Survival Prediction of Lung-Cancer Patients After One Year of Thoracic Surgery

Health Risk Prediction

MANAL ZNEIT

SABINA BHUIYAN

CSCI 795 Machine Learning Project
Team G5

Problem Description/Goals

- Implement ML algorithms to examine medical reports and predict survival rates of lung-cancer patients one year after thoracic surgery
- Devise comparable ML algorithms and perform model evaluation
- A supervised learning task - a classification problem with risk of mortality as a target variable.
- Imbalanced dataset posed a challenge to the learning problem
- SMOTE is the technique used to handle imbalanced data problem

Team Members Role

- Manal Zneit
  - Built ML models and performed the comparative analysis of the results
- Sabina Bhuiyan
  - Examined features that best contributed to risk of mortality and implemented some models
  - Visualization tool and video to demo the results

State-of-the-art/Related Work

- [1] proposed a boosted SVM model for survival prediction using an oracle-based approach to extract rules to solve the imbalanced data problem
- In [2], a dataset of heart failure examination records was used to implement ML models and predict the risk of patients' survival.
- Performed feature selection and concluded that two features were sufficient to perform accurate predictions instead of using the entire dataset.
- 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

- 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.
- The dataset utilized in this project demonstrates an imbalance in the class distribution of the target vector
- The positive instances are under-sampled, and the majority class (negative instances) introduced a bias during the training phase
- SMOTE (Synthetic Minority Oversampling TEchnique) synthesizes data from the existing samples in the minority class in order to solve the imbalanced data problem

Experiments/Evaluation

- 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.
- 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
- 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

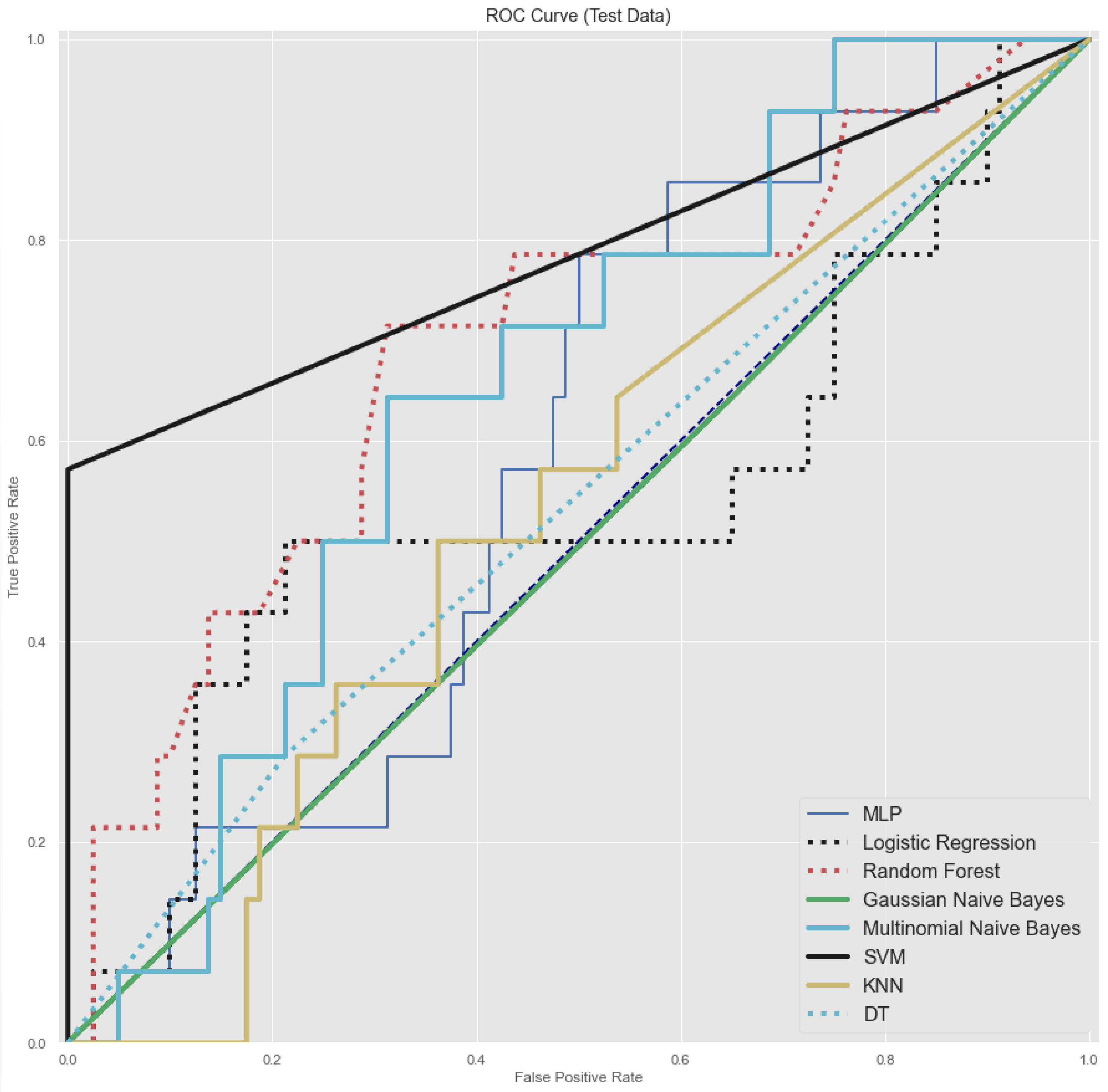
- The performance of the classifiers was compared based on the AUC performance metric
- The best performance was achieved by SVM (94%) followed by ANN (87%) and Random Forest (87%)
- DT and Multinomial Naïve Bayes had the lowest AUC values (64% and 72%) and NB was the weakest classifier in terms of precision, recall, F1 score, and accuracy.
- Logistic regression and KNN performed relatively the same (83% and 82%)

