Bharti Chauhan (BAC180006)

Machine Learning Assignment 3

To implement the following three learning algorithms on two datasets and to perform various experiments, compare and interpret the results: -
1. Artificial Neural Networks (ANN)
2. K Nearest Neighbours
Dataset Details:
Two datasets have been used to implement the above-mentioned algorithms.
Dataset Details:
Two datasets have been used to implement the above-mentioned algorithms: -
Dataset 1: Telecom Churn Dataset
Dataset 2: Seoul Bike count Dataset

Dataset Description and Data Pre-processing:

Telecom Churn Dataset

The dataset is downloaded from Kaggle and the link is: https://www.kaggle.com/blastchar/telco-customer-churn .
The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

There are 21 columns and 7043 rows.
The target column is Churn Yes/No variable. Customer ID is of no use for machine learning operations.

Seoul Bike count

The dataset is downloaded from UCI Machine Learning repository
None of the column contains any missing value, so no missing value imputation is required.
The target variable is Rented Bike Count.
There are 14 columns and 8760 rows.
Demographics and geographical conditions features are present. Date column is of no use for machine learning operations. Categorical features have been encoded for better analysis.

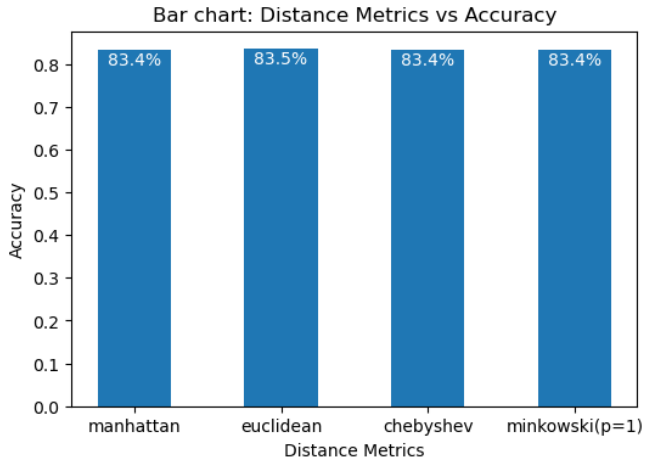
MODEL IMPLEMENTATION:

K Nearest Neighbours

The package used for implementing K Nearest Neighbours is KNeighborsClassifier which was downloaded from SciKit learn. The KNN model was implemented and the following experiments were performed:-

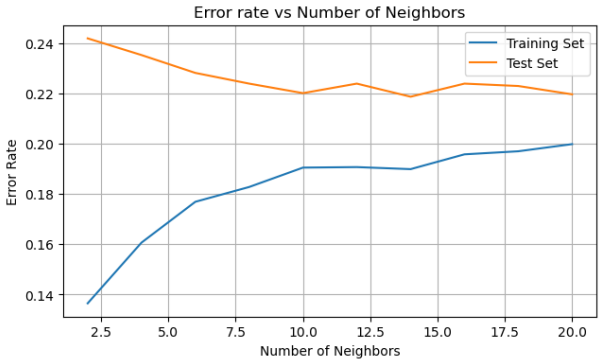
Experiments on Dataset 1 (Telecom Churn Dataset):-

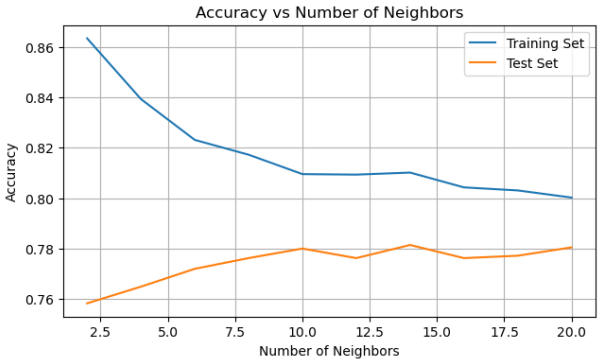
Experiment on the Distance Metric:



While all the distances are equally close, we can choose Euclidean to work for us in all the experiments since it is slightly more accurate.

