CSE 422
Lab Project
Session: Summer 2023
Topic: Lung Cancer Prediction

Submitted to:

Syed Zamil Hasan Shoumo (SZH)
and
Sumaiya Akter (SMA)

Group Members:

Atif Ronan
Id: 20201075

Asif Ali
Id: 20201049

Nafiun Al Amin
Id: 20201069

Nusaiba Zaman
Id : 20201104

Introduction:

In recent times, the mortality rate for lung cancer has been increasing rapidly. The use of Medical Science with Machine Learning could help predict the likeliness of lung cancer in a population. This project aims to predict lung cancer and to detect which existing conditions are predisposed to Lung Cancer, so that the population can avoid or treat their symptoms in the early stages.

Dataset Description:

Source: Kaggle.com/datasets

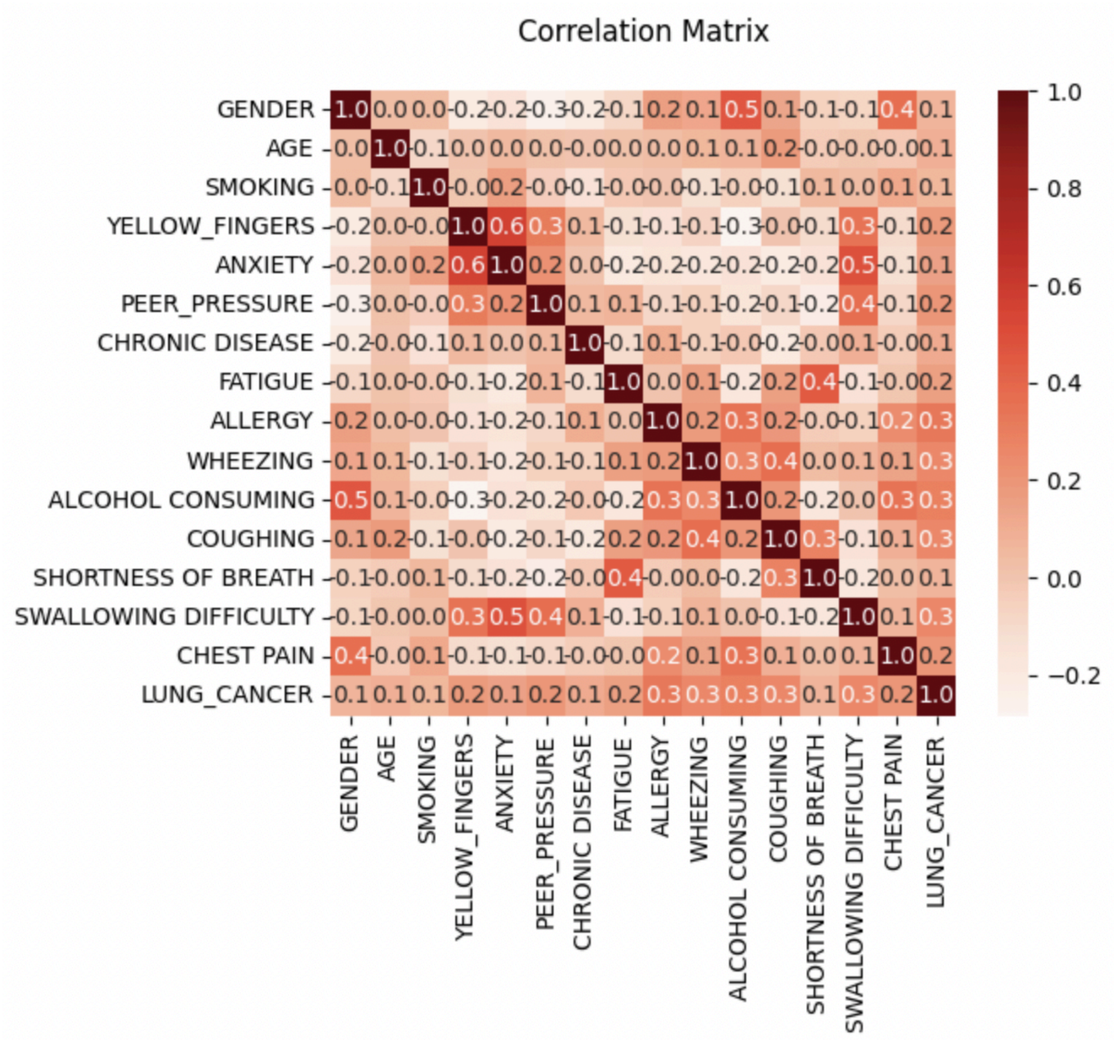
Dataset Link: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

