1 Abstract

In this project, we have implemented two models - K-Nearest Neighbours and Decision Trees - on two widely studied datasets: Hepatitis Data Set[1] and Diabetic Retinopathy Debrecen Data Set Data Set[2]. The goal is to get started with programming for Machine Learning, how to properly store the data, run the experiments, and compare different methods. The target of this project is to explore the application of the models in machine learning to real world data-sets. The report presents a descriptive overview of the performance of both the methods. We also discuss the effect of changing cost function and distance measures on the accuracy.

2 Introduction

2.1 Tasks

In this project, we have compared K-Nearest Neighbours (K-NN) and Decision Trees Classification models on Hepatitis and Diabetic Retinopathy Debrecen datasets to predict whether the person survives or not depending on different numerical features.

2.1.1 K-NN

K-NN is a non-parametric method. The quality of the results depends on the distances measured between datapoints, how distinct the classes are, and the value of the hyper-parameter K which represent the number of the nearest neighbors. K-NN is a lazy learner which keeps the data in reserve until the input data's class or the label is predicted, and does not train any internal parameters. We have used Euclidean and Manhattan method to calculate distance.

2.1.2 Decision Tree

Decision trees are widely used in machine learning as they give interpretable results. The Decision Tree model builds a decision tree where the root and internal nodes represents the features of the data set. The leaf nodes represent the class labels. We used the greedy heuristic at each step to select the optimal split of each node based on the Misclassification, Entropy, and Gini Index cost functions.

2.2 Dataset

In this project, we have used two widely studied UCI datasets – Hepatits Data Set and Diabetic Retinopathy Debrecen Data Set. The Hepatitis dataset is the smaller of the two, with 155 data points. The dataset captures 19 features of a Hepatits patient along with the binary label informing us whether the patient died or lived. The Diabetic Retinopathy Debrecen Data Set has 1151 instance of data and 9 features, of which several are irrelevant for us.

3 Datasets

3.1 Dataset 1 - hepatitis.csv (Hepatitis)

The Hepatitis dataset consists of 155 data points out of which, 23(20:6) DIE cases and 123(79:3) LIVE cases. There are a total of 20 attributes out of which, 19 can used for prediction and the Class feature is the labeled target. The hepatitis dataset has a mix of categorical and numerical features. There are six numerical features and thirteen categorical features. To render the dataset usable and without inventing data, we have removed the rows with missing values. Which resulted in a much smaller dataset with just 80 instances for model prediction. The features do not exhibit a high class correlation.

3.2 Dataset 2 - messidor_features.arff (Retinopathy)

The retinopathy dataset consists of 1151 data points with 20 attributes each, and no missing values. As with thehepatitis set, we have 19 features for prediction and 1 Class label as target, but many of the features here are unusable. The first, the binary result of quality assessment, is always 1 as only sufficient quality data points were added into the set. The models cannot be built on a constant feature. Additionally, the features 8-15 represent the same information as 2-7, but scaled differently, so there is a high correlation between the two. Equally, each group such as 8-15 are the number of MAs found at different levels of confidence, and tend to be correlated.

