CSE-587 – Data Intensive Computing
Project Phase – 2

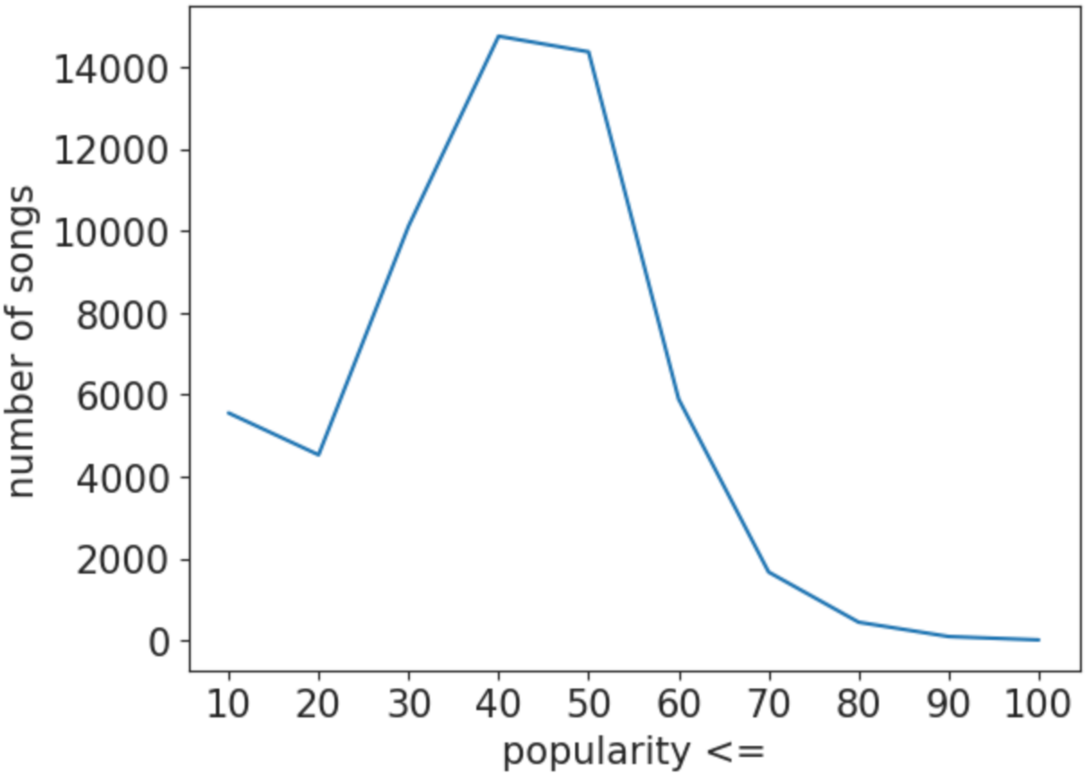
Thilak Reddy Dharam – thilakre – 50469154, Harsha Vardhan Bitra - hbirta - 50468952

Prediction of Songs Popularity on Spotify Data

Intro:

Popularity, acousticness, danceability, duration ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time signature, and valence are all included in the Spotify dataset. Based on the Correlation matrix we received in phase 1, the best features for the ML models we are using are **acousticness, danceability, energy, loudness, speechiness, and "popularity."**

We trained and tested five models: Logistic Regression, Neural Network, Support Vector Machine (SVM), Random Forest (RF) Architecture, and K-Nearest Neighbors. During training, these models examined a variety of song characteristics such as acousticness, danceability, energy, loudness, and speechiness.