References:

Bhat, M. A. (2021, October 1). *Lung cancer*. Kaggle. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

Course. buX. (n.d.). https://apps.bux-home.bracu.ac.bd/learning/course/course-v1:buX+CSE422+2023_Summer/home

The following dataset has 15 features. This dataset is a classification problem as it's Label, Lung Cancer is qualitative data that is answered with either a Yes or No. There are 309 data points. The features in this dataset are Gender, Age, Smoking, Yellow fingers, Anxiety, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consumption, Coughing, Shortness of breath, Swallowing difficulty and Chest pain. The feature, "Age" is quantitative and the rest are categorical features.



In the following correlation matrix, we infer that the features Gender, Age, Smoking, Anxiety, Chronic Disease and Shortness of breath are only 0.1 units correlated to the label lung cancer. The features Yellow fingers, Peer pressure, Fatigue, Allergy, Wheezing, Alcohol consumption, Coughing, Swallowing difficulty and Chest pain have a higher correlation to the label Lung Cancer. This matrix lets us know how much our features are dependant on Lung Cancer and to each other.

Imbalanced Dataset:

This dataset is heavily imbalanced, with the target label divided into two categories: 'yes' and 'no'. There are 238 instances classified as 'yes' and only 38 instances categorized as 'no'. Thus, it is evident that the dataset is imbalanced.

Dataset Preprocessing:

The dataset used does not contain any Null values. Hence, we have deleted values or created null values and then deleted the rows with the null values. The categorical values in features such as Gender and the label Lung Cancer have been converted to 0s and 1s. For the Gender feature, 0 is Female and 1 is Male while for the Lung Cancer label, No is 0 and Yes is 1. The other features were also changed to 0s and 1s from 1s and 2s, where 1s were encoded as 0 and 2s were encoded as 1.

Feature Scaling:

The feature “AGE” in our dataset is a continuos. We used MinMaxScaler to Scale our Dataset which makes sure all our features are on a comparable scale and not biased towards a feature that has more weight, which in our case is “AGE’.

Dataset Splitting:

We used Stratify, as the ‘No’ category in our label is very less, so it can make the prediction biased towards ‘Yes’. The training set has been divided keeping 70% of data points and the test set containing 30% of data points.

Model Training and Testing:

As the dataset we are working with falls under a classification problem, the models that have been implemented are:

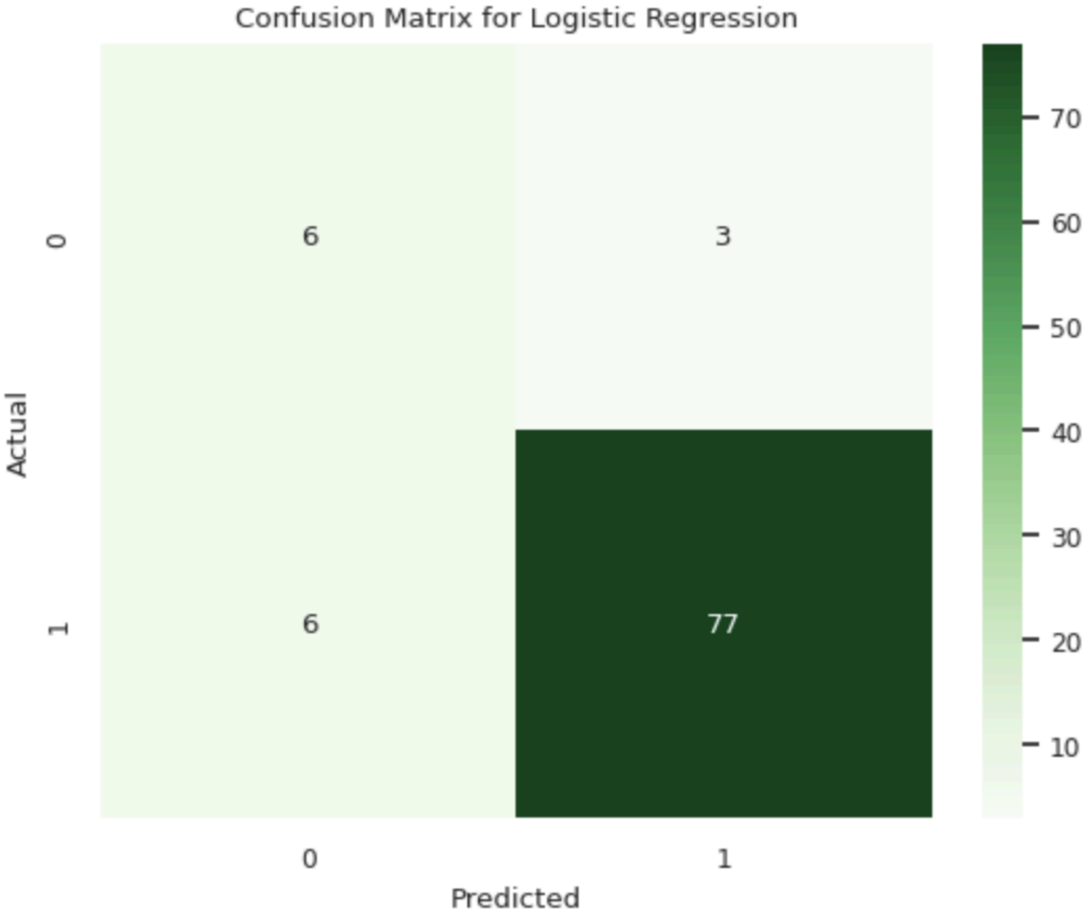
Logistic Regression:

Logistic Regression is used to solve Classification problems. It uses the logistic function to infer a value and classify it between 0 and 1 or the binary classifiers.

Accuracy:90.22%

Precision: 92.77%

Recall: 96.25%



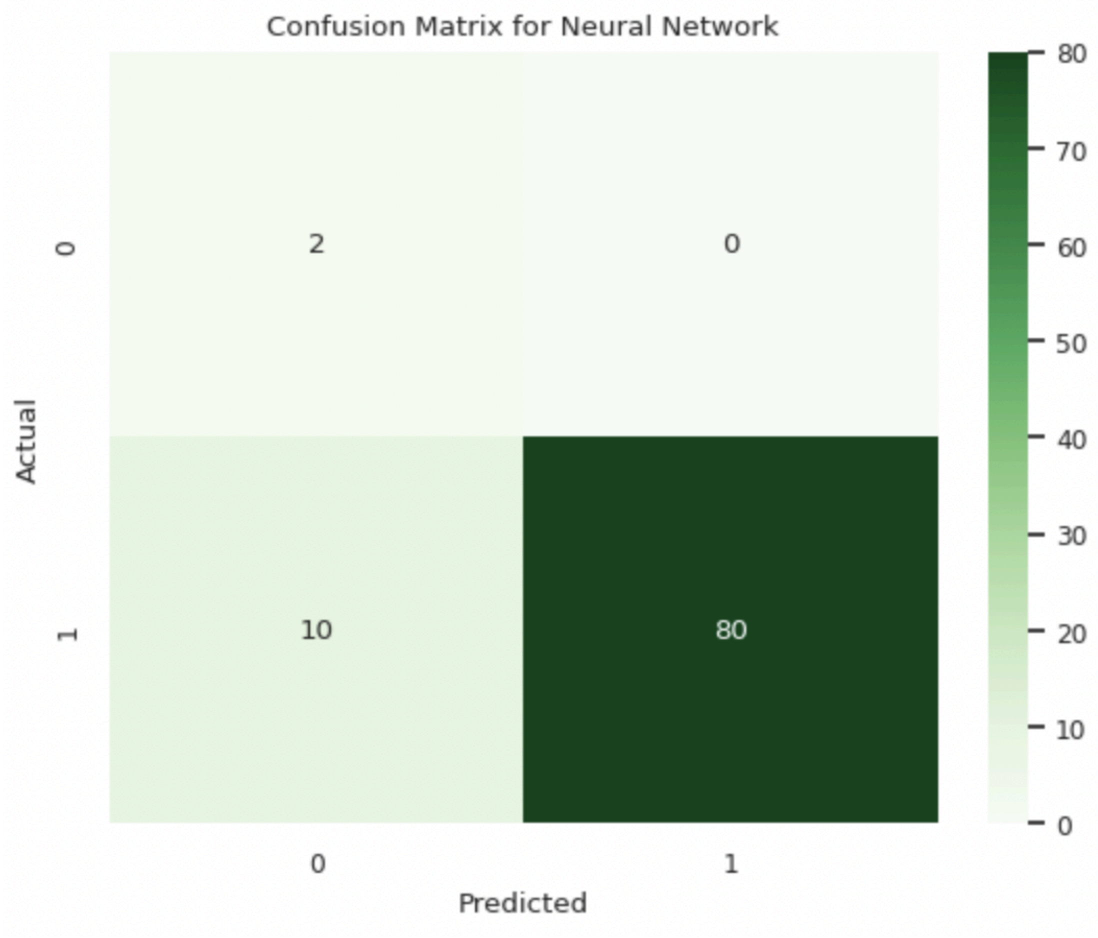
Neural Networks:

This model works for both classification and regression problems, it is inspired by the structure of the brain. It uses weights and biases alongside forward and backward propagation to predict outputs and then learn from itself.

Accuracy: 88.04%

Precision: 90.59%

Recall: 96.25%



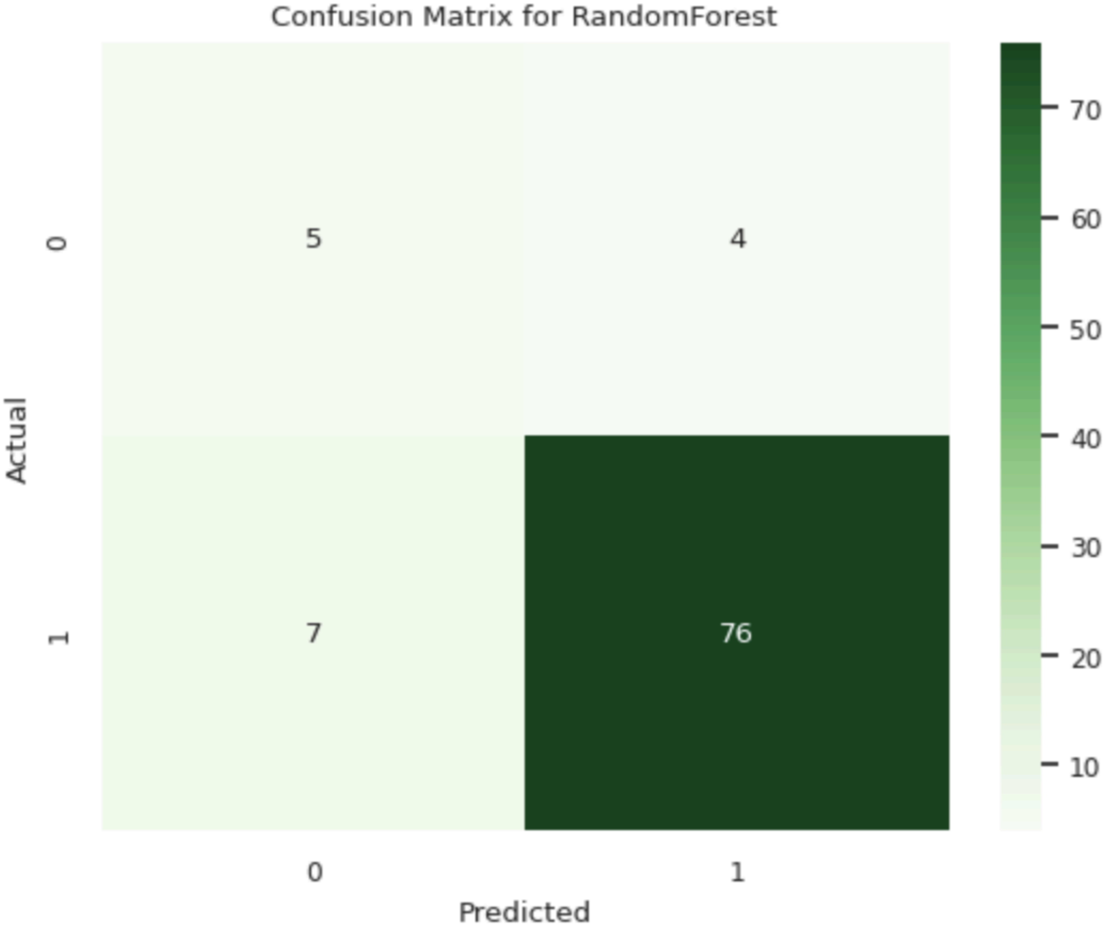
Ensemble Random Forest:

This model works for both classification and regression tasks, it uses multiple decision trees to improve it's predictive accuracy and decrease overfitting. Decision trees are generated with random feature selection and random data points. These decision trees are then combined to give the most optimal prediction.