4 Results

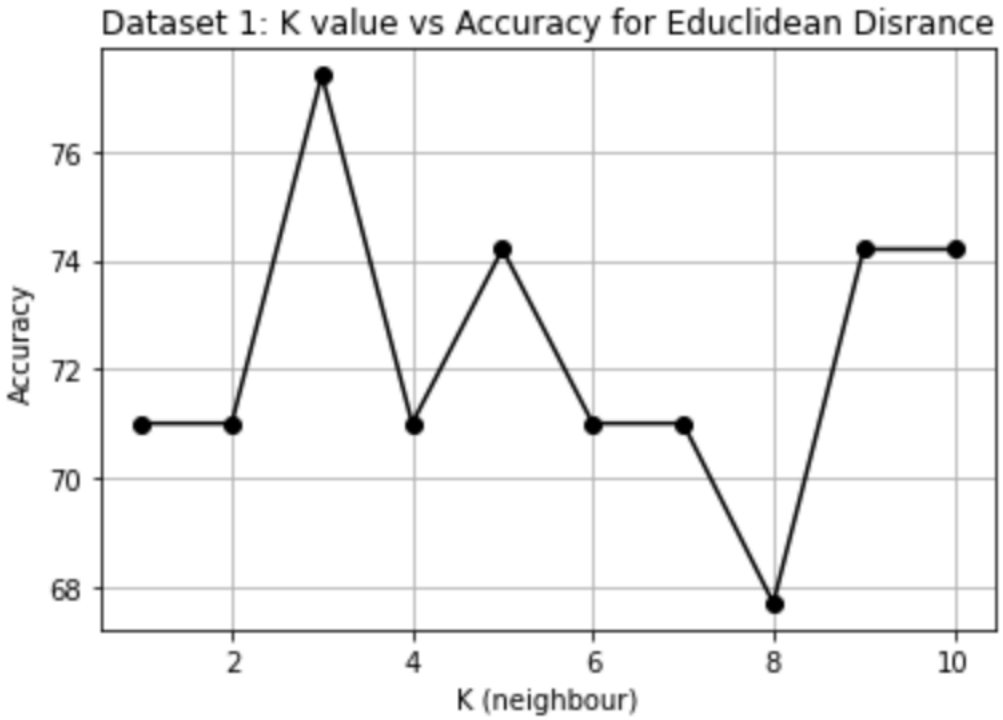
For implementation of the models, we have split the hepatitis data into test and training set 50:30, and the retinopathy (920:231).

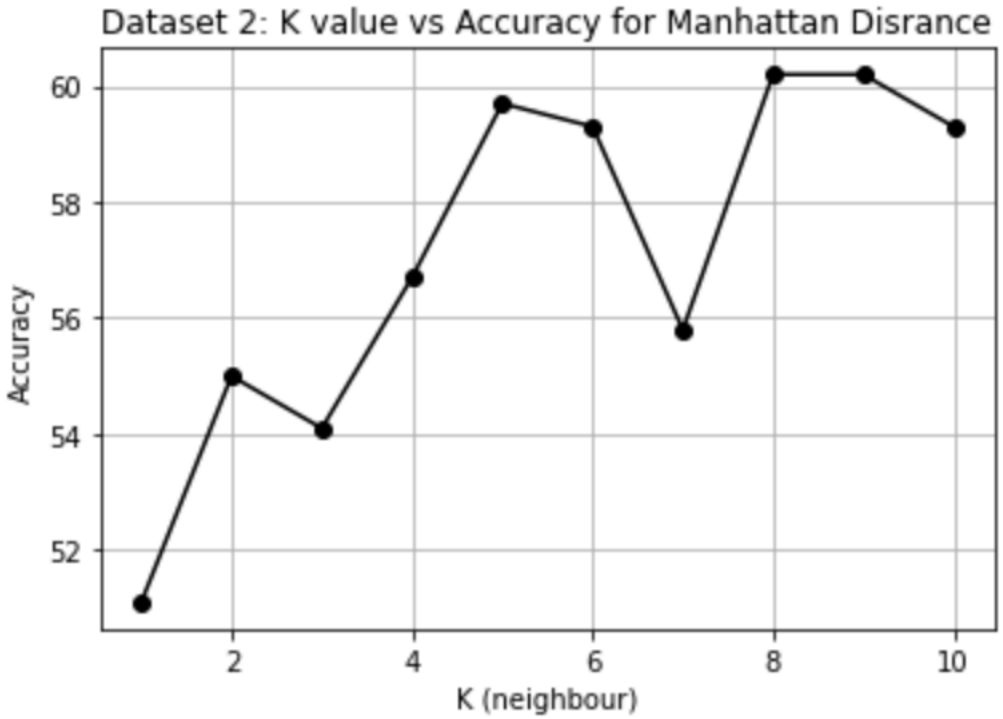
4.1 KNN Results

We used the Euclidean and Manhattan distances to calculate the distances between data points. Distance measures have been scaled to calculate the total distance. We have compared the resulting accuracy levels of K-NN classifier on both Diabetic Retinopathy Debrecen and hepatitis datasets.

