Lung Cancer Detection using Machine Learning

We present a solution for identifying lung cancer in patients using a binary classification model.

by Praveen Karthik Arumugam

Linkedin - www.linkedin.com/in/pk7779

Github Portfolio - https://github.com/praveenkarthika/data_science.git

Problem Statement and Type

We aim to detect lung cancer in patients using the given dataset of symptoms. This is a Binary Classification problem.

Data Collection

Data is collected from www.kaggle.com and preprocessed for further analysis.

Lung Cancer | Kaggle

EDA - Exploratory Data Analysis

Feature Engineering

We identify the key features and extract useful information to enhance the accuracy of our model.

Model Development

1

Splitting dataset for Training and Testing

We split the dataset into training and testing sets in order to evaluate the performance of our model.

2

Feature Selection

We select the most relevant features to use in the model to obtain the highest accuracy.

3

Model Selection

We use multiple models across Bagging and Boosting techniques and evaluate their performance using various accuracy metrics.

Model Performance Metrics

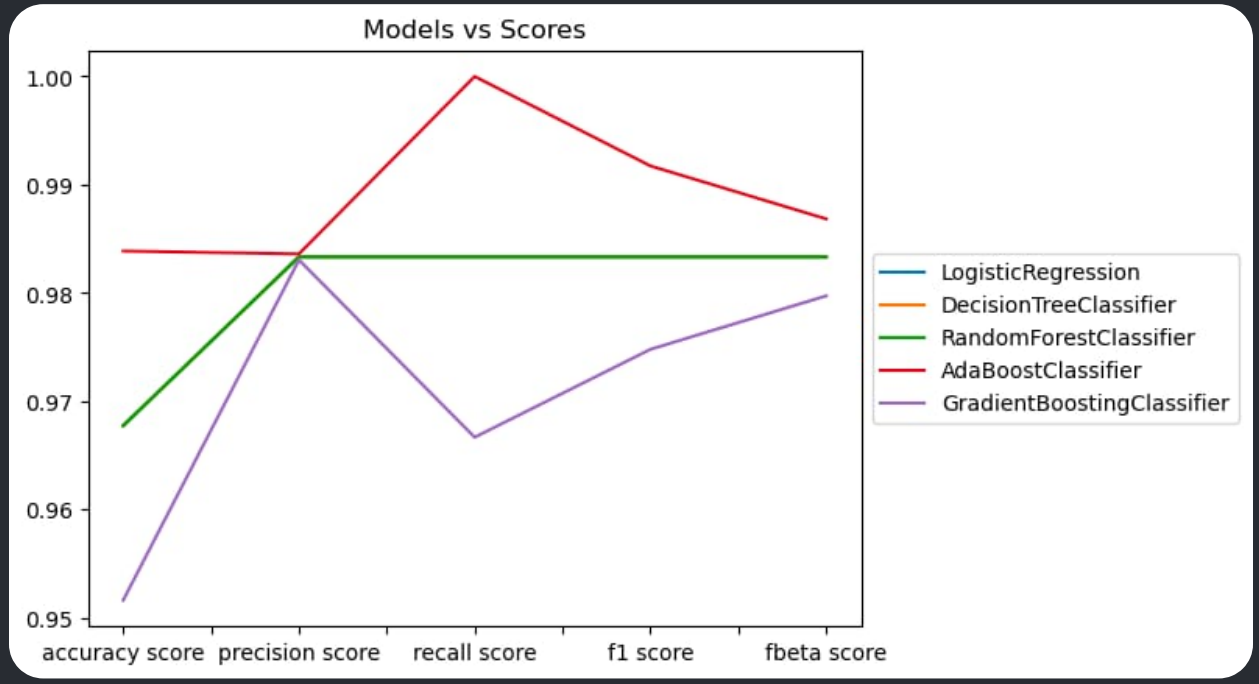
	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	AdaBoost Classifier	Gradient Boosting Classifier
accuracy score	0.967742	0.967742	0.967742	0.983871	0.951613
precision score	0.983333	0.983333	0.983333	0.983607	0.983051
recall score	0.983333	0.983333	0.983333	1.000000	0.966667
f1 score	0.983333	0.983333	0.983333	0.991736	0.974790
fbeta score	0.983333	0.983333	0.983333	0.986842	0.979730

Model Performance Metrics (Contd.,)