Experiment on the Number of Neighbors:

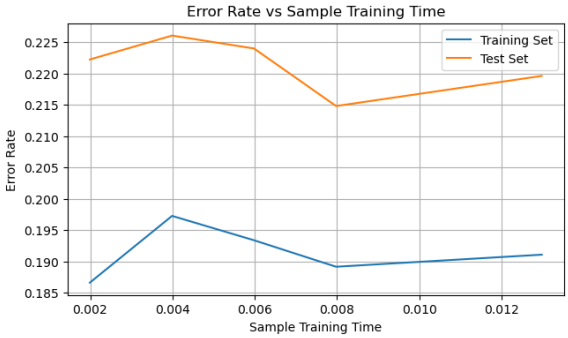




In order to find optimal number of clusters, we need to find minimum error for each cluster. We want to make sure we do not overfit the training set and check for a lowest error in test set which points out to 10 number of neighbours as the lowest error rate and with the highest accuracy.

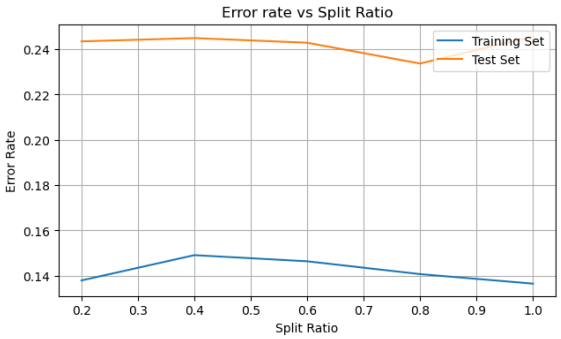
Experiment on the Sample Size and Clock Time: -

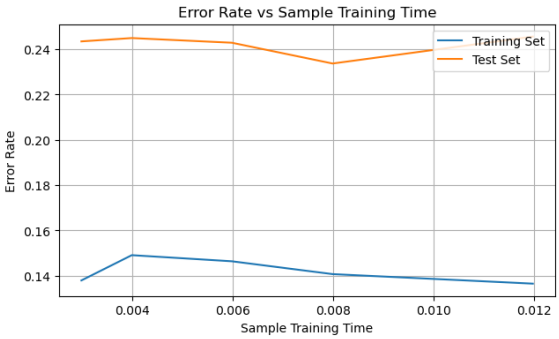
Checking for 10 neighbors we see the graph as:



For 10 clusters and knn calculation in euclidean distance we find these plots that shows error of the implemented model with different sample sizes and clock time for training the data. There is an upward and downhill trail which is somewhat expected as the number of clusters is more here.

Checking for 2 neighbors we see the graph as:





Here too the error rate increases initially and then decreases after a point.
I choose 10 neighors as that was a better estimate from previous graphs.

Artificial Neural Networks

For implementing Artificial Neural Networks “MLPClassifier” package was downloaded from Sci-kit Learn.

Artificial Neural Networks using MLPClassifier was implemented and the following experiments were performed:

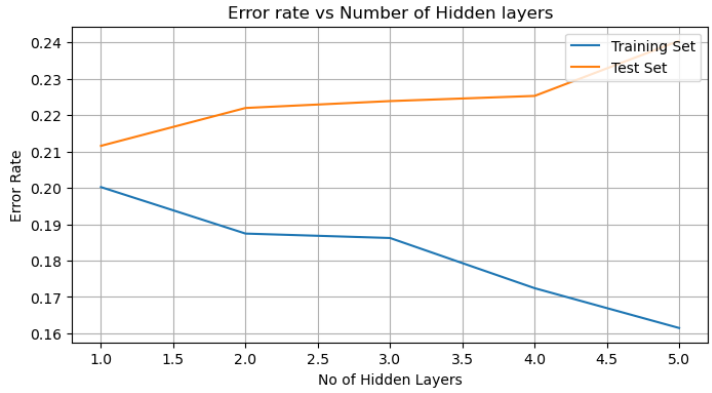
Experiment with max_iter variable in MLPClassifier:

To begin implementing ANN using MLPClassifier, an optimal value for the parameter max_iter was required for which this experiment was done. Max_iter is the maximum number of iterations. The solver iterates until convergence (determined by ‘tol’) or this number of iterations. The default value for this is 200. However, error rate was calculated for various values of this variable and the following plot was obtained to determine the optimal value of max_iter variable:

From the above graph it can be inferred that the error rate is decreasing with increase in the number of iterations for the training dataset. The minimum error in the above graph is obtained at 250 iterations where the convergence has occurred which will be considered for further analysis.

