ML algorithm for predicting lung cancer presence

Machine Learning Algorithm for Predicting Lung Cancer Presence

Jenni Emily Puga-Raya, *Citrus College*
Dr. Doina Bein, Department of Computer Science, California State University, Fullerton

Background

Lung cancer is quickly becoming the leading cause of cancer deaths and is one of the most common forms of cancer diagnosis with its high danger and risk. Beyond the most common factor of cigarette and smoking, there are many others issues and variables that can play into a diagnosis. Current research projects test out several variables to try and create a predictive algorithm that can accurately predict a lung cancer diagnosis based on these varying factors. Through the development and research of effective classification algorithms, diagnosing can be made easier.

Proposed Work

The goal of the current research project is to examine machine learning concepts and apply different types of classifiers on a existing lung cancer dataset to create an accurate and highly effective algorithm that can precisely detect results.

Potential Applications / Improvements in:

- Improvement of Illness Diagnosis
- Algorithm Designing Prevention Models
- Medical Data Analytics

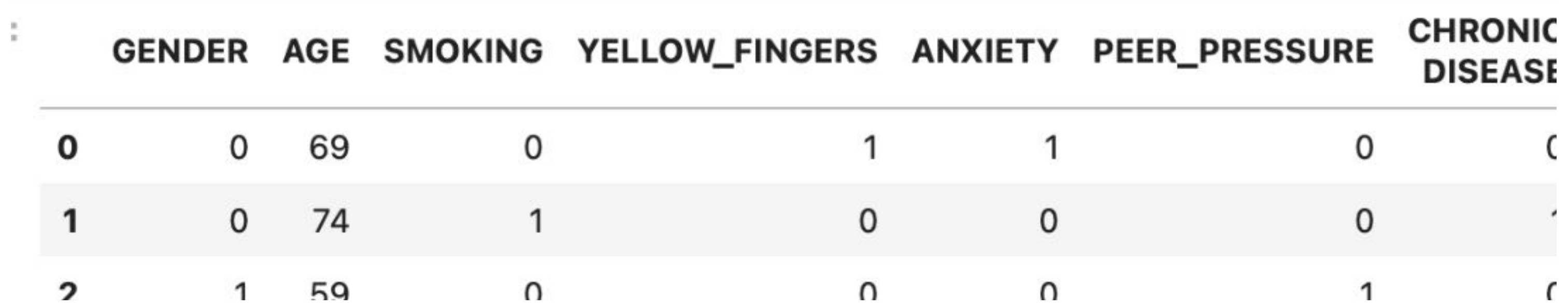
Methods

Utilizing a dataset [1] from Kaggle allowed us to analyze common factors that could lead to lung cancer in patients. The datasets were then collected and downloaded as csv files.

Jupyter Notebook through Anaconda configuration and Python programming was essential. Using the common Python libraries pandas and numpy, the data was read and then underwent cleaning through the following measures [2]:

- Checking for any missing values.
- Duplication and then deletion.
- Changing data to numerical from previous categorical.

