature, Risk vs No Risk

After performing SMOTE, a heat correlation graph was generated to determine which attributes were more correlated to one another, as seen in Figure 2.

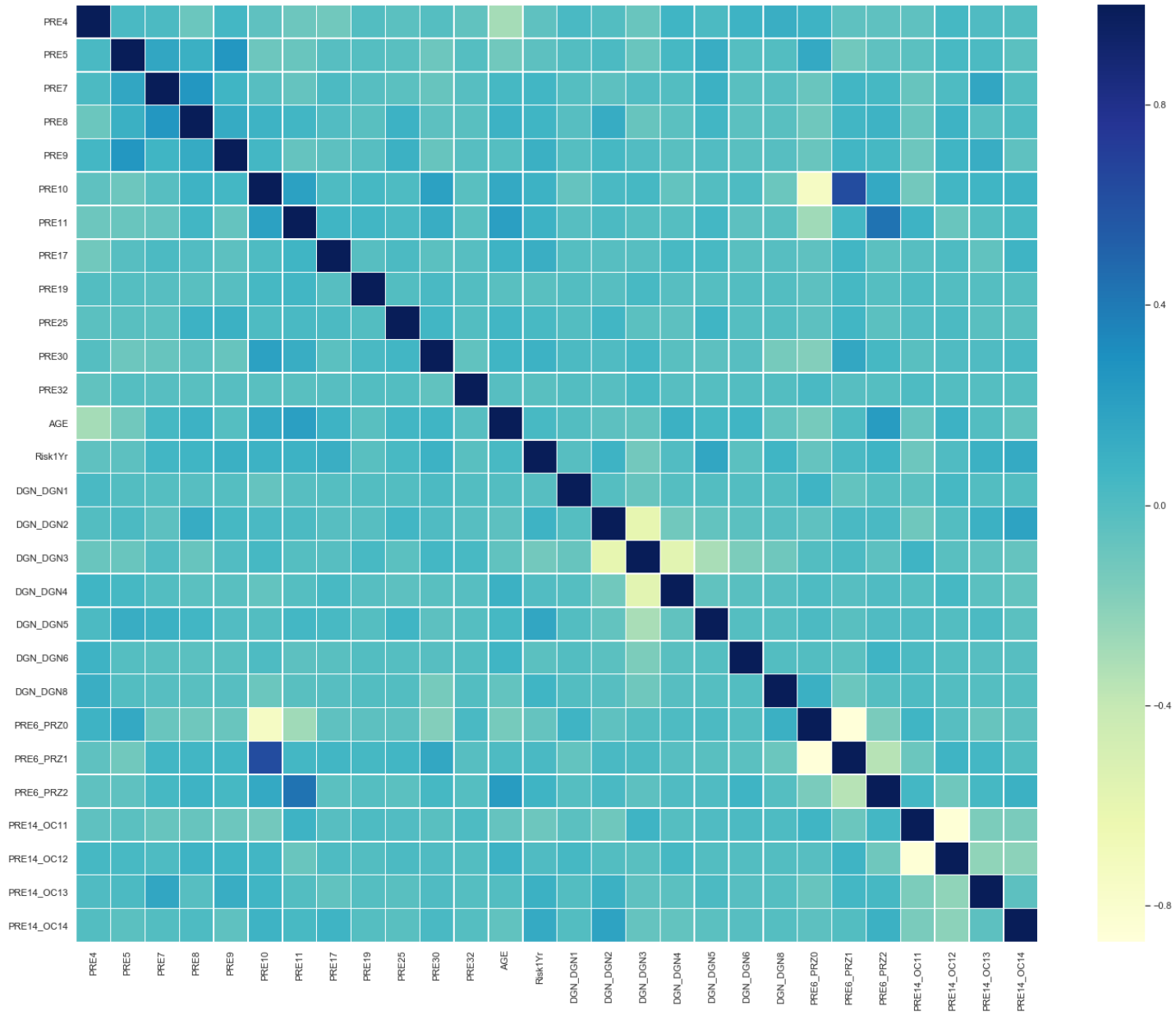


Figure 2: Heat Map Depicting Correlation between Attributes

The results of the heat map were compared to those mentioned in past studies, which mentioned that the attribute PRE14 (size of original tumor), PRE30 (smoking), PRE17 (type 2 diabetes mellitis), PRE8 (hemoptysis before surgery), and PRE 9 (dyspnea before surgery) were most relevant to the risk analysis.