Discussion of Results

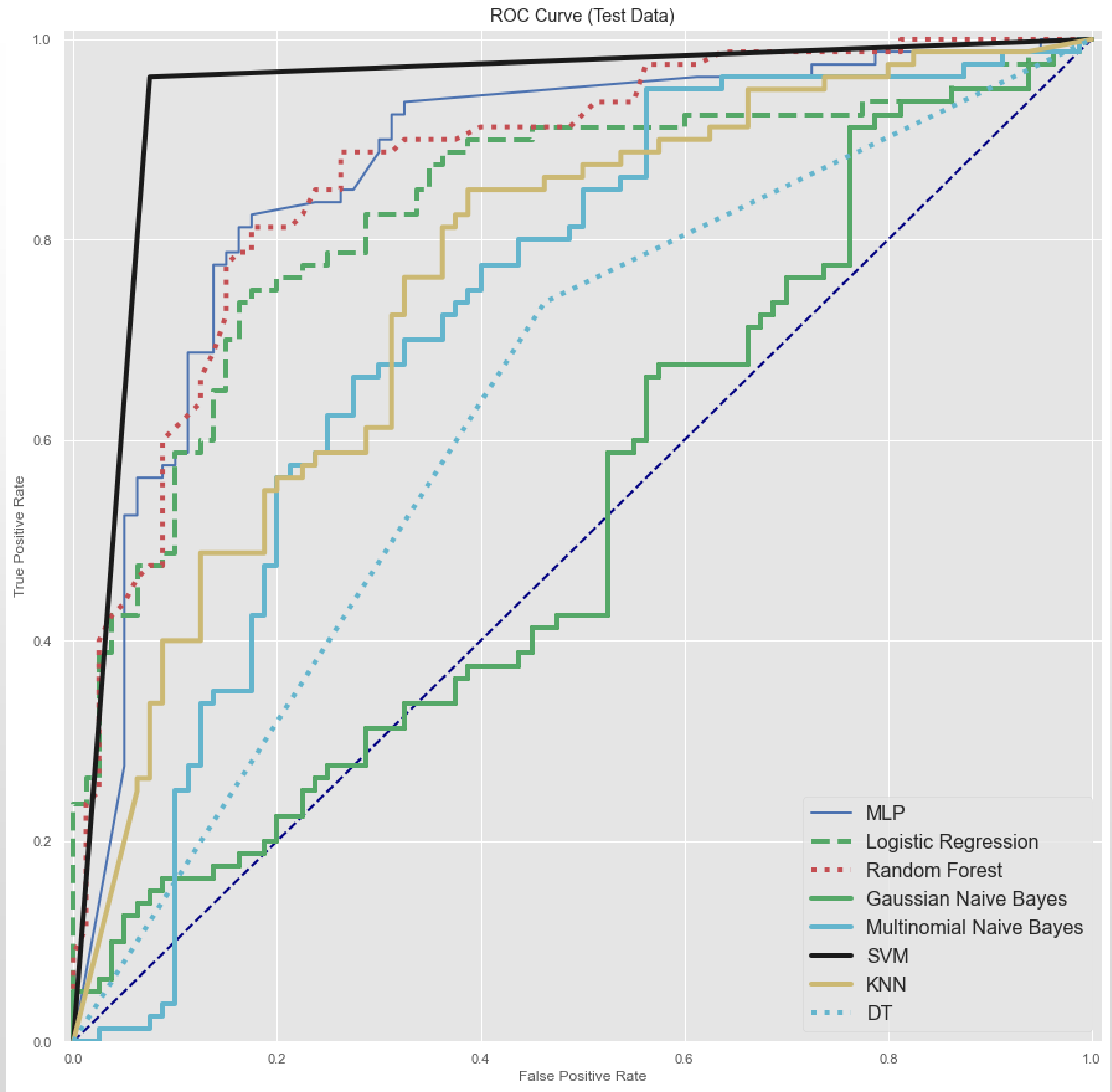
- The weak performance of Naïve Bayes is due to the simplifying assumption that presupposes conditionally independent features given the class label.
- In this dataset, it is unreasonable to assume the independence of features since the complications that a patient may experience are not simply independent of each other
- DT had the lowest performance measure (64% AUC score for the test data)
- 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 the final decisions at the leaf level.
- Non-linear dataset, SVM and MLP had the best performance

Results

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



ROC curves for classifiers (Imbalanced dataset)



ROC curves for classifiers (Balanced dataset)

Lessons Learned about ML Topics due to Project

- A variety of ML algorithms can be applicable to different learning tasks
- Some algorithms are task-dependent
  - This requires a thorough understanding of the dataset and the goals of the task
- Data preparation by scanning through the dataset and data-cleaning (EDA) are crucial before implementing a predictive model (apply SMOTE, missing data analysis methods…)
- Handling missing values and imbalanced data problems are necessary to avoid skew and biased outcomes
- A variety of learning algorithms potentially generate more reliable results (diverse and independent models)

Challenges Faced/Goals not Achieved

- Small dataset
- Imbalanced class distribution
- Accuracy of the results

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.