Accuracy: 86.96%

Precision: 91.46%

Recall: 93.75%



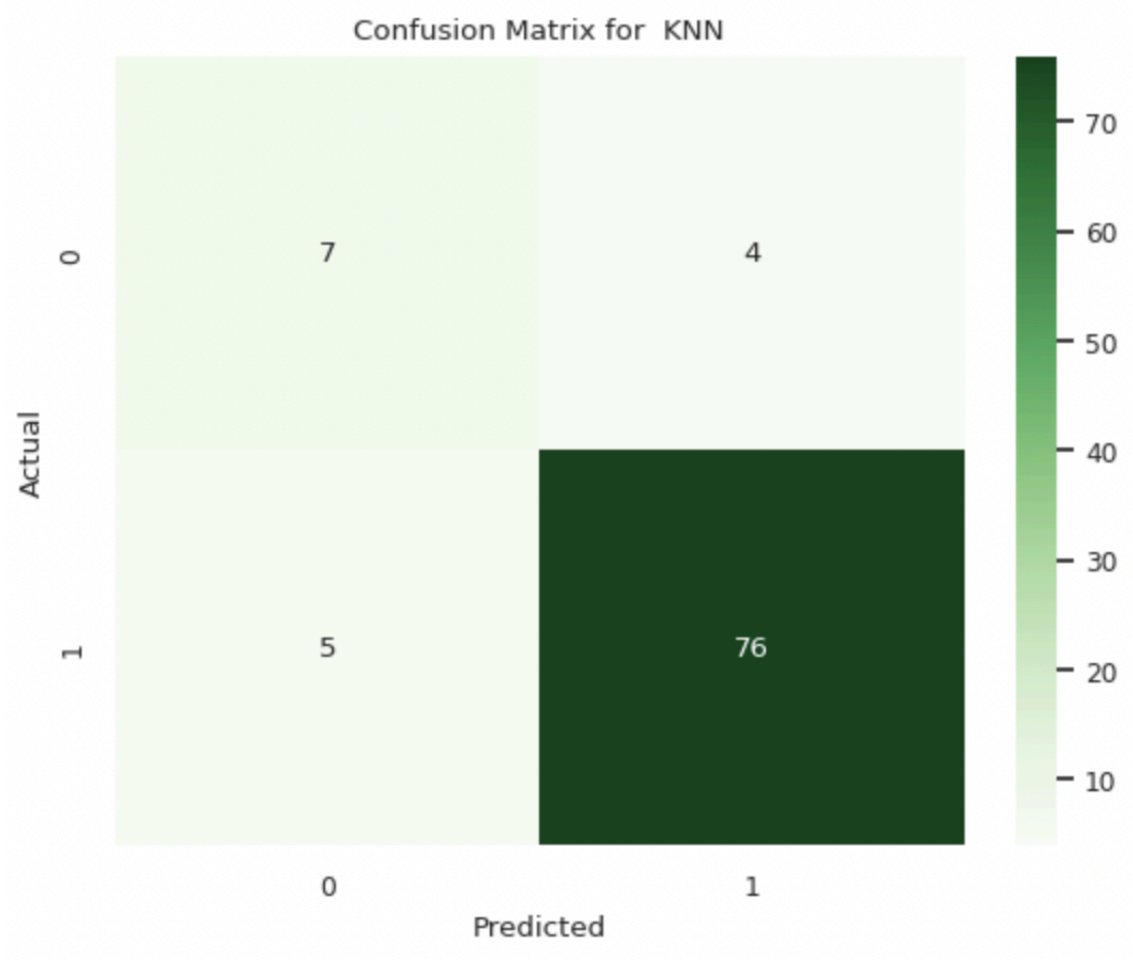
K Nearest Neighbours:

It is a machine learning algorithm that works for both classification and regression tasks. It predicts the label or value of a data point based on the majority class/value among its k nearest neighbors in the feature space. Below is the result produced when the value of k is 10.

