MANIPAL INSTITUTE OF TECHNOLOGY

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

DMPA PROJECT NAME:

CAR PRICE PREDICTION USING HYBRID LEARNING

TEAM MEMBERS:

Sl#	NAME	Regno	RollNo	Section/Batch
1	Anmol Agarwal	190953140	42	CCE-B Batch 4
2	Ishika Jaiswal	190953138	41	CCE-B Batch 4

TABLE OF CONTENTS

ABSTRACT

Hybrid learning is based on combining two different machine learning techniques. For example, a hybrid classification model can be composed of one unsupervised learner (or cluster) to pre-process the training data and one supervised learner (or classifier) to learn the clustering result.

Prediction methods combining clustering and classification techniques have the potential of creating more accurate results than trying to predict in the linear data space.

We would be using a hybrid prediction method consisting of clustering and a decision tree classifier. We will use the predicted classes to find the nearest neighbor from the training set belonging to the same class and we will consider its label as the predicted label.

INTRODUCTION

In this project we are implementing it into two stages:
The first learning stage is used for processing the data and breaking it into multiple class labels on ranges of selling prices using weighted K-Means Clustering.
The second stage is used for making the final prediction by creating a decision tree that predicts the cluster class values. We use these predicted cluster values and find the nearest neighbor of the data point in its given cluster which is used to predict the corresponding car price
When we are dealing with the linear ranges in the output, we are essentially giving the model infinite possible outputs and thus making it difficult for the model to fully understand and map the relationship between the attributes and the respective output.

LITERATURE SURVEY

The most commonly used method for constructing a hybrid model is to combine clustering with a decision tree algorithm. For example, Gaddam et al. first proposed a hybrid model called K-means+ID3 [1]. With this method, K-means clustering is applied to partition the training samples into K disjoint clusters; an ID3 decision tree is then trained on each cluster.
The aforementioned propositions have made important contributions to the development of hybrid models. However, we found experimentally that the existing studies, and their outcomes, still have a number of shortcomings.

Most of the existing hybrid models first partition the training set into multiple clusters, and then, train a single classifier on each cluster. Finally, for each test sample, a classifier is selected according to certain classification criteria [2], [1], [3]. Intuitively, the advantage of this approach is that it breaks down a complex classification problem into many simpler problems such that each classifier is more focused on the classification of samples in a specific region. However, a potential problem caused by this type of approach is that many samples may be located near the boundaries of the clusters when given a training set that is not-well-separated [4], and such samples are often difficult to classify correctly using nearby classifiers because they are far from any cluster center

We took inspiration from these existing models and came up with our own idea in which we first cluster the dataset into the desired cluster after various statistical checks like sillhouette coefficient and then use those cluster as class labels for training the decision tree and then obtained the predicted class, ultimately finding the selling price by finding the nearest neighbor of the data point in its given cluster

METHODOLOGY

Data Preprocessing Step:

We first clean our data by dropping all the null and the unknown values from our dataset.
After that, we convert all the numerical attributes of the dataset into the float data type so that they can be used for further numerical calculations.
Even before clustering, we will split the data set into train set and test set.
Here since we are focusing on the predicting selling_price attribute of the data, we try to implement weighted K-Means by giving this attribute a higher weight by 5 times.
The rest of the attributes are reduced into a single attribute by applying Principal Component Analysis (PCA). We join our weighted selling_price with the result obtained by PCA.

Now our data has been mapped into a 2D NumPy array which can be better visualized into clusters after clustering it using K-Means.

Selecting the number of clusters:

We use the Silhouette Coefficient to select a particular number of clusters.
Silhouette Coefficient is a metric used to calculate the goodness of the clustering.
After obtaining the value of K from the silhouette coefficient we use that value to get our cluster and visualise them.

