Contents

Project Abstract

Three classifiers namely; Naïve Bayes, ID3 Decision Trees and Random Forest are to be implemented on a dataset of discrete valued data instances. A comparison of the accuracies of the three classification models are to be defined. Confusion matrices of the three classification models are to be created. And the classification accuracies of the classification models when trained on 50%, 75% and 90% is to be reported. Initially, the models are to be trained on 60% and tested on 40% of the given dataset.

For computation and building models, Python will be used. Google Collab is utilized as the web IDE.

Dataset description

The dataset used in this project is named 'Car Evaluation Database', created in June 1997. The authors and donors of this dataset are Marko Bohanec and Blaz Zupan.

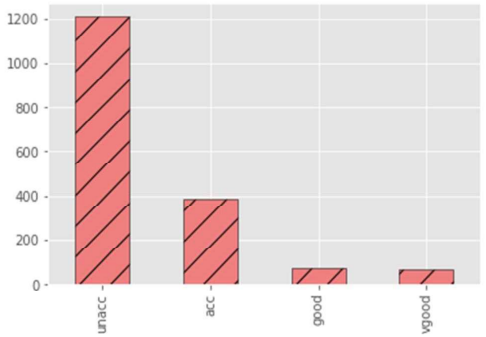
This dataset consists of a total of 1728 instances, 6 attributes and 1 output class variable. The 6 attributes contain 'buying' , 'maint' , 'doors' , 'persons' , 'lug_boot' , 'safety'. The attribute description of the 6 attributes are as follows;

```
buying: vhigh, high, med, low.

maint: vhigh, high, med, low.

doors: 2, 3, 4, 5more.

persons: 2, 4, more.

lug_boot: small, med, big.

safety: low, med, high.
```

Further, the class values are `unacc, acc, good and vgood.`

The 'Cars Evaluation Dataset' is discrete categorical data.

The individual value counts for the output class are follows;




```
unacc     1210
acc        384
good        69
vgood       65
```

Label Encoding.

The concept of 'label encoding' was leveraged in this project study. It is an important data pre-processing task which converts the string labels into numeric form, so that it can be expressive, for the models to compute.

Naïve Bayes Classifier

The Naïve Bayes classifier is a classification model based on the Bayes theorem. This classifier is known for achieving great computational accuracy. The Naïve Bayes classifier is labelled 'naïve' due to the fact that it is based on the assumption that, all the input instances are independent in nature.

In the following sections, three variants of the Naïve Bayes Algorithm and their outcomes will be discussed.

(I) Gaussian Naïve Bayes.

The Gaussian Naïve Bayes is a variant of Naïve Bayes that supports the computation of datasets with continuous values and works with the assumption of the input classes having a normal distribution.

Since, our dataset is discrete categorical data, the Gaussian Naïve Bayes is not a great fit for the same.

The accuracies after implementing the Gaussian Naïve Bayes are as follows;

Split / Model	60% - 40%		50% - 50%		75% - 25%		90% - 10%
Gaussian NB	65.173%		63.35%		64.35%		68.208%

The highest accuracy can be observed when 90% of the data is assigned for training. And the lowest accuracy is found when 50% of the data is assigned for training.

(II) Multinomial Naïve Bayes.

The Multinomial Naïve Bayes is a variant of the Naïve Bayes that works well with the datasets with multinomial distribution. Being an adaptation of the Naïve Bayes, it also considers that the input features are not related to features.

The accuracies after implementing the Multinomial Naïve Bayes are as follows;

Split / Model	60% - 40%		50% - 50%		75% - 25%		90% - 10%
Multinomial NB	71.387%		70.370%		72.916%		72.832%

The highest accuracy can be observed when 75% of the data is assigned for training. And the lowest accuracy is found when 50% of the data is assigned for training.

(III) Categorical Naïve Bayes.

The Categorical Naïve Bayes is a variant of Naïve Bayes which is well aligned with the datasets which contain discrete input features and when the features are categorically distributed. Since our dataset is discrete categorical data, the Categorical Naïve Bayes is a good fit for the same.

The accuracies after implementing the Categorical Naïve Bayes are as follows;

Split / Model	60% - 40%		50% - 50%		75% - 25%		90% - 10%
Categorical NB	85.260%		87.5%		85.416%		82.080%

The highest accuracy can be observed when 50% of the data is assigned for training. And the lowest accuracy is found when 90% of the data is assigned for training.

The confusion matrix for the Categorical Naïve Bayes, when 60% of the data is assigned for training is as follows;

108	3	25	0
18	3	0	1
33	1	471	0
22	0	0	7

Total: 692

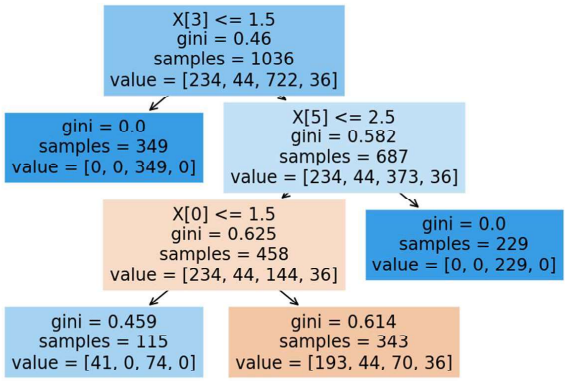
Correctly classified: 589

Misclassified: 103

Decision Tree ID3.