Before performing any classification, the balanced dataset was loaded onto Tableau to gain a better understanding of the relationships between the attributes. First, the relationship between PRE14 and the target variable was explored, as shown in Figure 3.

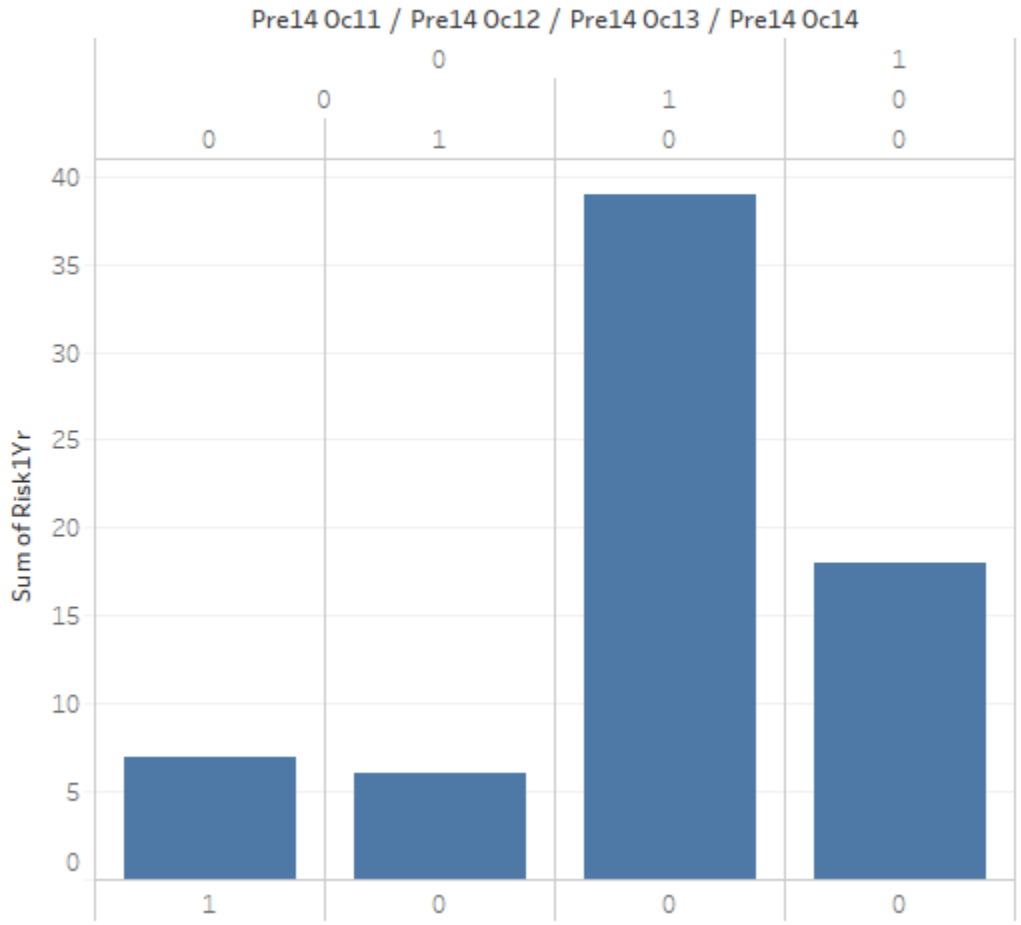


Figure 3: Relationship Between PRE14 Subgroups and Risk1Yr

The plot illustrates that among the 4 subgroups of PRE14, PRE14-OC12 was found correlating to the largest number of patients dying, 39 out of 70. PRE14-OC11 was the second highest at 18 people, followed by PRE14-OC13 (7) and PRE14-OC14 (6). The totals can be found in Table 1.