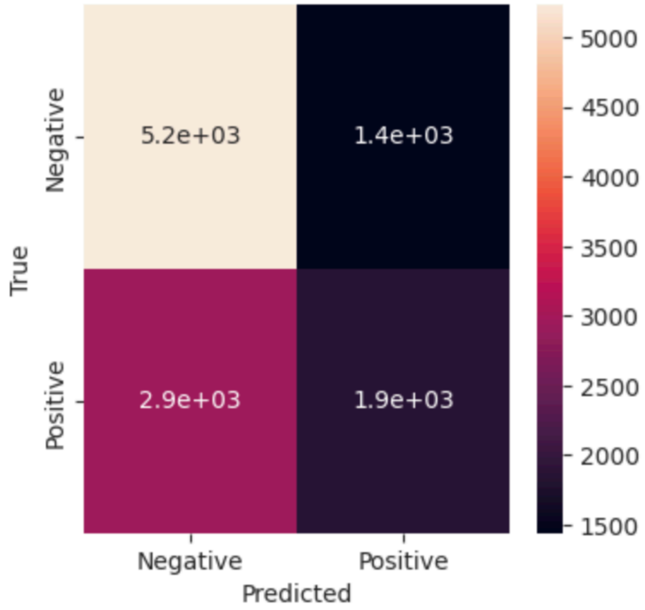


As we can see from the above graph, there are more songs at popularity of 40 and then decreases gradually. Therefore we are using binary classification as labelling 0 for the songs before popularity level of 40 and 1 for the rest of the popularity levels

Random Forest Classifier:

Random Forest is a well-known machine learning algorithm that can be used for classification and regression tasks. It is an ensemble method in which many decision trees are created, and their predictions are combined to make a final prediction. Each tree in a Random Forest Classification model is constructed using a random subset of features and a random subset of data. This reduces overfitting and improves the model's generalizability.
Because the regression was not giving the expected results, I used the classification approach because the target feature is discrete and by making a range of popularity values and labelling that range. We used a grid search method for observing the best parameters {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 100} and used those values for getting an accuracy of **61.90%.**

Confusion Matrix:



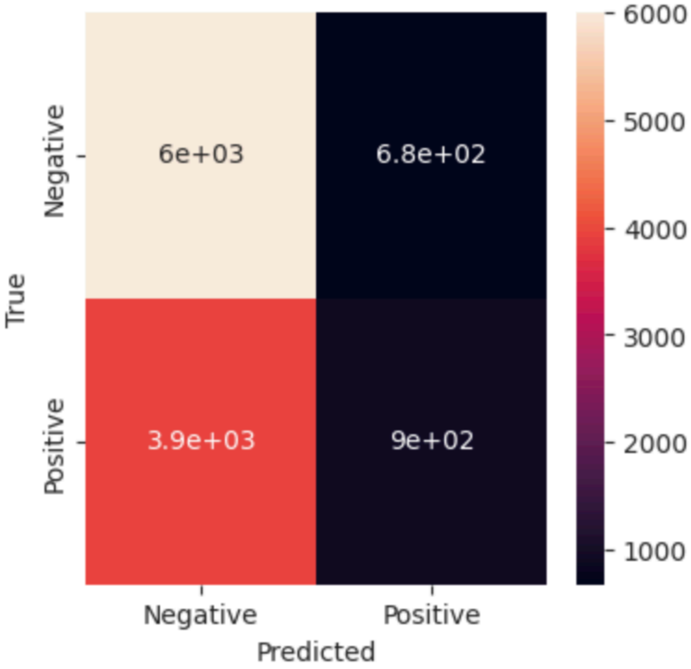
SVM (Support Vector Machines):

A Support Vector Machine (SVM) is a versatile and powerful supervised machine learning algorithm that is used for classification and regression tasks. It is a type of discriminative classifier that determines the best boundary, or hyperplane, between two data classes. The SVM algorithm finds the hyperplane with the greatest margin between the two classes.

SVMs can also handle data that is not linearly separable by transforming it into a higher-dimensional space where it is more likely to be linearly separable. This is accomplished through a technique known as kernel trick, which maps data into a higher-dimensional space without explicitly computing the data's coordinates in that space.