Experiment with Number of Hidden Layers:

Another important parameter here in MLPClassifier is the number of hidden layers for the ANN model built. For this, after setting the maximum number of iterations to 250, an experiment was conducted by varying the number of hidden layers in the model. The activation function used is Rectified Linear Unit and this was plotted against the error rate for each hidden layer built and the following plot was obtained:

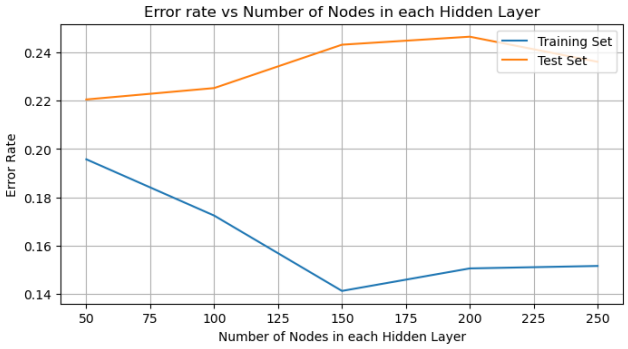


In the above plot the minimum error rate was determined at 4 hidden layers for the training as well as test data set. Hence, for the given dataset, the number of hidden layers will be taken as 4 for further analysis.

Experiment on the Number of Neurons in each hidden layer: -

Now that we have determined the optimal number of hidden layers in the neural network to be 4 with the given dataset, we must determine the number of neurons in each of these hidden layers.

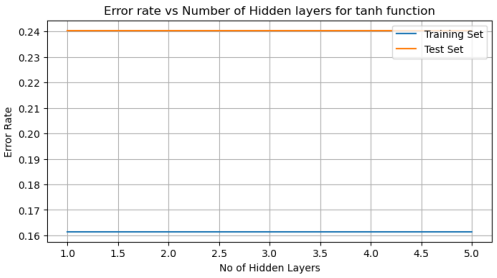
For this an experiment was conducted with different values of nodes that will be parsed to MLPClassifier to determine the optimal number of neurons in each hidden layer with minimal error rate and the following plot was obtained:

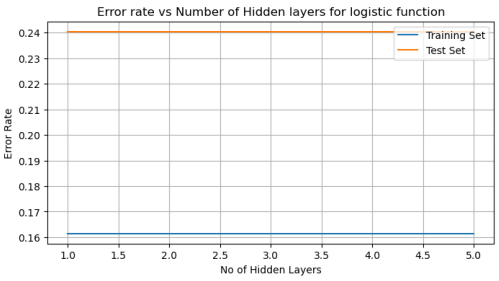


From the above plot we have determined the minimal error rate for 150 nodes in each hidden layer of the Neural Network which can be considered as optimal for conducting further experiments.

Experiment with various activation functions: -

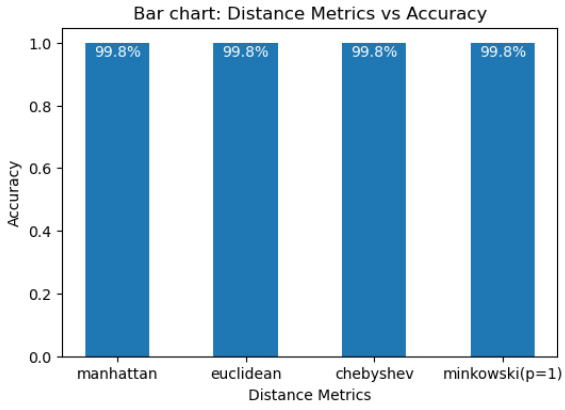
Also, experimenting with tanh and logistic activation function we get the below graphs: We can use any of the activation functions to continue with ann experiments.