Table 1: Sum of Each PRE14 Subgroup in Relation to Risk

Pre14 Oc11	Pre14 Oc12	Pre14 Oc13	Pre14 Oc14	Sum of Risk1Yr
0	0	0	1	7
0	0	1	0	6
0	1	0	0	39
1	0	0	0	18

PRE14-OC12 was then analyzed in relation to the target variable, with 39 patients in the risk group, vs 218 in the no-risk group, compiling 56% of the total risk patients.

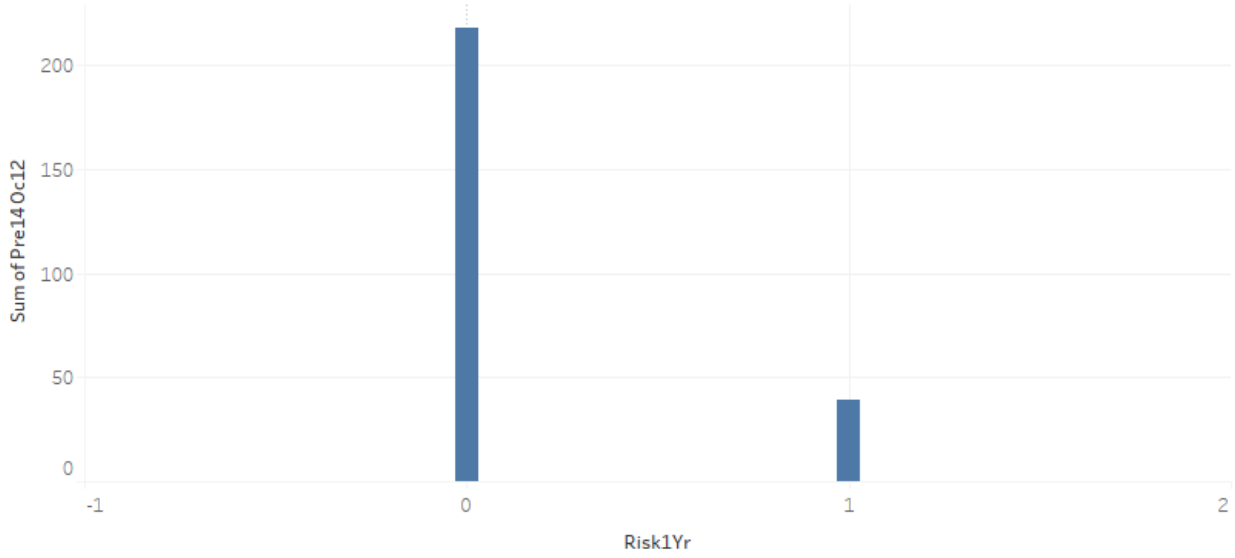


Figure 4: Risk vs No Risk Patients of Previous PRE14-OC12 Risk Factor

The relationship between the other relevant attributes listed in the paper was then investigated in regard to PRE14-OC12. First, PRE30 was plotted against Risk1Yr with PRE14-OC12 added as another dimension, as shown in Figure 5. All 70 patients were depicted from the risk group. Of these 70 people, 37 had both PRE30 and PRE14-OC12 as risk factors before surgery; 26 out of 70 people only had PRE30 as a risk factor, not PRE14-OC12; 2 people had only PRE14-OC12 as a risk factor, and 5 people did not have either of these pre-existing conditions as risk factors.

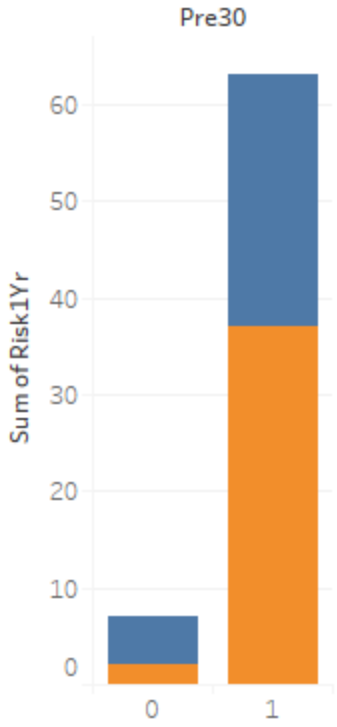


Figure 5: Plot of PRE30 vs Risk1Yr, in Relation with PRE14-OC12

The same was done for PRE17, as shown in Figure 6. However, there was less correlation between the two risk factors. Of the 70 patients at risk, 6 had both PRE14-OC12 and PRE17; 33 only had PRE14-OC12, but not PRE17; 4 only had PRE17, and 27 did not have either of these conditions.