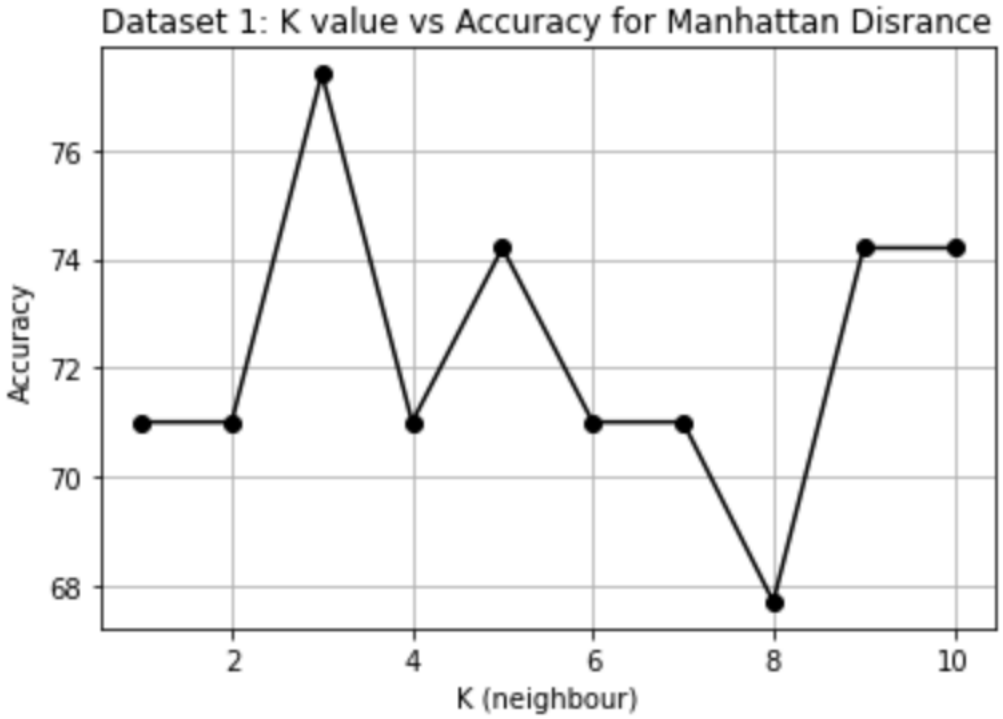
From the graph we can infer that as the hyper-parameter value increases, the test accuracies tend to increase for both datasets.

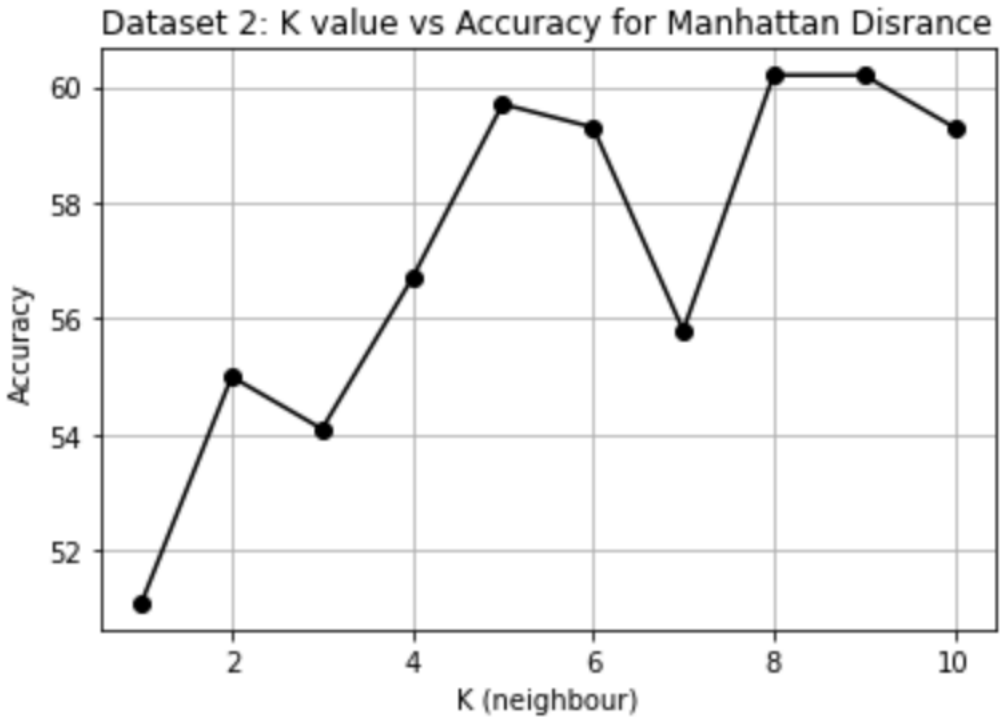
For the Euclidean Method, the hepatitis the accuracy reaches up to 77.4 at K=3 and then it remains the same, where else in Diabetic Retinopathy Debrecen, accuracy reaches up to 62.3 at k = 8 and then it remains same.





