Machine Learning (CS60050)

Assignment 1

Decision Tree and Bayesian (Naïve Bayes) Classifier

Group 43
Umang Singla - 20CS10068
Utsav Mehta - 20CS10069

Dataset: (Insurance Policy Dataset)

This dataset contains the predictions of whether a customer is interested in vehicle insurance or not based on the measurements of the certain attributes. The dataset has the following attributes:

- **Gender -** Gender of the customer
- **Age -** Age of the customer
- **Driving License -** Whether a customer has Driving License or not
- **Region Code -** Unique code for the region of the customer
- **Previously Insured -** 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
- **Vehicle Damage -** 1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
- **Vehicle Age -** Age of the Vehicle
- **Annual Premium -** The amount customer needs to pay as a premium in the year
- **Policy Sales Channel -** Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
- **Vintage -** Number of Days, Customer has been associated with the company
- **Response -** 1 : Customer is interested, 0 : Customer is not interested

The dataset has 131689 training examples and is provided in '.csv. Format.

1. Decision Tree

Tasks:

1. Implement a Decision Tree to predict whether a customer would be interested in vehicle insurance using ID3 algorithm and Information gain as impurity measure.
2. Report the best test accuracy and tree depth over 10 random 80/20 train-test splits.
3. Prune the tree obtained from Task-2 to obtain a tree with max accuracy.
4. Study the depth effects on the accuracy of the Decision tree.

The Decision Tree Algorithm Used

We have used the ID3 algorithm for constructing the decision tree. However, the conventional ID3 algorithm is restricted to attributes that take on discrete values. To take into account the continuous range of values into a discrete set of intervals we have divided the continuous data into 10 classes. The data in the ith section lie between $[a_i , b_i]$.
where,
a_i = *min + (i-1)*width*
b_i = *min + i*width*
min = minimum possible value of the attribute
max = maximum possible value of the attribute
width = (max - min)/10
Thus, we have discretized the continuous data. Now, the conventional ID3 algorithm can be applied on this data.

Some Important Terms and Definitions

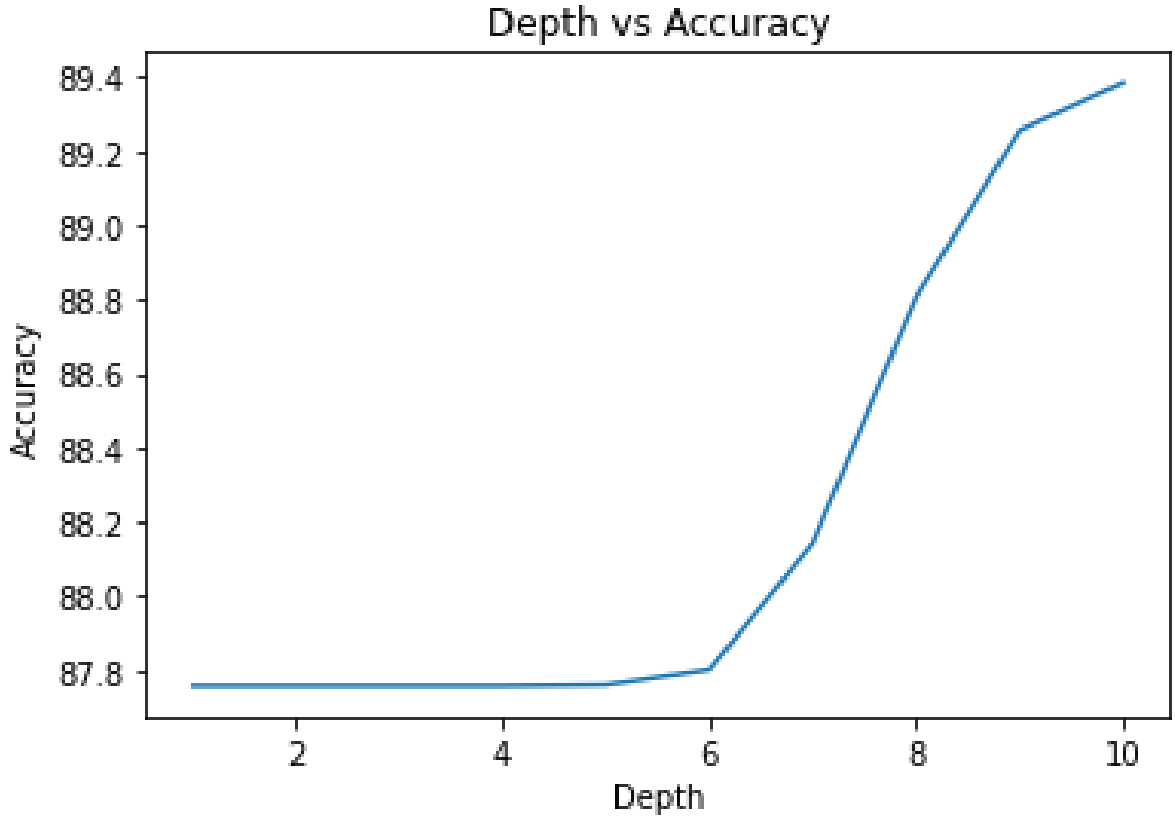
- **Entropy:** For a collection S, $Entropy(S) = -\Sigma p(i) log_2 p(i)$.
- **Information Gain:** The information gain is the reduction in entropy after choosing an attribute A. Mathematically, it can be written as:

$$InformationGain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v).$$

- **Accuracy:** $Accuracy = \frac{No.\ of\ examples\ correctly\ classified}{Total\ no.\ of\ examples}$

Results:

1. **Best Test Accuracy over 10 test-train data splits:** 86.41886248006682 %
2. **Depth of the tree:** 10
3. **Accuracy after pruning:** 87.74394411117017 %
4. **Depth vs Accuracy Plot:**



2. Naive Bayes

Tasks:

1. Randomly divide Dataset C into 80% for training and 20% for testing. Encode categorical variables using appropriate encoding methods.
2. A feature value is considered as an outlier if its value is greater than mean + 3 x standard deviation ($\mu + 3 \times \sigma$). A sample having maximum such outlier features must be dropped. Print the final set of features formed. Normalize the features as required.
3. Train the Naïve Bayes Classifier using 10-fold cross validation (no packages to be used for Naïve Bayes Classifier). Print the final accuracy.
4. Train the Naïve Bayes Classifier using Laplace correction on the same train and test split. Print the final accuracy.

Theory:

1. Naive Bayes uses Maximum Likelihood Estimation methods for calculating probabilities.
2. Naives Bayes with Laplace corrections uses an ‘alpha’ to estimate the probability of a data value, using Maximum a Posteriori Estimation.
3. K-fold cross validation divides the training dataset into k folds and uses one of the folds as test, the rest as train. The best fit among the k-models is used on the testing dataset to determine the final accuracy.

Procedure and Results:

1. **Procedure** - Split the dataset into two parts in the ratio 4:1 (or 80:20). The bigger of the two parts is used as the Training Dataset and the smaller part is used as the Testing Dataset. Categorical variables are also encoded using the Label Encoding Algorithm, to convert their values to numeric data, in order to apply the model.

Result - The dataset is split into four parts - X_train and y_train (used as the training dataset), X_test and y_test (used as the testing dataset).The following features were encoded:

- ➔ Gender
- ➔ Vehicle_Age
- ➔ Vehicle_Damage

2. **Procedure** - Calculate the mean and deviation for each feature in the dataset. If for any data point, the absolute value between the mean and it is greater than 3 x standard deviation, then we need to remove that data point from the dataset. Normalize the features with continuous data values.

   Result - The dataset size is reduced from 131689 to 130725 after the outliers are removed.
   The extracted feature names are printed in the output file.
   The following features are normalized -

   - ➔ Age
   - ➔ Region_Code
   - ➔ Annual_Premium
   - ➔ Policy_Sales_Channel
   - ➔ Vintage

3. **Procedure** - Divide the training dataset into 10 parts (for 10-fold cross validation) and choose 9 folds as the training data, and one of them as testing data, determine the accuracy of this test set. Similarly, choose each fold as the testing set and find the maximum accuracy among the generated trained models. Then use the model with maximum accuracy and test the testing data (20% of the original dataset) to find the accuracy.

   Result - The obtained accuracy for the naive bayes algorithm with 10-fold cross validation is - 76.57295850066934 / 100

4. **Procedure** - Apply the Laplace correction on the data points, followed by the Naive Bayes algorithm. Train the dataset with this algorithmic improvisation and test again the test set split.

   Result - The obtained accuracy for the naive bayes algorithm with laplace correction is - 76.62268120099445 / 100

The final accuracies obtained are -

1. For Naive Bayes with 10-fold repetition - 76.57295850066934 **(76.54%)**
2. For Naive Bayes with laplace correction - 76.62268120099445 **(76.62%)**