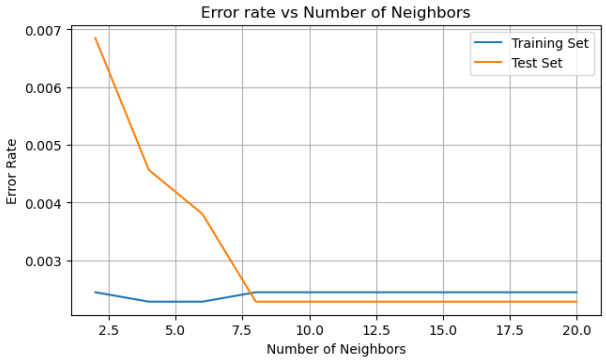
Experiments on Dataset 2 (Seoul Bike Count Dataset):-

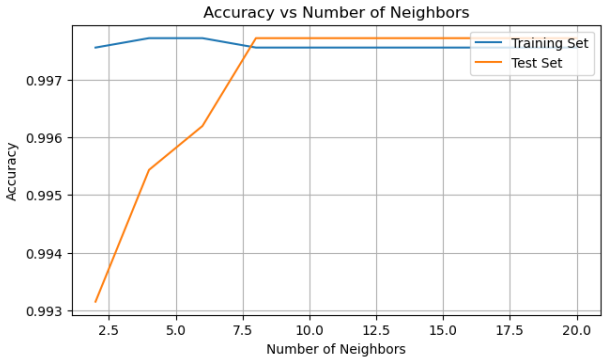
Experiment on the Distance Metrics:



While all the distances are equally close, we can choose Euclidean to work for us in all the experiments since it is slightly more accurate.

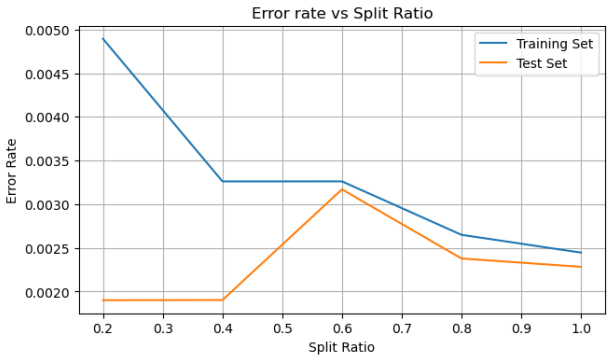
Experiment on the Number of Neighbors:

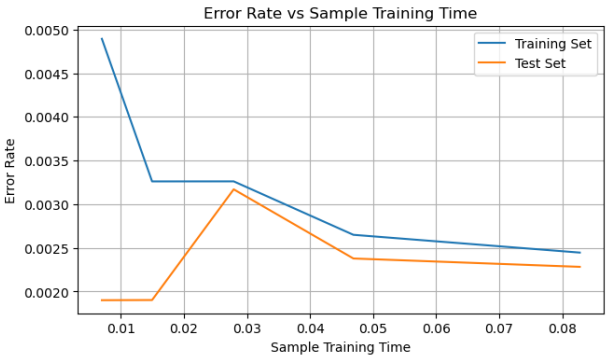




In order to find optimal number of clusters, we need to find minimum error for each cluster. We want to make sure we do not overfit the training set and check for a lowest error in test set which points out to 8 number of neighbours as the lowest error rate and with the highest accuracy.

Experiment on the Sample Size and Clock Time: -





For 8 clusters and knn calculation in euclidean distance we find these plots that shows error of the implemented model with different sample sizes and clock time for training the data. The error rate increases and decreases after a certain point. But training and test datasets error rates are closer after a certain time. This validates our nearest neigbours calculation. The test accuracy is around 75%

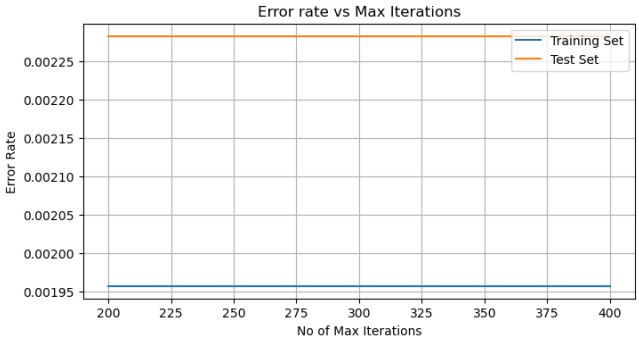
Artificial Neural Networks

For implementing Artificial Neural Networks “MLPClassifier” package was downloaded from Sci-kit Learn.