For Manhattan distance, the hepatitis the accuracy reaches up to 77.1 at K=3 and then it remains the same, where else in Diabetic Retinopathy Debrecen, accuracy reaches up to 60.1 at k = 8 and then it remains same.





We infer that there is no significant differences in the accuracy levels over different distance calculation methods.

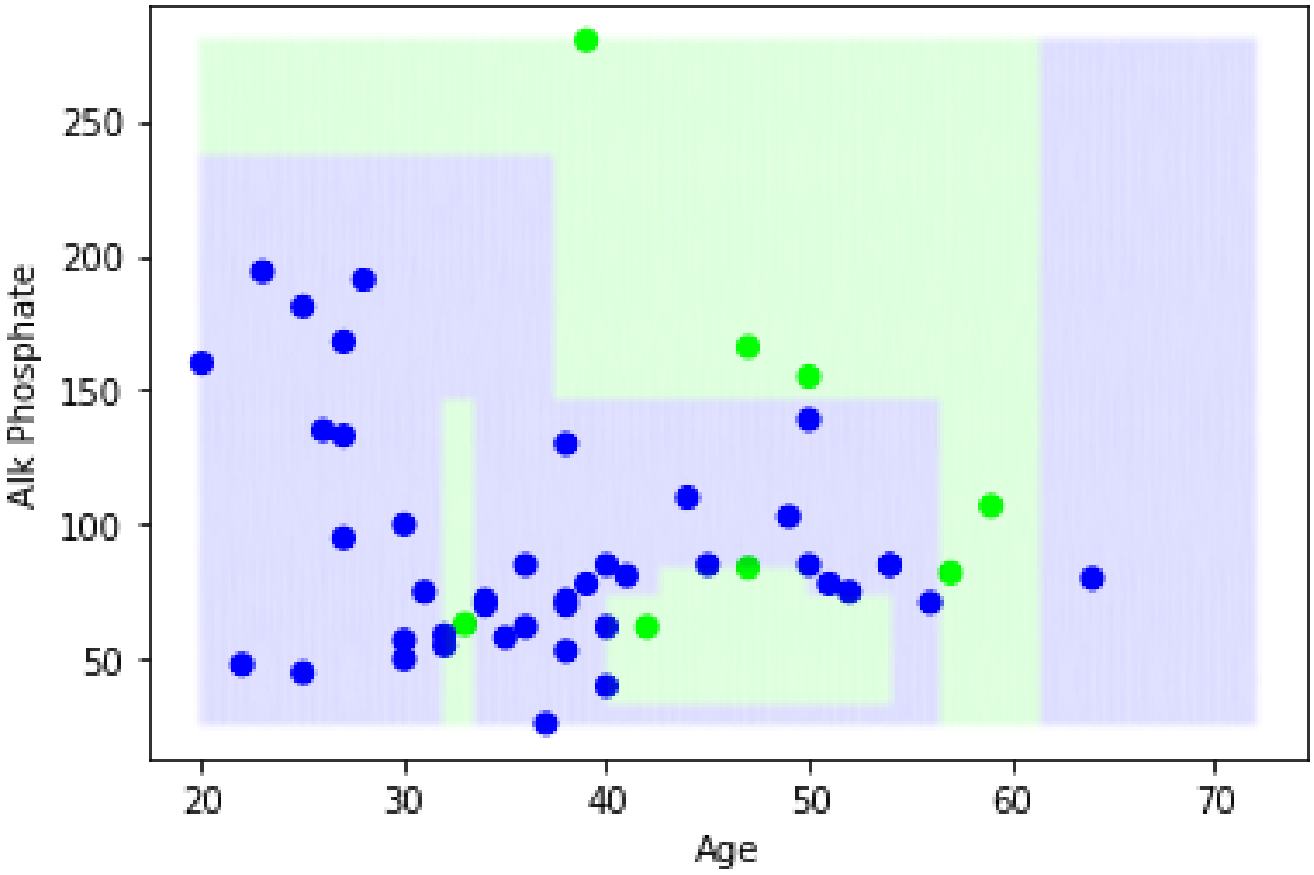
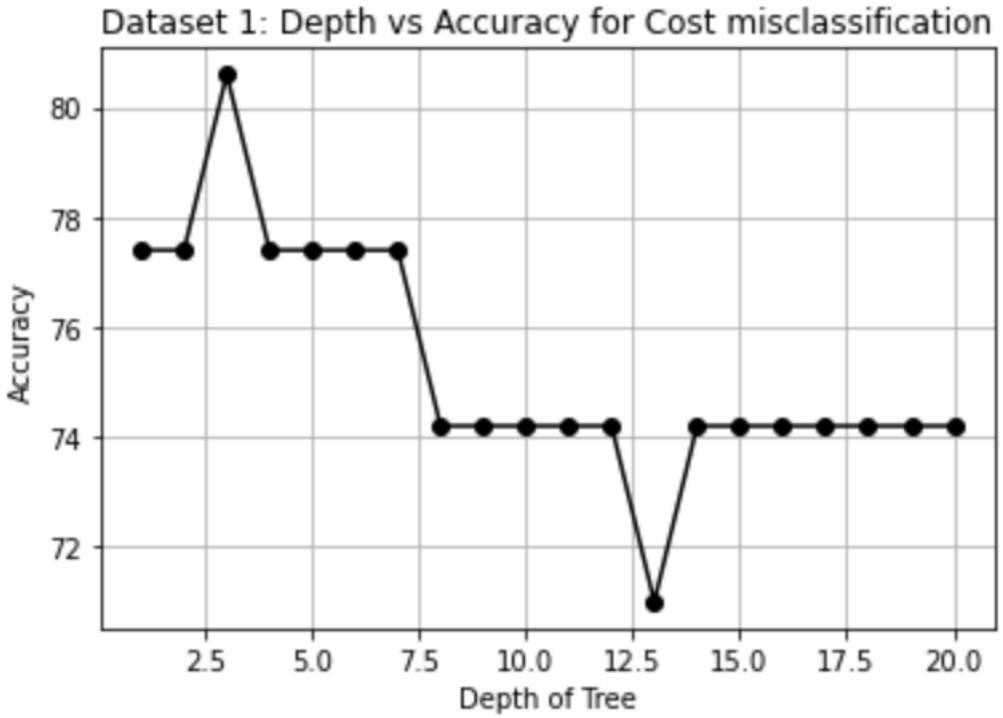
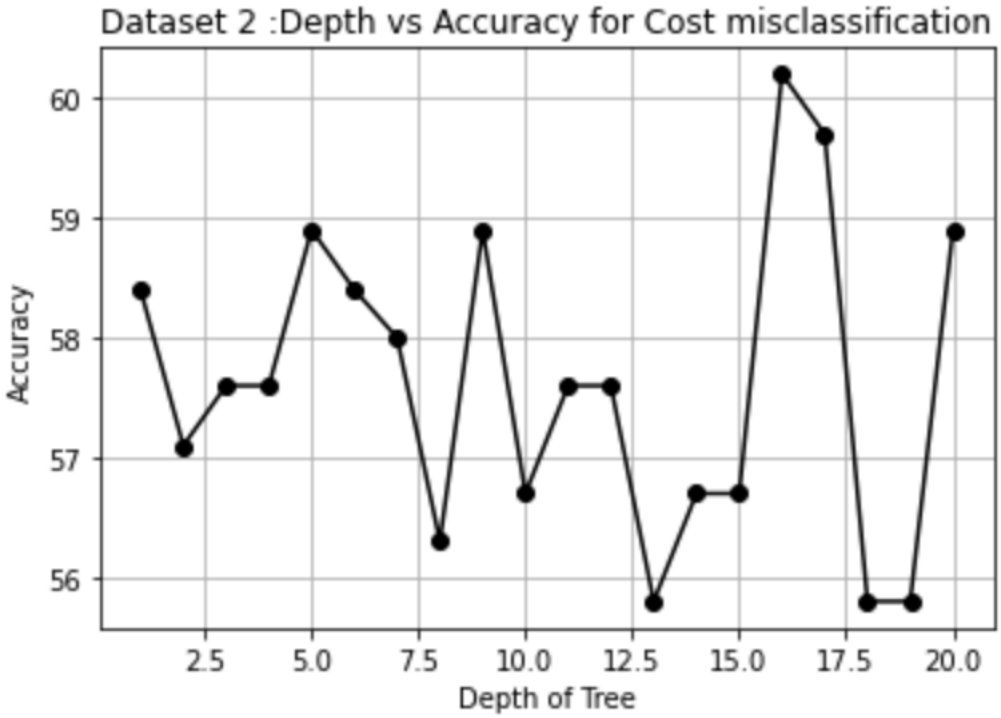