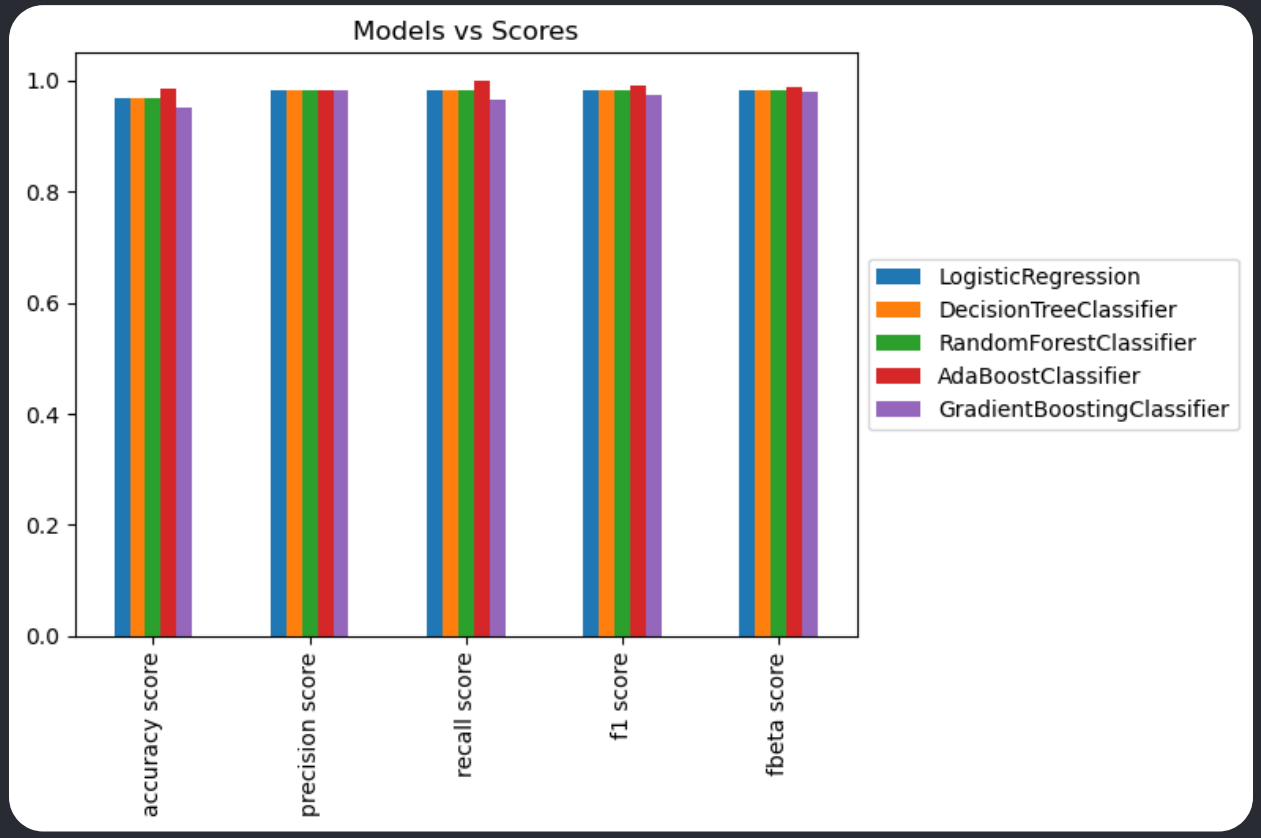
Even though the scores are given across diff. metrics for each ML algorithm in the above dataframe, taking a mean of all the score for a given model to understand which ML model is more accurate. In this case, AdaBoostClassifier is the most accurate ML algorithm to be used, as you can see recall_score is 1 and a mean for this ML algorithm in the below table is the highest to 0.989211

	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	AdaBoost Classifier	Gradient Boosting Classifier
count	5.000000	5.000000	5.000000	5.000000	5.000000
mean	0.980215	0.971170	0.980215	0.989211	0.971170
std	0.006973	0.012553	0.006973	0.006861	0.012553
min	0.967742	0.951613	0.967742	0.983607	0.951613
25%	0.983333	0.966667	0.983333	0.983871	0.966667
50%	0.983333	0.974790	0.983333	0.986842	0.974790
75%	0.983333	0.979730	0.983333	0.991736	0.979730
max	0.983333	0.983051	0.983333	1.000000	0.983051

Models vs Scores - Line



Models vs Scores - Bar



About the Presenter

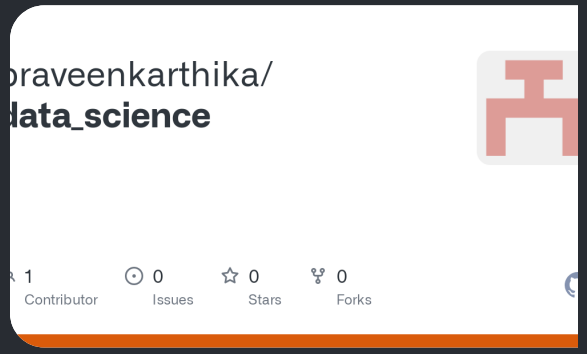
PraveenKarthik Arumugam

With a background in data science, Machine learning, Data Analytics and Automation/Manual Testing, PraveenKarthik Arumugam is an experienced professional in Information Technology.

Connect with him on LinkedIn at

www.linkedin.com/in/pk7779

Github Portfolio



GitHub

GitHub - praveenkarthika/data_science

Contribute to praveenkarthika/data_science development by creating an account on GitHub.