Table 1: Example of Columns of Prepared Data



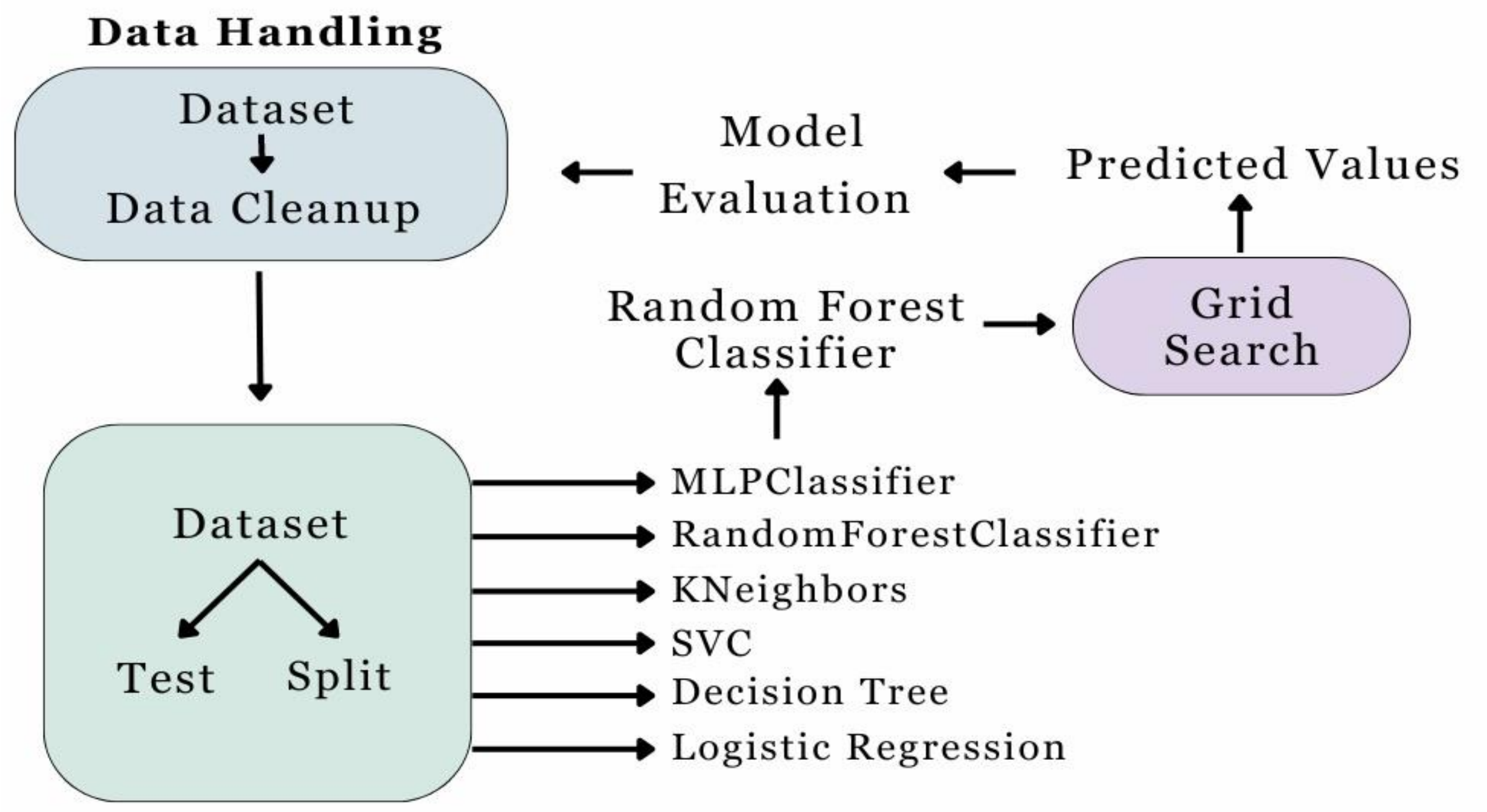
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE
0	0	69	0	1	1	0	0
1	0	74	1	0	0	0	1
2	1	59	0	0	0	1	0

A training model was created using the prepared data with the Python libraries Scikit-Learn through the use of a 20/80 train and test split–this allows data to learn for future data sets by stopping the algorithm from overfitting.

Due to the imbalance of the dataset (86% true positive versus 14% true negatives), certain adjustments were made to reduce the imbalance that could have skewed our data results through Synthetic Minority Over-sampling Techniques (SMOTE).

Methods (Continued)