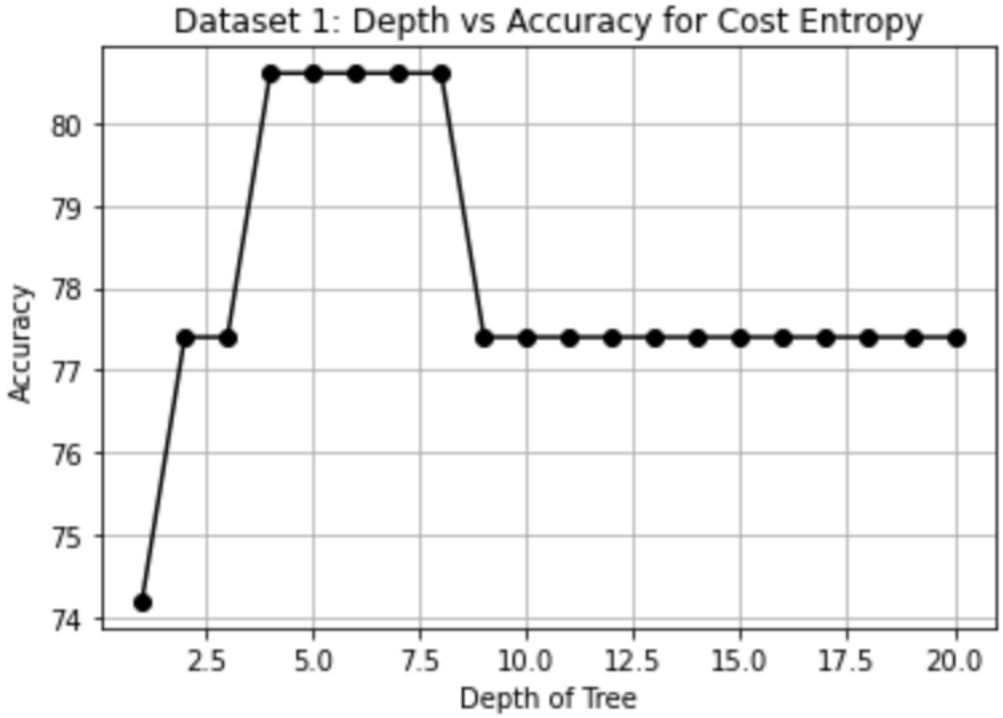
Figure 1: Decision Boundry for Hepatitis(Class: LIVE - blue, DIE - green). In general, younger patients with less alk phosphate were more likely to be alive.