Accuracy: 90.22%

Precision: 93.83%

Recall: 95.00%



Model Selection/Comparison Analysis:

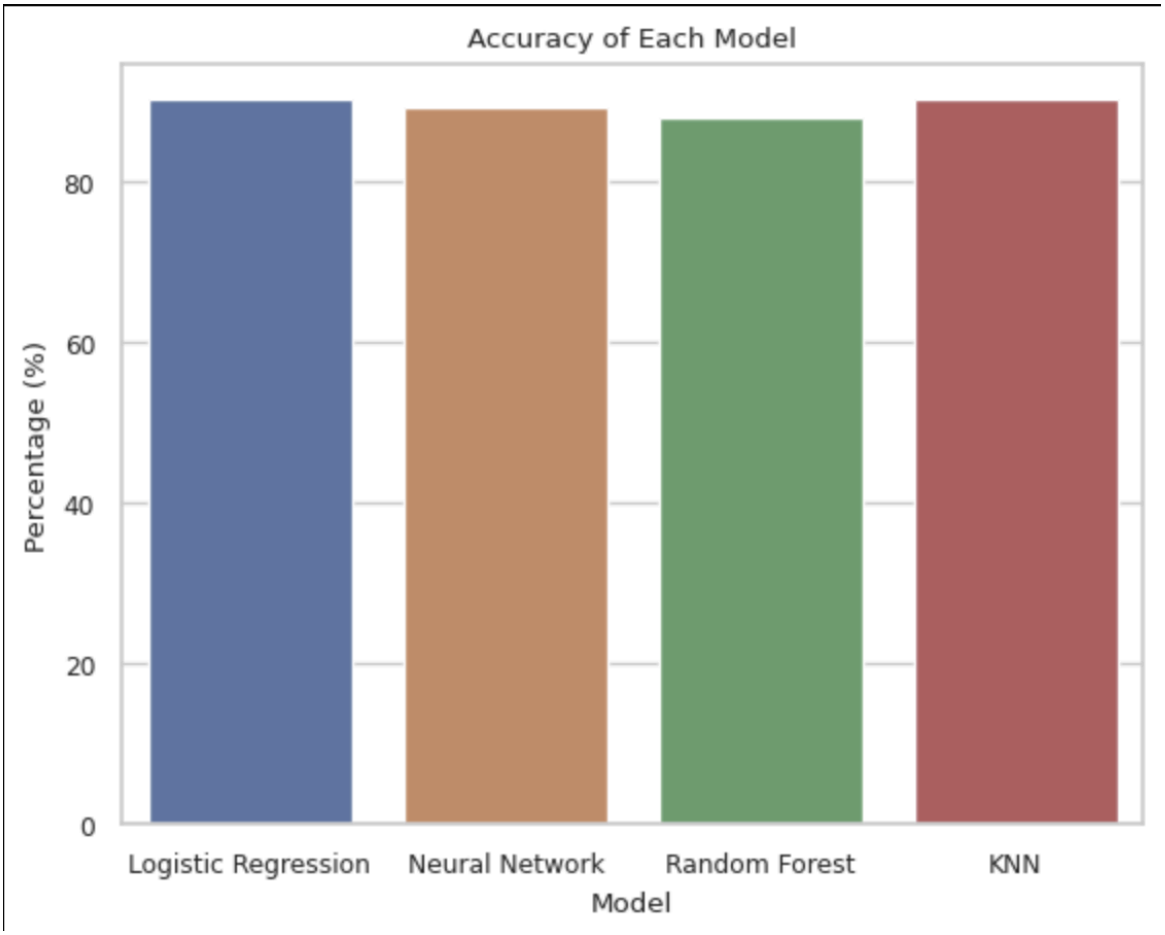


Fig. Model vs Accuracy percentage bar-chart

The Bar-chart above represents the Accuracy of each model and we can see that Logistic Regression and KNN produces the highest accuracy which is 90.22% whereas the lowest accuracy was produced by Neural Network which is 88.04%.