Modeling the classifier:

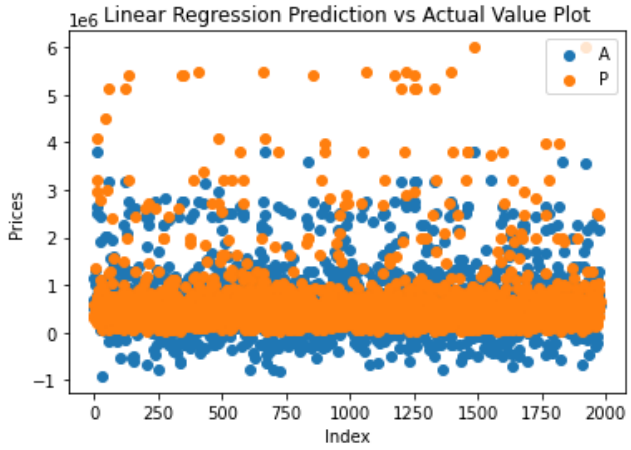
We take the price ranges of the clusters and use them as the class labels for our decision tree model.
We preprocess our dataframe to give numerical class numbers to all the nominal attributes. We also assign numerical class numbers to our price ranges according to the clusters obtained in the prior step.
We use these to create a decision tree classifier which predicts the cluster values (cluster price range) of our input data

Finding the nearest neighbour in the predicted cluster:
We searched for the nearest data point in the predicted cluster and considered its label as our predicted label, hence predicting the car price by using the hybrid machine learning algorithm.

RESULTS

We try to draw a comparative analysis between the prediction accuracy of the linear regression and our hybrid machine learning model.

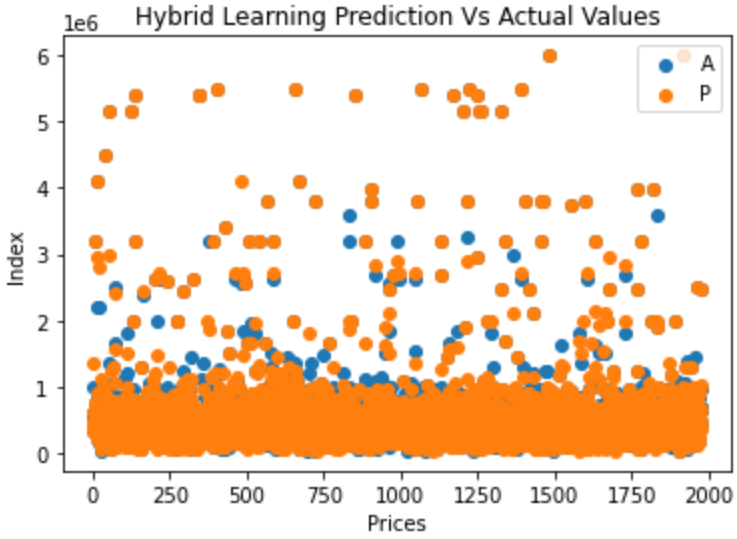
In the linear regression model, we obtain the following scatter plot between the predicted prices and the actual prices of the test data set.



The mean squared error sqrt between the predicted values (P) and the actual values (V) of the test data set was **425251.64618672157**.

In our hybrid machine learning model after performing K-Means clustering for ten clusters, we obtain the following clusters

After performing the training of our decision tree model, we obtain the following scatter plot between the predicted prices and the actual prices of the test data set.



The mean squared error sqrt between the predicted values (P) and the actual values (V) of the test data set was **172757.24672984425**

CONCLUSION

It has been observed that our hybrid machine learning model has performed significantly better than the normal regression model and the mean squared error sqrt is almost half in the hybrid machine learning model compared to the linear regression model.

REFERENCES:

[1] S. R. Gaddam, V. V. Phoha, and K. S. Balagani, "K-Means+ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods," IEEE Trans. Knowl. Data Eng., vol. 19, no. 3, pp. 345–354, Mar. 2007
[2] Y. Huang and T. Kechadi, "An effective hybrid learning system for telecommunication churn prediction," Expert Syst. Appl., vol. 40, no. 14, pp. 5635–5647, Oct. 2013
[3] T. Ma, F. Wang, J. Cheng, Y. Yu, and X. Chen, "A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks," Sensors, vol. 16, no. 10, pp. 1701–1724, Oct. 2016