The Decision Tree ID3 is a classification model which divides the input features into two or more sets or groups at each iteration/step. The ID3 uses a top-down greedy approach for the computation.

The following is the tree generated when assigning 40% of the data for testing.



The accuracies after implementing the Decision Tree, ID3 are as follows;

Model \ Split	60% - 40%		50% - 50%		75% - 25%		90% - 10%
Decision Tree ID3	78.612%		79.745%		80.092%		71.098%

The highest accuracy can be observed when 75% of the data is assigned for training. And the lowest accuracy is found when 90% of the data is assigned for training.

The confusion matrix for the Decision Tree ID3, when 60% of the data is assigned for training is as follows;

127	0	31	0
32	0	0	0
50	0	417	0
35	0	0	0

Total: 692

Correctly classified: 544

Misclassified: 148

Random Forest.

Random forests are the ensemble learning models used for classification of datasets. It creates decision tree from the input dataset and produces the predictions. The random Forest uses the extension of the concept of bagging, which randomly selects subsets of the input features from the dataset.

Model \ Split	60% - 40%	50% - 50%	75% - 25%	90% - 10%
Random Forest	93.208%	93.287%	95.370%	98.265%

The highest accuracy can be observed when 90% of the data is assigned for training. And the lowest accuracy is found when 50% of the data is assigned for training.

The confusion matrix for the Random Forest, when 60% of the data is assigned for training is as follows;

126	1	28	3
6	21	0	1
6	1	474	0
1	0	0	24

Total: 692

Correctly classified: 645

Misclassified: 47

Project Analysis.

The Naïve Bayes classifier was the first classifier which was implemented. The three variations of the same were worked upon namely; Gaussian Naïve Bayes, Multinomial Naïve Bayes and the Categorical Naïve Bayes. The highest accuracy for the Gaussian Naïve Bayes was when 90% of the data was assigned for training. The accuracies from the different training sample sizes for the Gaussian Naïve Bayes, had a standard deviation of 1.814. The highest accuracy for the Multinomial Naïve Bayes was when 75% of the data was assigned for training. The accuracies from the different training sample sizes for the Multinomial Naïve Bayes, had a standard deviation of 1.060. The highest accuracy for the Categorical Naïve Bayes was when 50% of the data was assigned for training. The accuracies from the different training sample sizes for the Multinomial Naïve Bayes, had a standard deviation of 1.939.

The Decision Tree ID3 classifier had the highest accuracy can be observed when 75% of the data is assigned for training. For the results of the different training-testing splits, the standard deviation was 3.671.

The highest overall accuracies were found while implementing the Random Forest classifier. The highest accuracy was achieved when allocating 90% of the data for training. For the results of the different training-testing splits, the standard deviation was 2.057.

Conclusion.

Altogether, all the three classification models varied substantially and in other situations slightly, in their output results to changes in the training and testing size. The ‘Cars Evaluation Dataset’ is discrete categorical data. Hence, the machine learning models which handle categorical data, best suited the dataset. The highest classification accuracies were generated by the Random Forest classifier, where the accuracy reached 98.27%. And the lowest accuracies were generated by the Gaussian Naïve Bayes, with an average accuracy of 65.270%

The Decision Tree ID3 classifier proved to be the most sensitive to the training and testing split sizes. Great differences can be observed while adjusting the training sizes for the

Decision Tree ID3 classifier. The least sensitive/affected classifier to the training and testing split sizes was the Multinomial Naïve Bayes classifier.

References.

[1] Ramezan, C.A.; Warner, T.A.; Maxwell, A.E.; Price, B.S. Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. Remote Sens. 2021, 13, 368. https://doi.org/10.3390/rs13030368

[2] Płoński, P. (2020, June 22). *Visualize a Decision Tree in 4 Ways with Scikit-Learn and Python*. MLJAR. https://mljar.com/blog/visualize-decision-tree/

[3] Horbonos, P. (2020, February 15). *Comparing a variety of Naive Bayes classification algorithms*. Medium. https://towardsdatascience.com/comparing-a-variety-of-naive-bayes-classification-algorithms-fc5fa298379e

[4] Horbonos, P. (2020, February 15). *Comparing a variety of Naive Bayes classification algorithms*. Medium. https://towardsdatascience.com/comparing-a-variety-of-naive-bayes-classification-algorithms-fc5fa298379e

[5] *Introduction to Random Forest in Machine Learning*. (2020, December 11). Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-ducation/introduction-to-random-forest-in-machine-learning/