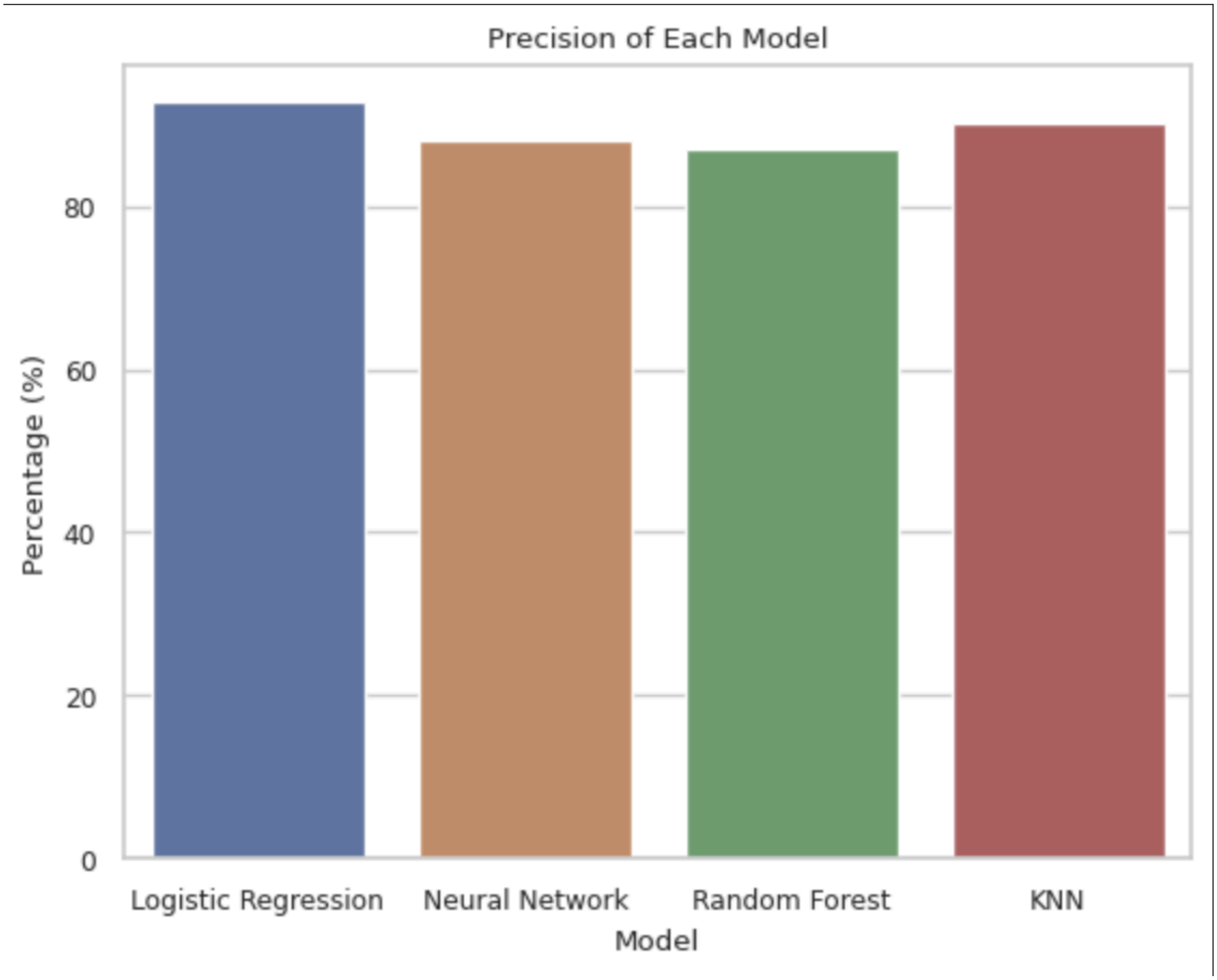


Fig. Model vs precision percentage bar-chart

The following bar chart compares the precision between the models we have used. Here, we can see that the precision for KNN is the highest at 93.83% and then Logistic Regression at 92.77%. The lowest precision at 90.59% belongs to Neural Networks.