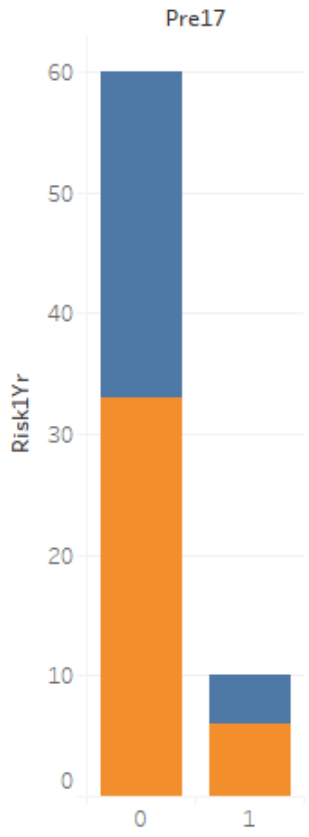


Figure 6: Plot of PRE17 vs Risk1Yr, in Relation with PRE14-OC12

We then conducted a cluster analysis of these three attributes in relation to the target variable. Choosing k = 3 for the number of clusters, the sum of Risk1Yr for each PRE14-OC12 was broken down with PRE17 and PRE30 as the supplementary attributes, depicted in Figure 7. We found that patients at risk were more likely to have both PRE14-OC12 and PRE30 as risk factors, followed by 23 patients with solely PRE30 as a pre-existing condition. There were only 6 patients who had all three risk factors.

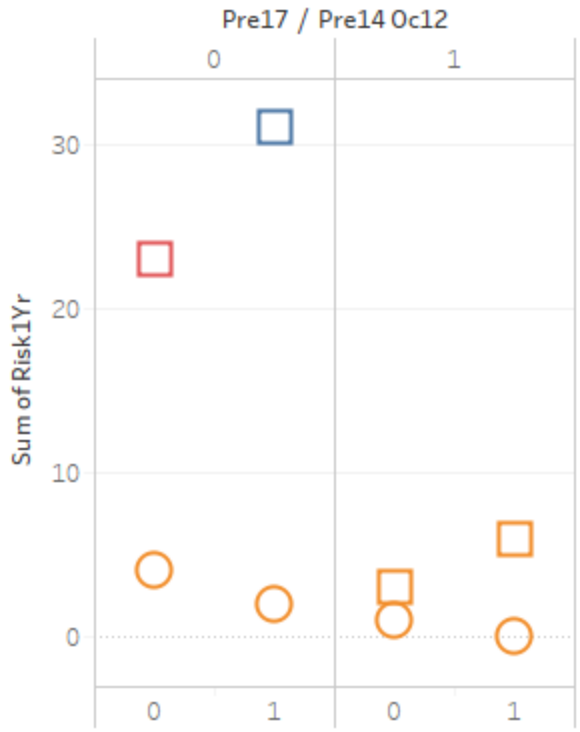


Figure 7: Sum of Risk1Yr for Each PRE14-OC12 Broken Down by PRE17 and PRE30

Results:

After conducting the exploratory data analysis, the classifiers were implemented using the newly balanced data to avoid impartiality. The accuracy and AUC scores of each classifier were obtained using the test data to compare each model's performance on the dataset. Figures 8 – 11 and Table 2 portray our findings.