For our problem statement, we tested the kernels on “linear” first and then “rbf”. Rbf kernel gave the best accuracy possible when compared to the linear kernel because it is more flexible and can capture complex, nonlinear relationships between the input features and the output variable. Using rbf kernel the model achieved an accuracy of **60.10%.**

Confusion Matrix:

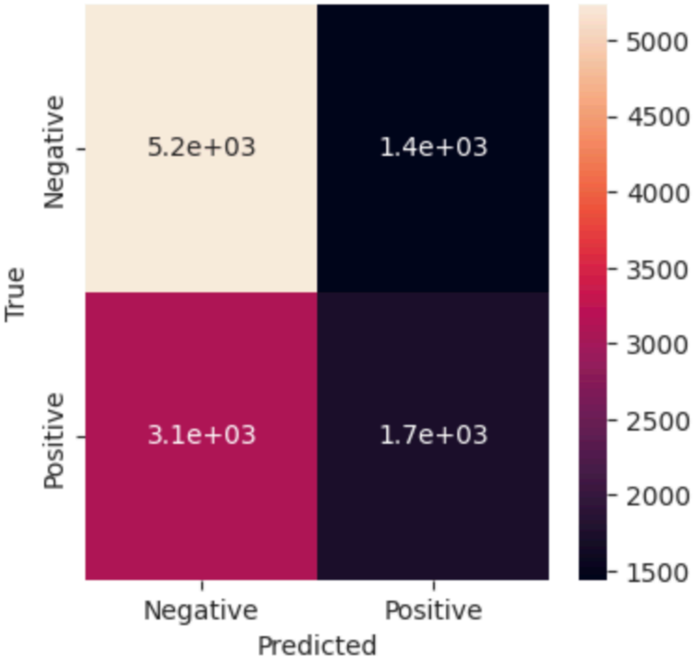


Logistic Regression:

Logistic regression is a method for binary classification tasks that seeks to determine the likelihood of a binary outcome based on one or more predictor variables. In the context of predicting song popularity, logistic regression can estimate the likelihood of a song's popularity based on the relationship between the input features and the binary outcome variable of popularity.

Furthermore, logistic regression is a parametric algorithm because it makes assumptions about the underlying distribution of data, which is a significant benefit when dealing with linear relationships between features and outcome variables. Using Logistic Regression, we achieved an average accuracy of **60.70%.**

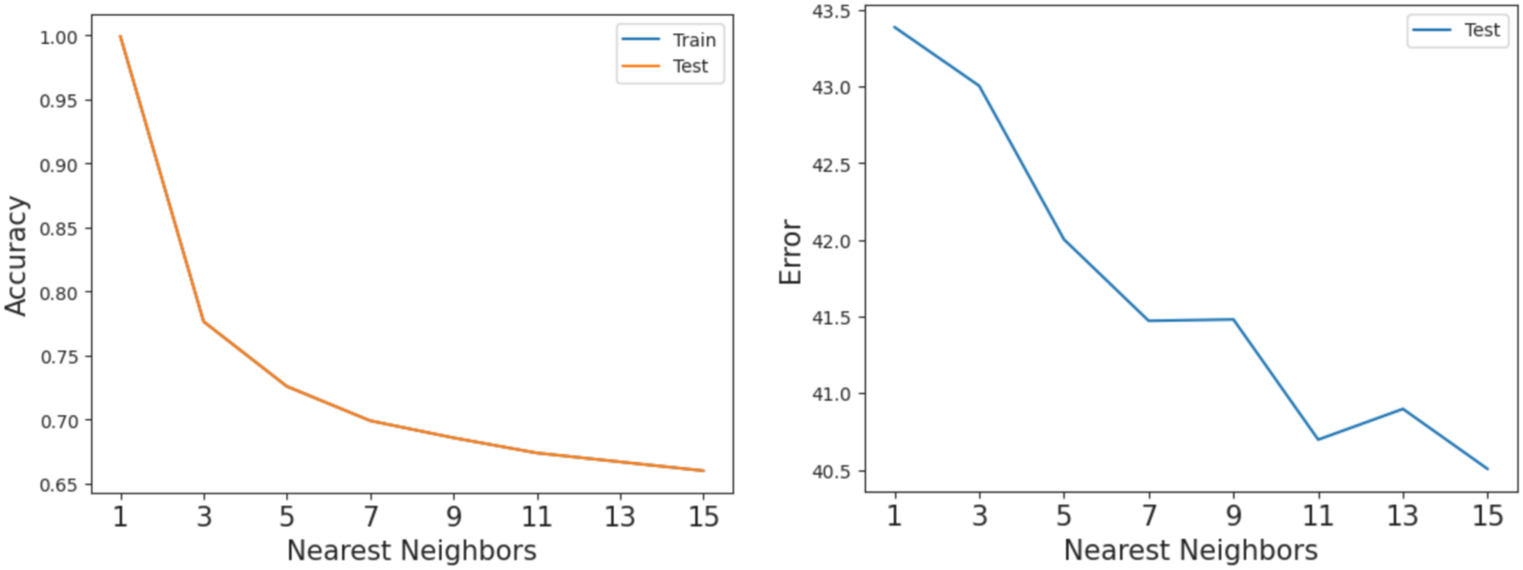
Confusion Matrix:



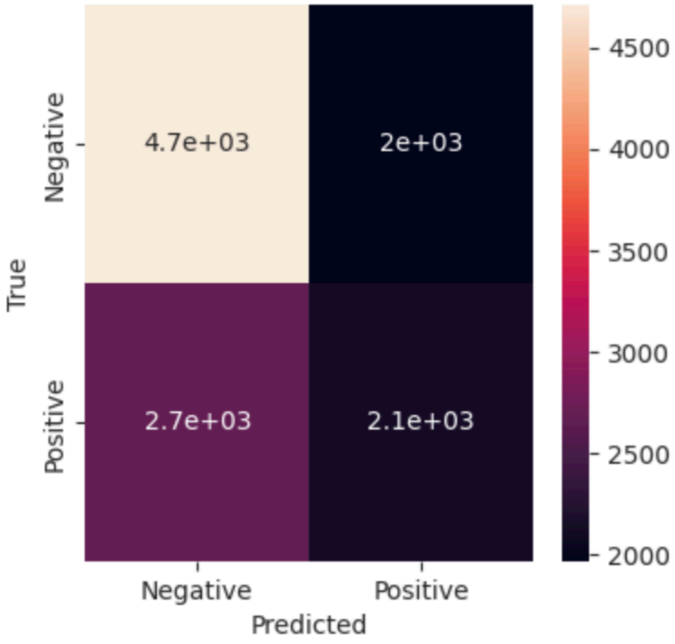
K-Nearest Neighbors:

K nearest neighbors, abbreviated as KNN, is a supervised algorithm used for classification and regression. KNN finds songs that are most similar to a given song based on various features, with the goal of predicting the class of a new data point based on the classes of its nearest neighbors. If k is chosen roughly, it will provide good results for predicting the popularity of a song on the Spotify dataset.

The performance of the KNN algorithm is determined by various factors in the data set as well as the values of k; the main advantage here is that there is no training period. Although, KNN performs poor among the other algorithms and the main reason is that Spotify contains a lot of features which results in high dimensionality which is making KNN computationally expensive and less effective. The accuracy achieved by KNN is **59.5%**. By looking at the below graphs we can interpret that as the neighbors increases, the accuracy decreases.



Confusion matrix for K-NN:

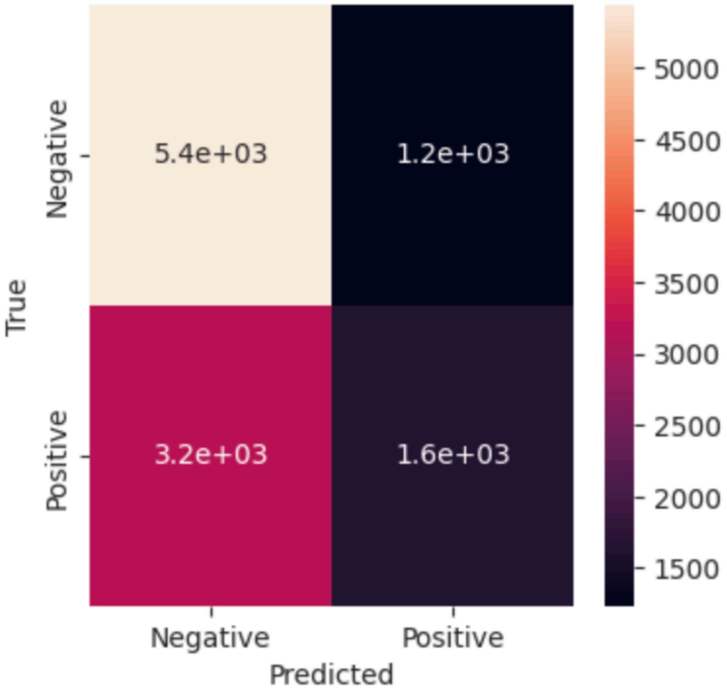


Neural Networks:

A neural Network is a computational learning system that employs a network of functions to comprehend and translate a data input in one form into a desired output, typically in another. The artificial neural network concept was inspired by human biology and how neurons in the human brain work together to understand inputs from human senses.

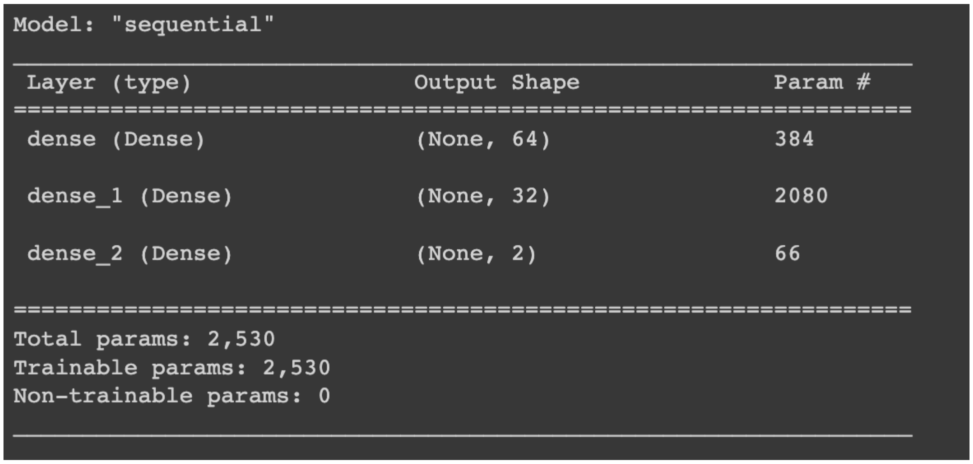
In machine learning algorithms, neural networks are just one of many tools and approaches. The neural network can be used as a component in a variety of machine learning algorithms to convert complex data inputs into a space that computers can understand. The Neural network gives the 2nd best result after Random Forest as their ability to learn complex patterns and relationships from large amounts of data achieving an accuracy of **61.64%** over 100 epochs.

Confusion Matrix:



There are three hidden layers in the neural network we implemented, in addition to the input layer and output layer. The first hidden layer has 64 neurons with a ReLU activation function. The second hidden layer has 32 neurons with a ReLU activation function. The output layer has 2 neurons with a sigmoid activation function, which is used for binary-class classification problems. So, the total number of layers in the neural network is 5, including the input layer, three hidden layers, and the output layer.

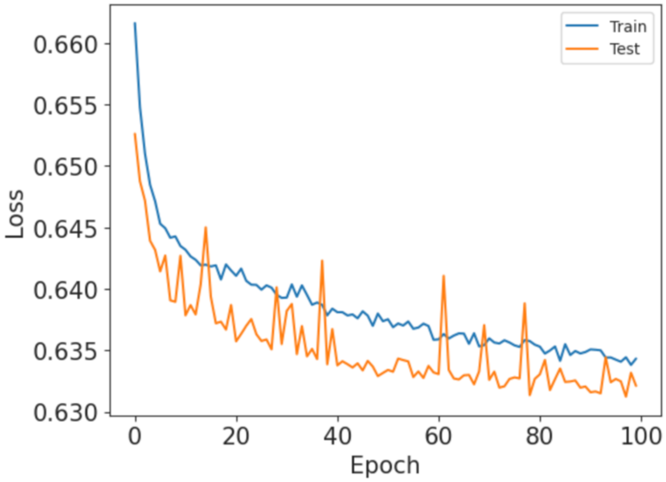
Model Summary:

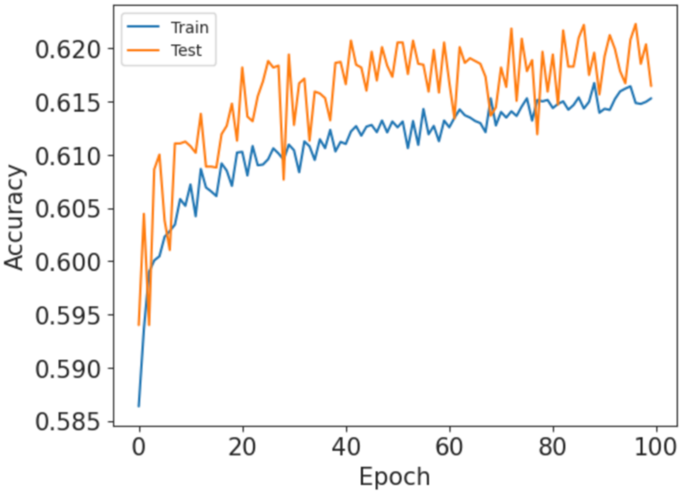


```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                384

 dense_1 (Dense)             (None, 32)                2080

 dense_2 (Dense)             (None, 2)                 66

=================================================================
Total params: 2,530
Trainable params: 2,530
Non-trainable params: 0
_________________________________________________________________
```

From the below graphs, the first graph shows the training loss and testing loss over epochs. The y-axis represents the value of the loss function, which measures the difference between the predicted output and the actual output. The x-axis represents the number of epochs, which is the number of times the network has iterated over the training dataset. The train and test loss initially started from 66.5% and 65.3% and reaches down to 63.6% and 63.4% respectively.

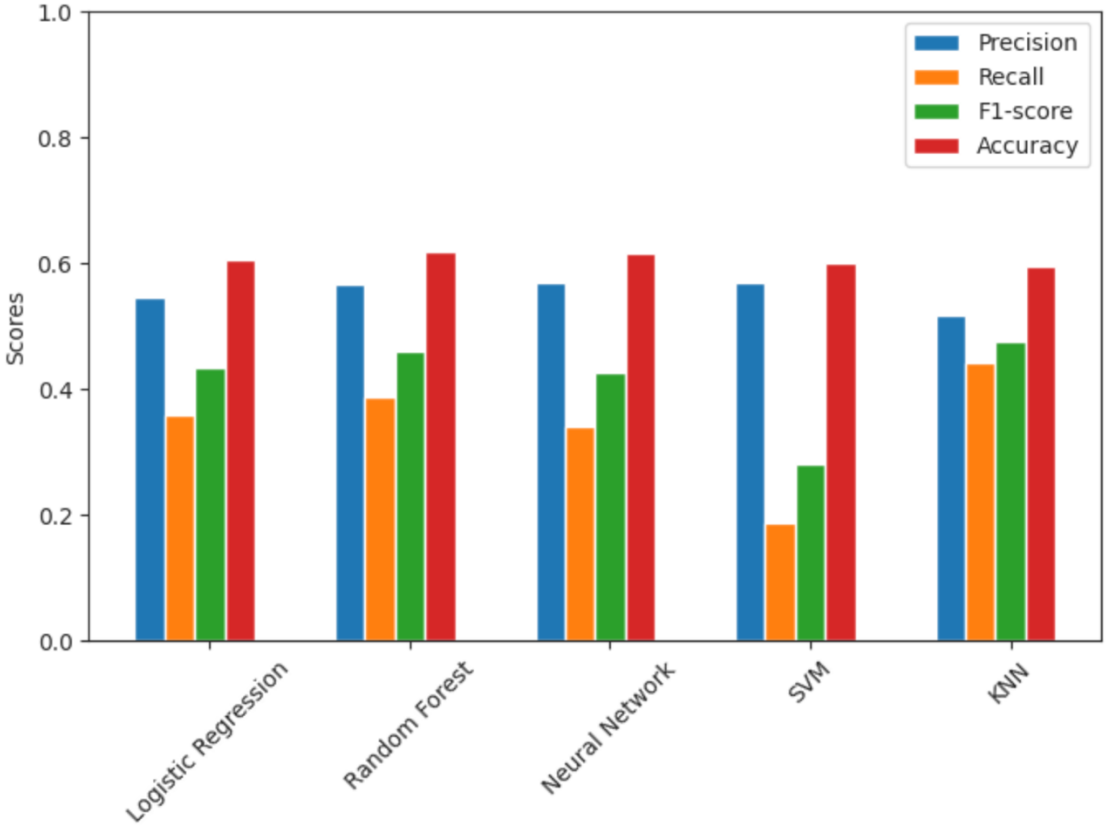
The second graph shows the training accuracy and testing accuracy over epochs. The y-axis represents the accuracy of the network, which is the proportion of correctly classified samples out of the total number of samples. The x-axis represents the number of epochs. The train and test accuracy initially started from 58.5% and 59.5% and by the end of the epochs the train and test accuracies are 61.3% and 61.5%.