Classical ML and Model Overview

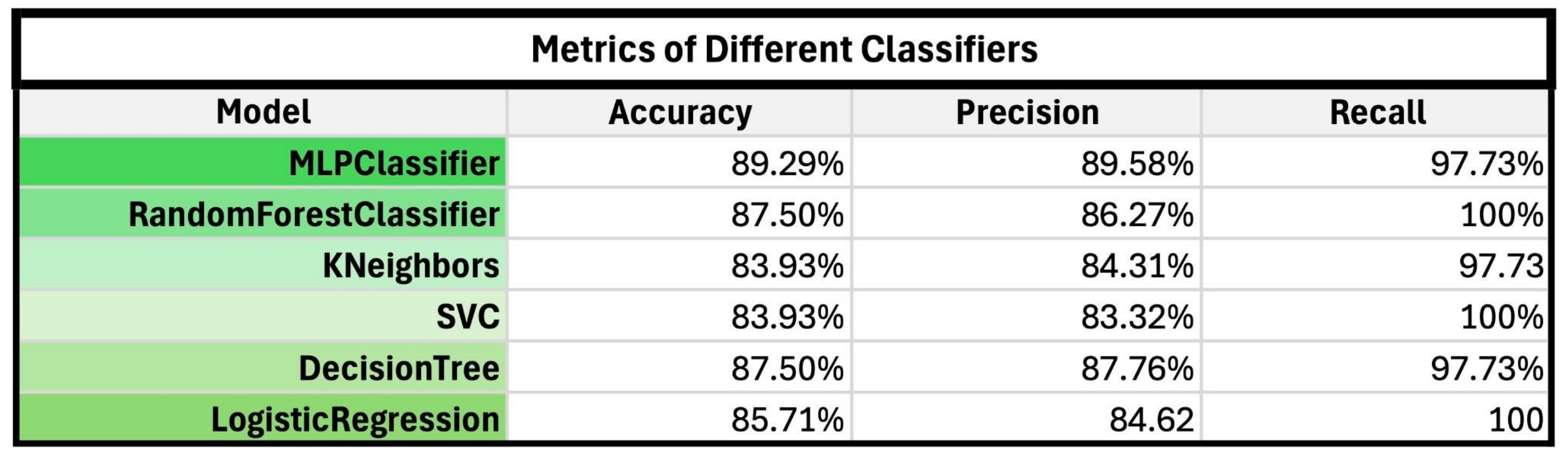


Different classifications models were tested to analyze the metrics before moving forward with the RandomForestClassifier, which focused on a grid search and hypertuning the parameters for evaluation.

Results

The table below contains the general metrics of each classification model used to test the algorithm.

Table 2: Scikit-Learn Performance Success Metrics



Metrics of Different Classifiers			
Model	Accuracy	Precision	Recall
MLPClassifier	89.29%	89.58%	97.73%
RandomForestClassifier	87.50%	86.27%	100%
KNeighbors	83.93%	84.31%	97.73
SVC	83.93%	83.32%	100%
DecisionTree	87.50%	87.76%	97.73%
LogisticRegression	85.71%	84.62	100

Specifically focusing on one model, the RandomForestClassifier was tuned with parameters to achieve more accurate and precise results.