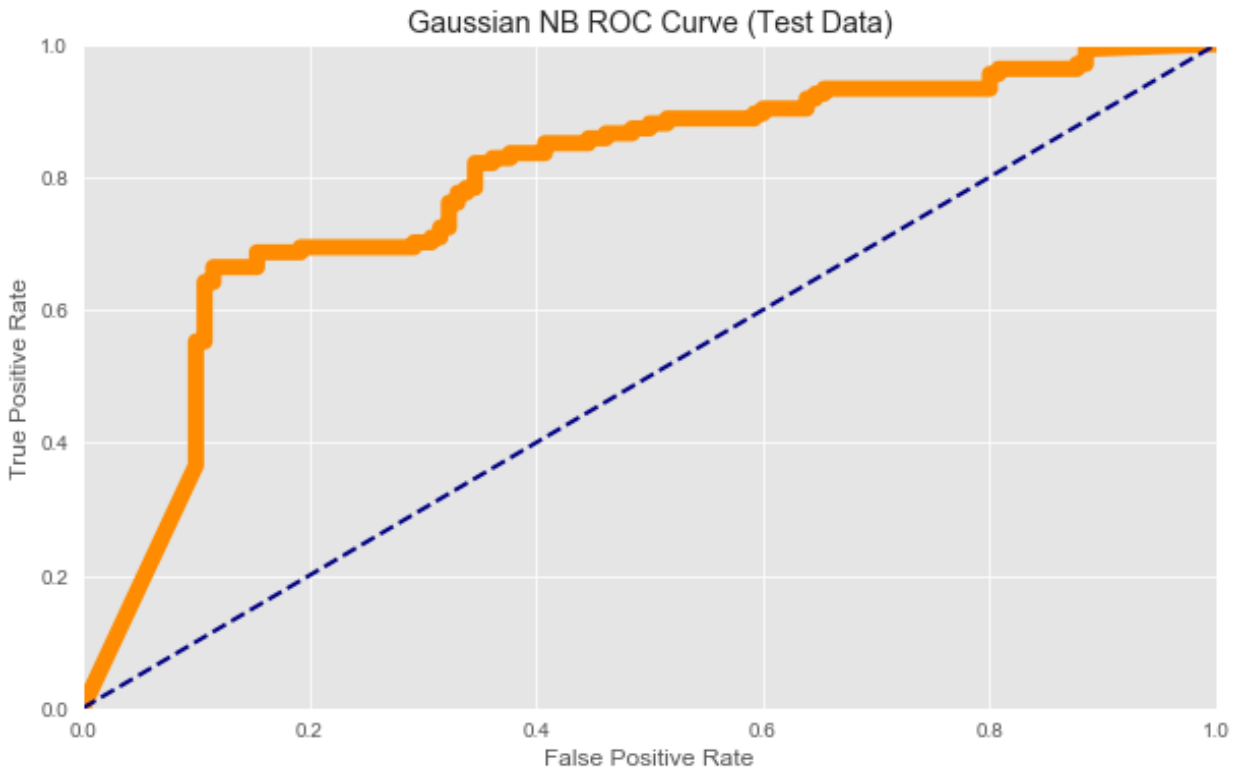


Figure 8: ROC Curve of Gaussian NB Classifier Implemented on Test Data

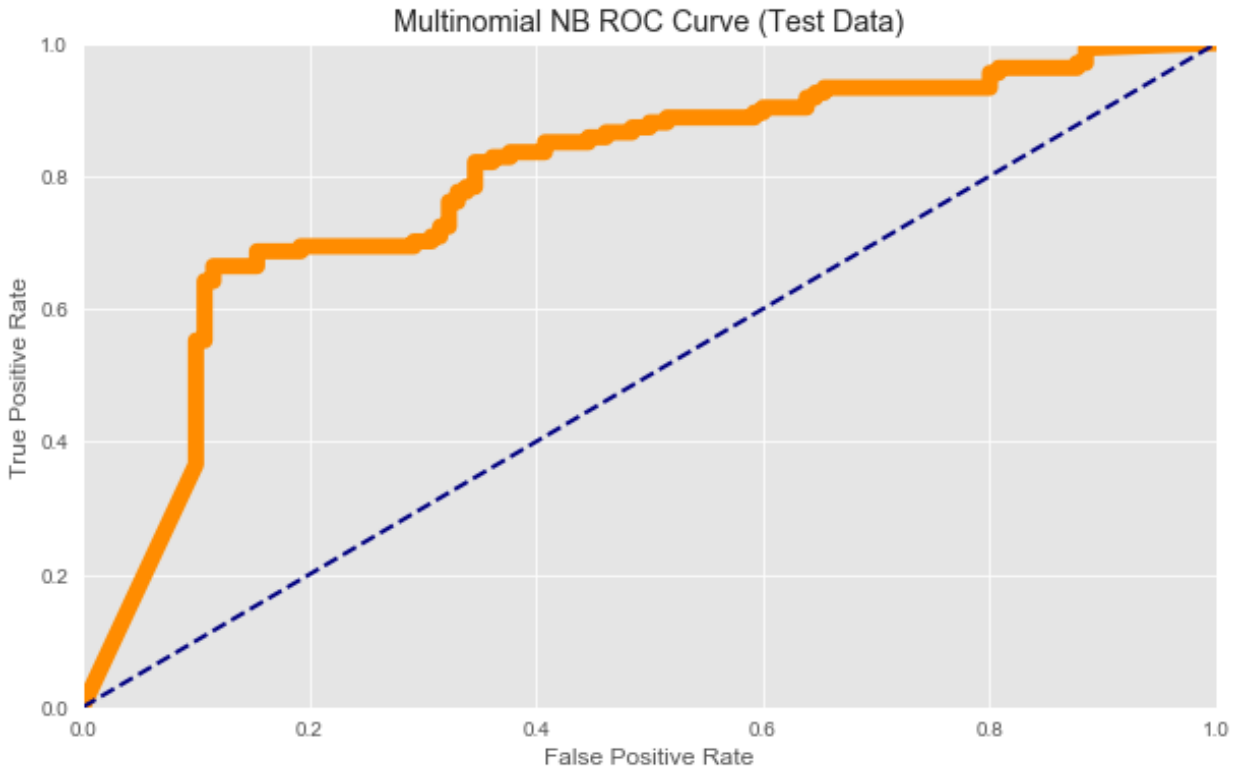


Figure 9: ROC Curve of Multinomial NB Classifier Implemented on Test Data

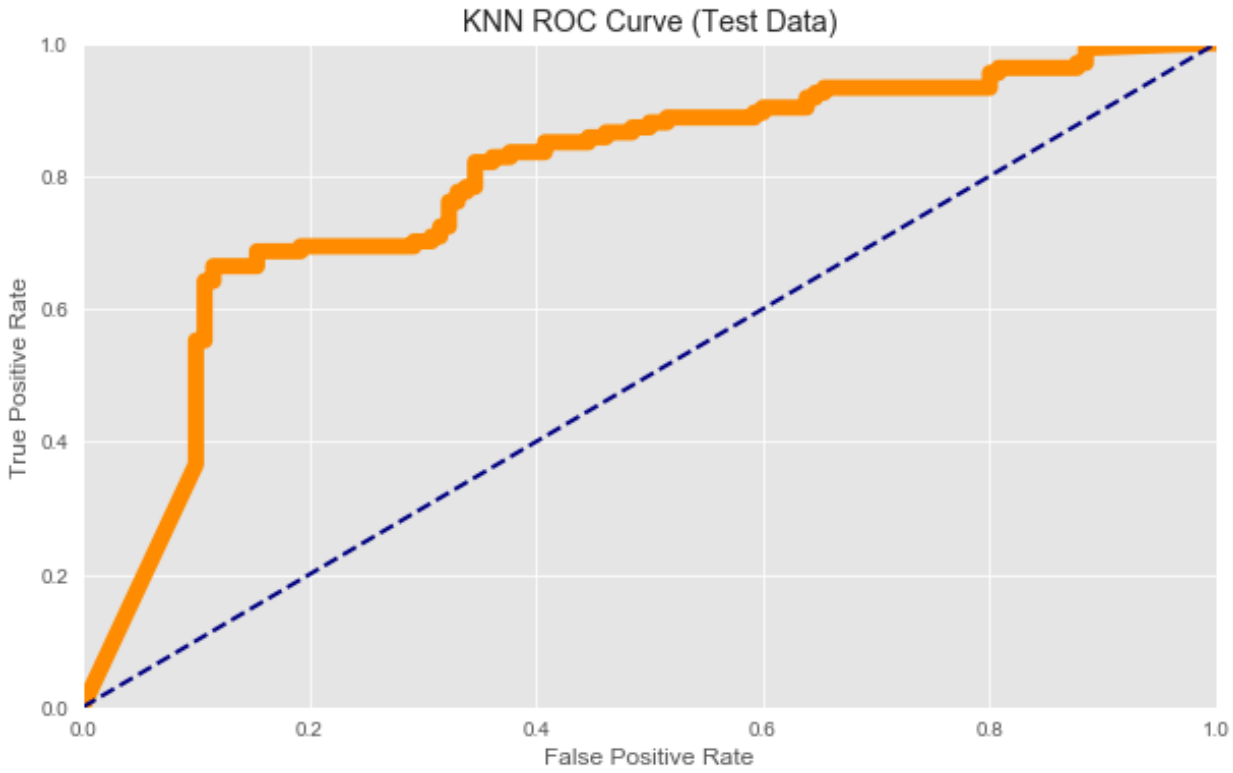