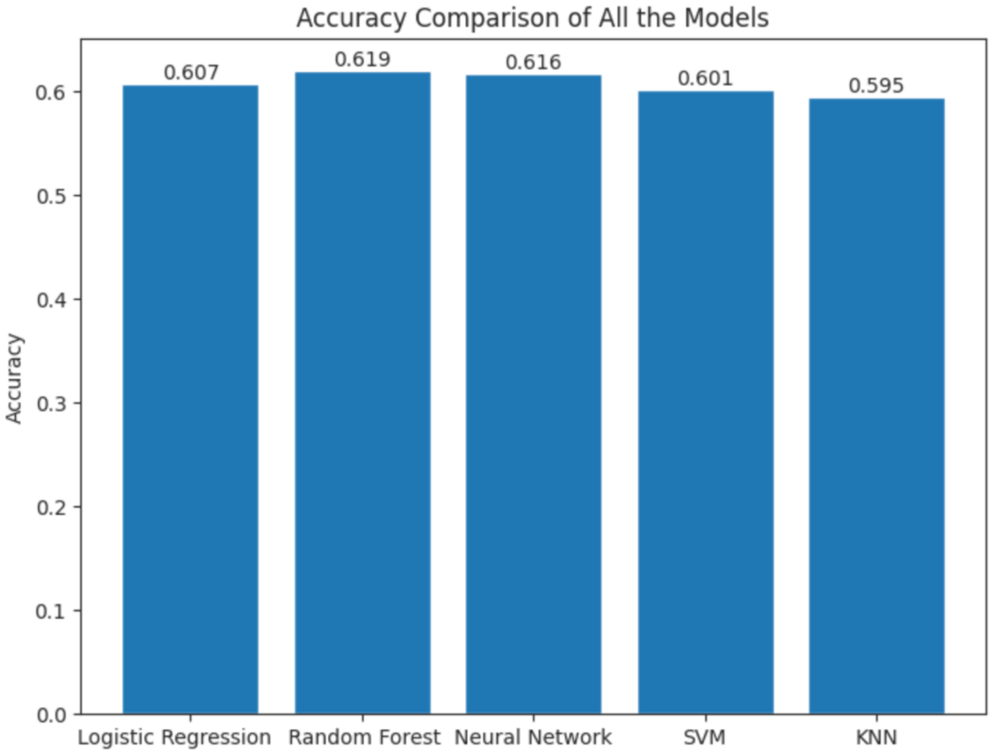




Final Thoughts:

The accuracies achieved by Random Forest and Neural Network are similar as they are handling complex data very efficiently, next the Logistic Regression and Support Vector Machines are getting similar accuracies not as good as the above two models and at last the KNN performs poor among the all as there are limitations such as number of features, complexity of data and Imbalanced data.





References:

1) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
2) https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
3) https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets
4) https://www.ibm.com/topics/neural-networks
5) https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
6) https://scikit-learn.org/stable/modules/svm.html