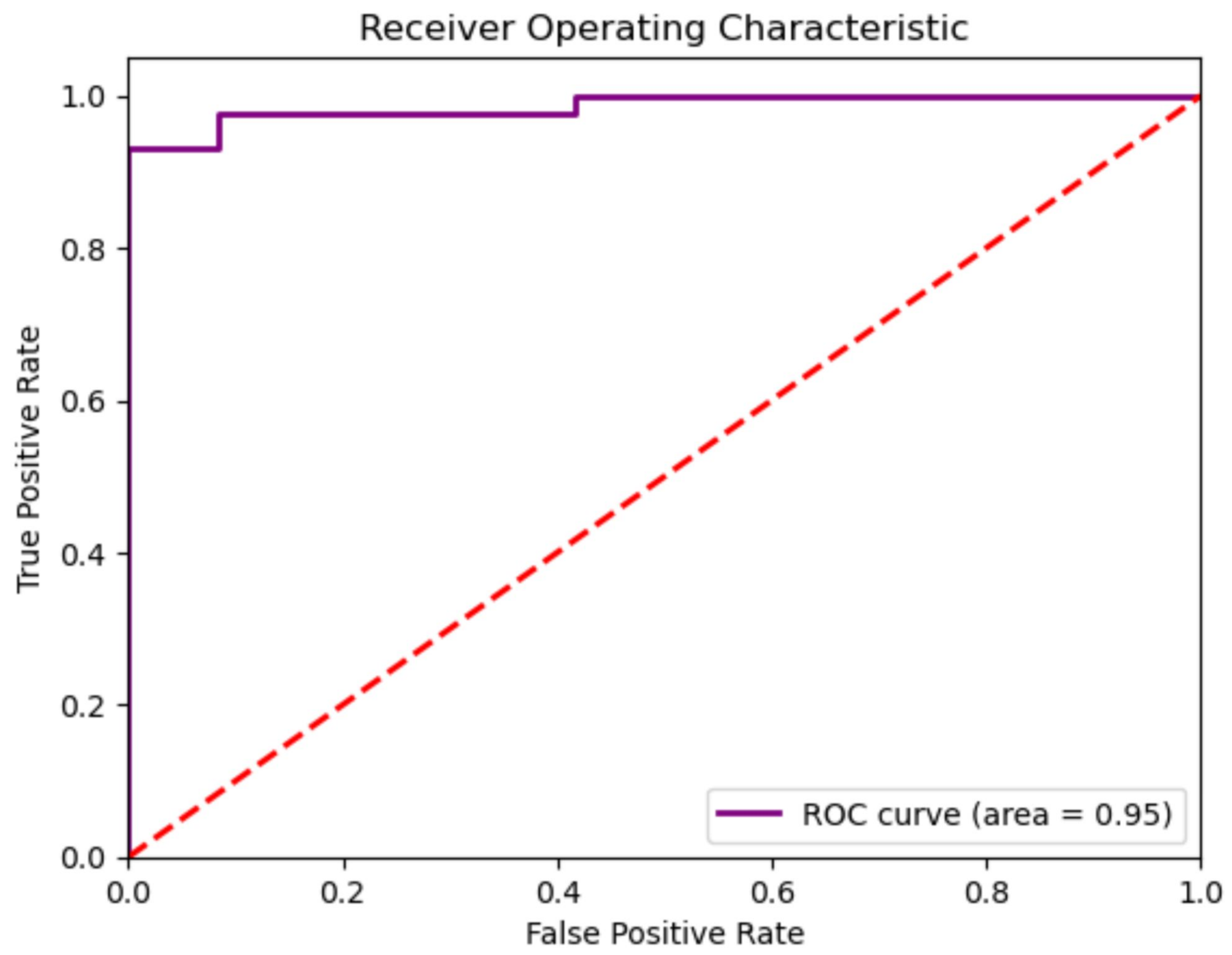


Figure 1: Scikit-Learn Performance Success Metrics

The figure above depicts the RandomForestClassifier (with hyper parameters) through an AUC-ROC graph. *Higher AUC-ROC = better results [3].

Conclusion

Based on the metrics obtained during testing, one can conclude that the algorithms tended to produce a high accuracy with the lowest being above 80%. One noteworthy conclusion is that the recall was near or at 100% for all the classifiers tested–extremely important for diagnosing lung cancer at early stages as even false negative diagnosing means all potential true positives will be caught.

When further introspecting a classifier–in this instance, the RandomForestClassifier –the AUC-ROC graph produces a 0.95 area of curve depicting exceptionally well and high effectivity. The program thus achieves or excels in the following:

- properly classifies potential patients,
- good distinguishment of true positives and false negatives,
- sensitivity to recall power and precision.

The application of the RandomForestClassifier on the Machine Learning algorithm model provides a effective prediction of lung cancer presence between potential cases.

Sources of Error / Limitations:

- Equipment used could not handle MLPClassifier running iterations and therefore RandomForestClassifier was the substituted for the AUC-ROC graph.
- Data imbalance of true positives and false positives could cause issues with bias and misdiagnosing.

Future Work

Future direction would likely allow many other ventures into properly diagnosing lung cancer such as:

- Hypertuning additionally more parameters to increase further the precision and accuracy importantly.
- Expand current emphasize from numerical datasets to diagnosing images and larger scale data.

References

1. Al Aswad, M. (2022, April) Lung Cancer, Version 6. Retrieved July 15, 2024 from https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer
2. Hong, Z.Q. and Yang, J.Y. "Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane", Pattern Recognition, Vol. 24, No. 4, pp. 317-324, 1991
3. Pepe, M. S. (2000). Receiver Operating Characteristic Methodology. Journal of the American Statistical Association, 95(449), 308–311. https://doi.org/10.2307/2669554

Acknowledgements

Project supported by Project RAISE, U.S. Department of Education HSI-STEM award number P031C160152 and Project RAISER U.S. Department of Education HSI-STEM Award P031C2100118