Figure 10: ROC Curve of KNN Classifier Implemented on Test Data

Table 2: Comparison of Accuracy and AUC Scores for Each Classifier

Classifier	Accuracy	AUC Score
SVM	0.88	0.956
KNN	0.82	0.789
Gaussian NB	0.64	0.683
Multinomial NB	0.67	0.697
Random Forest	0.9	0.877

The classifier that performed the best was SVM with the highest AUC score, followed by Random Forest, KNN, Multinomial NB and Gaussian NB, respectively.

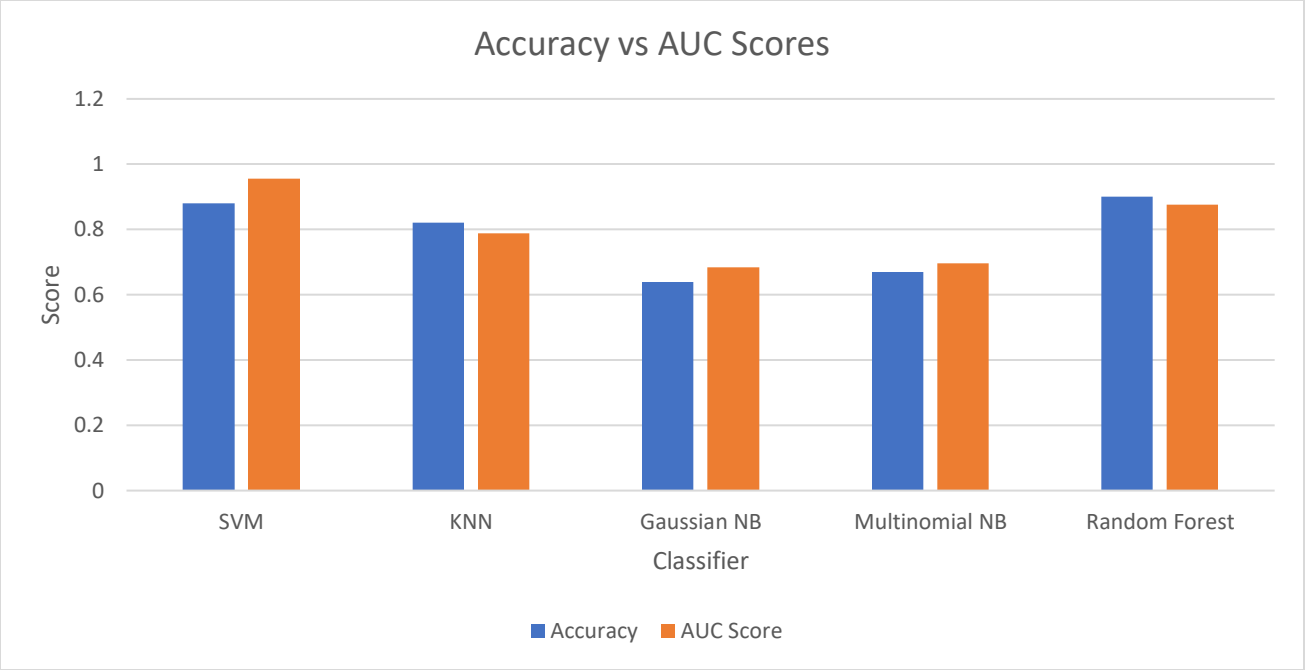


Figure 11: Comparison of Classifiers Using Accuracy and AUC Score