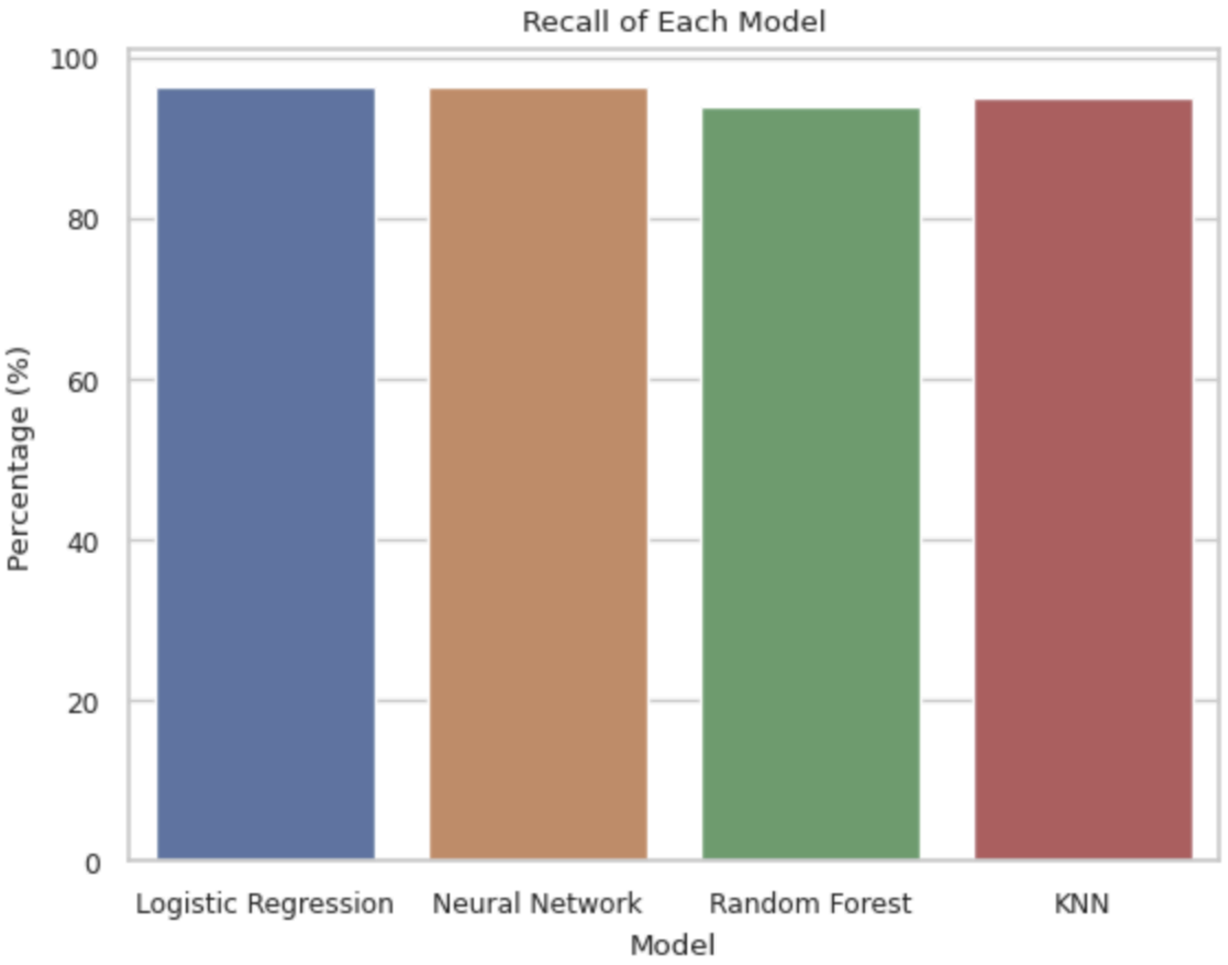


Fig. Model vs Recall percentage bar-chart

Here, we can infer that the recall is similar for Logistic Regression and Neural Networks. The recall is 96.25% for both.

Conclusion:

Predicting Lung Cancer can decrease mortality rates by warning the population through early detection. Our project uses a Lung Cancer dataset and a few models to accurately predict whether a person is more likely to have lung cancer or not. The model that produces the most accurate results are Logistic Regression and KNN. In terms of precision, the model KNN has the highest precision and for Recall both Neural Networks and Logistic Regression perform the best. Thus, we can use Logistic Regression or the KNN model to analyze data and predict the likelihood a person has towards Lung Cancer taking into account their lifestyle and attributes.