Jenni Emily Puga-Raya
California State University, Fullerton
Machine Learning Algorithm for Predicting Lung Cancer Presence

Background: Lung cancer is quickly becoming the leading cause of cancer deaths and is one of the most common forms of cancer diagnosis with its high danger and risk. Beyond the most common factor of cigarette and smoking, there are many others issues and variables that can play into a diagnosis. Current research projects test out several variables to try and create a predictive algorithm that can accurately predict a lung cancer diagnosis based on these varying factors. Through the development and research of effective classification algorithms, diagnosing can be made easier

Proposed work: The goal of the current research project is to examine machine learning concepts and apply different types of classifiers on a existing lung cancer dataset to create an accurate and highly effective algorithm that can precisely detect results.

Potential Applications / Improvements in:

- Improvement of Illness Diagnosis

- Algorithm Designing Prevention Models

- Medical Data Analytics

Methods: Utilizing a dataset [1] from Kaggle allowed us to analyze common factors that could lead to lung cancer in patients. The datasets were then collected and downloaded as csv files.

Jupyter Notebook through Anaconda configuration and Python programming was essential. Using the common Python libraries pandas and numpy, the data was read and then underwent cleaning through the following measures [2]:

- Checking for any missing values.

- Duplication and then deletion.

- Changing data to numerical from previous categorical.

A training model was created using the prepared data with the Python libraries Scikit-Learn through the use of a 20/80 train and test split–this allows data to learn for future data sets by stopping the algorithm from overfitting.

Due to the imbalance of the dataset (86% true positive versus 14% true negatives), certain adjustments were made to reduce the imbalance that could have skewed our data results through Synthetic Minority Over-sampling Techniques (SMOTE)

(a) Classical ML and Model Overview

Different classifications models were tested to analyze the metrics before moving forward with the RandomForestClassifier, which focused on a grid search and hypertuning the parameters for evaluation.

Results: The table below contains the general metrics of each classification model used to test the algorithm.

(b) Table 2: Scikit-Learn Performance Success Metrics

Specifically focusing on one model, the RandomForestClassifier was tuned with parameters to achieve more accurate and precise results.

(c) Figure 1: Scikit-Learn Performance Success Metrics

The figure above depicts the RandomForestClassifier (with hyper parameters) through an AUC-ROC graph. *Higher AUC-ROC = better results [3].

Conclusion: Based on the metrics obtained during testing, one can conclude that the algorithms tended to produce a high accuracy with the lowest being above 80%. One noteworthy conclusion is that the recall was near or at 100% for all the classifiers tested–extremely important for diagnosing lung cancer at early stages as even false negative diagnosing means all potential true positives will be caught.

When further introspecting a classifier–in this instance, the RandomForestClassifier –the AUC-ROC graph produces a 0.95 area of curve depicting exceptionally well and high effectivity. The program thus achieves or excels in the following: - properly classifies potential patients, - good distinguishment of true positives and false negatives, - sensitivity to recall power and precision.

The application of the RandomForestClassifier on the Machine Learning algorithm model provides a effective prediction of lung cancer presence between potential cases.

Sources of Error / Limitations:

- Equipment used could not handle MLPClassifier running iterations and therefore RandomForestClassifier was the substituted for the AUC-ROC graph.

- Data imbalance of true positives and false positives could cause issues with bias and misdiagnosing.

Future Work: Future direction would likely allow many other ventures into properly diagnosing lung cancer such as:

- Hypertuning additionally more parameters to increase further the precision and accuracy importantly.

- Expand current emphasize from numerical datasets to diagnosing images and larger scale data.

References:

1. Al Aswad, M. (2022, April) Lung Cancer, Version 6. Retrieved July

15, 2024 from https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer

2. Hong, Z.Q. and Yang, J.Y. "Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane", Pattern Recognition, Vol. 24, No. 4, pp. 317-324, 1991

3. Pepe, M. S. (2000). Receiver Operating Characteristic Methodology. Journal of the American Statistical Association, 95(449), 308–311. https://doi.org/10.2307/2669554

Acknowledgements: Project supported by Project RAISE, U.S. Department of Education HSI-STEM award number P031C160152 and Project RAISER U.S. Department of Education HSI-STEM Award P031C2100118