Artificial Neural Networks using MLPClassifier was implemented and the following experiments were performed:

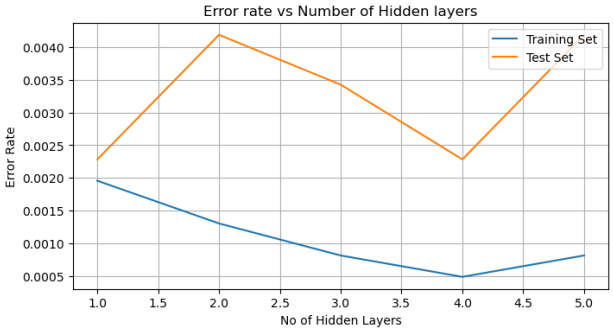
Experiments on Dataset 2 (Seoul Bike Count Dataset):-

Experiment with max_iter variable in MLPClassifier:



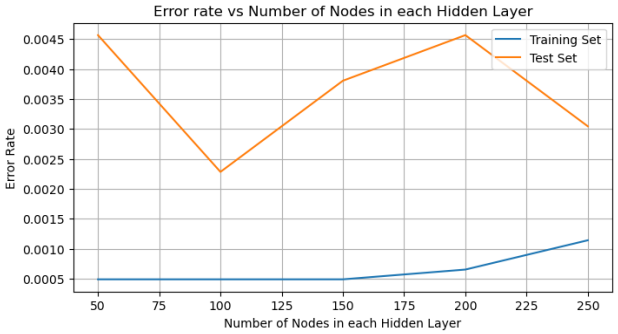
For above graph we see equal error rate for all the iterations, we consider 300 iterations for further experiments on a safer note.

Experiment with Number of Hidden Layers:



The error rate for number of hidden layers increases past 4, therefore we choose 4 hidden layers for further experimentation.

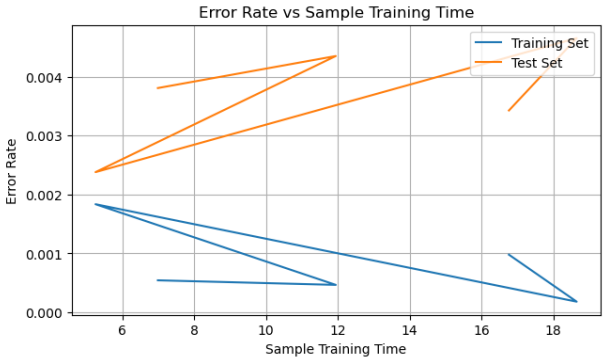
Experiment on the Number of Neurons in each hidden layer: -



Number of neurons as 150 is optimum for further analysis and reducing of error rate.

Experiment on the Length of the sample and the clock time: -

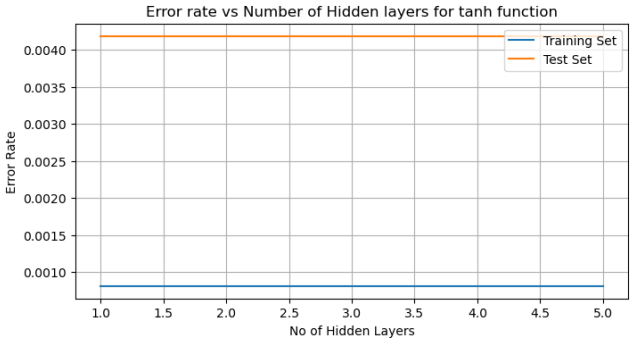
With the calculated values of maximum iterations (300), Number of Hidden Layers in the Neural Network (4) and the number of neurons in each hidden layer (150), an experiment was conducted to determine different error rates with different sample sizes and the training time for each of these samples and the following plot was obtained:

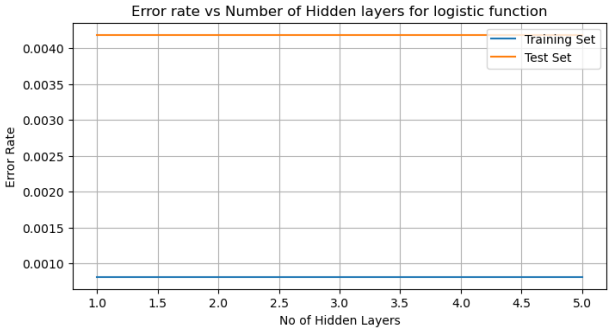


The test accuracy is 99% for this dataset.

Experiment with various activation functions: -

Also, experimenting with tanh and logistic activation function we get the below graphs: We can use any of the activation functions to continue with ann experiments.