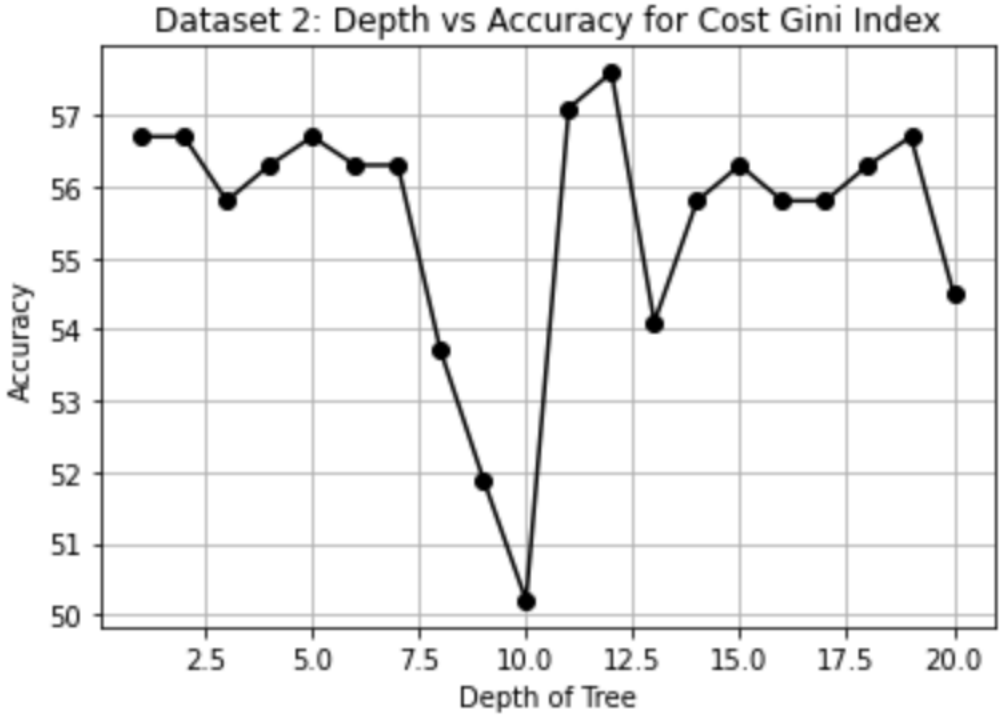
4.2 Decision Tree Results

The decision tree method is not sensitive to the scale of features and can handle out-liners. However, it tends to over-fit the data.









We can observe that as the Maximum depth increases, the training accuracy reaches to 80 percent. This is because of the overfitting of the model, and this can be improved by pruning,

which has not been considered in this work.

The test accuracy increases and then stabilises at a certain level. We have observed certain peaks in the test accuracy especially for the hepatitis dataset which might be due to the small sample size. We use the Entropy and Gini Index cost functions as as well as the Misclassification cost function. However, all the cost functions achieve a similar accuracy on both the datasets.

5 Discussion and Conclusion

The primary objective of this exercise was to explore the application of two widely used classification methods K-Nearest Neighbors and Decision Tree to real world problems. The results show that the algorithms are able to achieve a high accuracy on both the datasets.

We observed that the KNN dataset is sensitive to the scale of the features as the method depends on a distance measure to identify nearest neighbours. The features should be standardised pre-processing the data on KNN. As we increase the value of K, the model starts to generalize better, and we see an increase in test accuracy. Furthermore, we observed that both Euclidean and Manhattan distance measures give almost same results on both the datasets, which is likely due to the low dimensionality of the input. We have tried to increase the number of features used for classification but noticed no significant changes in performance while also breaking all secondary methods such as plotting.

We can observe a similar pattern for Decision tree models. The decision tree model overfits the training data if we overincrease the depth of the tree. In certain instances, the test accuracy decreases with an increase in maximum depth. A comparison of the the application of different cost functions clearly show that the entropy and Gini Index cost functions achieve a hundred percent training accuracy faster than the misclassification cost. All three methods give comparable accuracy.

For the Hepatitis dataset, we tried replacing the missing values with the mode of the columns for binary features and the means for continuous ones. However, that led to a lower accuracy compared to removing the missing values. This is possibly due to the skewing of data and is particularly noticeable due to the dataset's small size.

6 Statement of Contributions

Sagar Nandeshwar(ID: 260920948) : Worked on Data cleaning, K-NN classification model, Decision Tree Model. Wrote Abstract, Introduction, Datasets, Results, Discussion and Conclusion for the report.

Svyatoslav Sklokin (260953256): Worked on dataset cleaning, preprocessing and data standardization, statistical analysis and feature selection, along with the K-NN model. Wrote Datasets, Results, Discussion and Conclusion for the report.

Weibang Han(OP):Worked on Data cleaning, K-NN classification model, Decision Tree Model. Wrote Datasets, results and Discussion and Conclusion for the report.

7 References

[1]: http://archive.ics.uci.edu/ml/datasets/Hepatitis
[2]: https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set