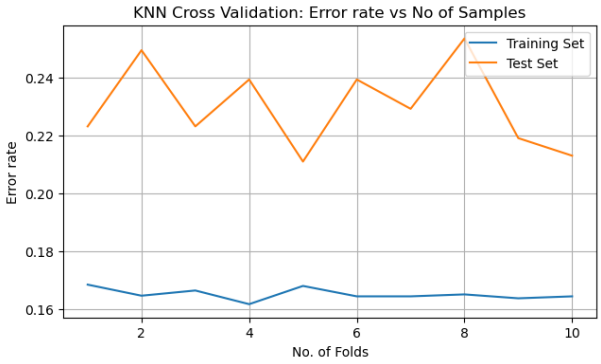


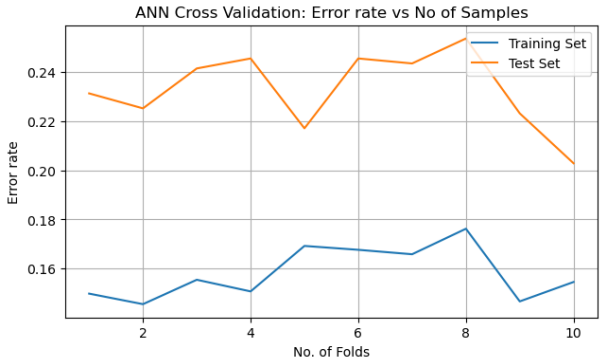


Cross Validation of KNN and ANN models for dataset1 and dataset2

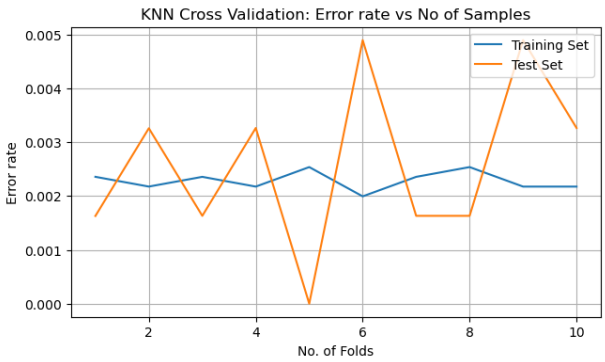
Now that we have implemented the best KNN and ANN models implemented for dataset 1 and dataset 2 by performing various experiments and parsing best values for the respective parameters, lets do Cross Validation for the given datasets with the implemented models. The package that is used for Cross Validation is “cross_validate” imported from sklearn.model selection. Cross Validation was implemented and the following plots for the errors was obtained:

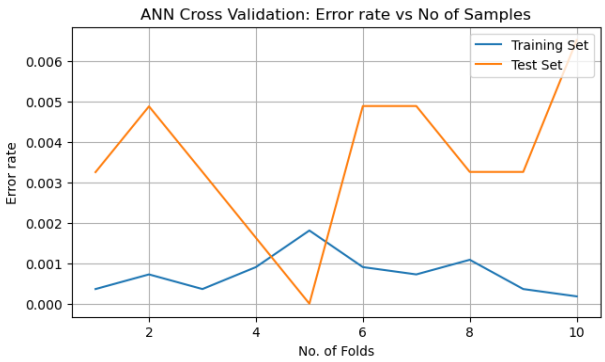
Cross Validation for Dataset 1 (Left-side graph for KNN and right-side for ANN):-





Cross Validation for Dataset 2 (Left-side graph for KNN and right-side for ANN):-





The above Cross Validation was performed with an ideal value of k=10. For dataset 1, it can be seen that the error rate is not exceeding 24% for KNN model, while with the same dataset the error rate is not exceeding 25% for ANN model. For dataset 2, we can see that the error rate is not exceeding 5% for the implemented KNN model, while with the same dataset the error rate is not exceeding 6% for the implemented ANN model.

Model Comparison, Interpretation and Conclusion

In this assignment, KNN and ANN models were implemented for the two datasets: Telecom churn and Seoul Bike dataset. Accuracy of the model well describes how good a model is. The best models developed, that were obtained by conducting various experiments on the given dataset, using which the following table with the respective accuracies was formed: -

Training Accuracy	KNN	ANN
Dataset 1(Telecom)	81	75
Dataset 2(Seoul Bike)	99	99