Group_3

Name: Daniel Ndubisi Eze

Student ID: 22009281

Naive Bayes Classification Report

Introduction to the Topic

Naïve Bayes Classifier is one of the most simple and effective classification algorithms which assumes that features for the description of observation are conditionally independent, given the class label. It can also be viewed as a simple and probabilistic model used for classification tasks that are based on Bayes' theorem which make assumptions of independence between attributes or variables used to describe data point. The Naïve part of the classifier was derived from assumption that all features are unrelated. The setback with this classifier arises when there is violation of its assumption of independent features.

Let's consider mathematical representation of Naïve Bayes classifier model based on Bayes' Theorem to calculate the conditional probability of a class © given a set of features (N). P(C|N) = probability of class C given the features N, P(N|C) = likelihood of observing N given class C

P© = Prior probability of class C, P(X) = probability of observing features N. (Evidence)

Then: $P(C|N) = \frac{P©P(N|C)}{P(N)}$

In the other hand, Decision Tree Classifier is a Tree-like model which splits data based on features. They are very useful in classification and regression problems. Random Forest classifier is an ensembled Decision Trees. They are more accurate in classification and achieve this by constructing numerous decision trees during the training process and ultimately predicting the class that appears most frequently across these individual trees. The main disadvantage is overfitting and less interpretable.

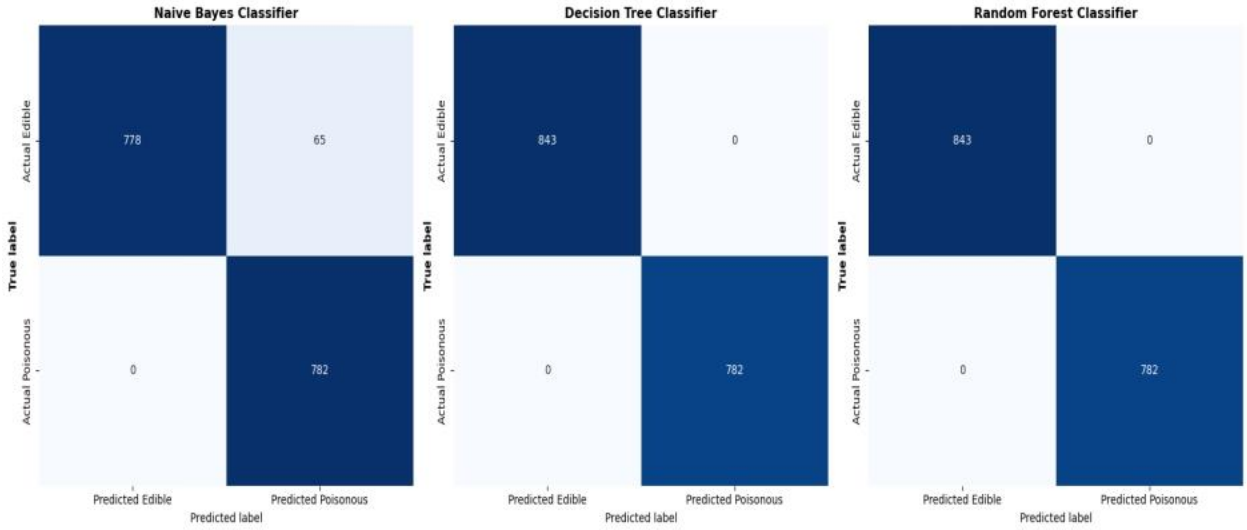
Data Loading and Preprocessing:

The data for the exercise was gotten from: https://archive.ics.uci.edu/ml/datasets/Mushroom

The aim is to process the data and subsequently evaluate model on a dataset, specifically the "Agaricus-Lepiota.data" mushroom dataset. The objective is to classify mushrooms as edible or poisonous using machine learning models: Naive Bayes, Decision Tree, and Random Forest classifiers. The mushroom dataset loaded into pandas Dataframe was without column header which was processed to include descriptive column names from an accompanying 'agaricus-lepiota.names' file. Also, attribute names were also extracted from the .names file. For better readability and analysis, these attributes including the target variables (class: Edible or poisonous) were assigned to the Dataframe.

Lastly, OneHotEncoder was employed to creates a new binary variable for each level of the categorical variable. Mushroom data then is split into training and testing sets.

Data Analysis and visualization



Comparing and Contrasting Results.

The accuracy of Naïve Bayes' Classifier on the dataset is 96% and has some misclassifications thus: 65 false positives and zero false negatives. The custom threshold (normally 0.5) was reduced to 0.4 to ensure that no potentially poisonous mushrooms are misclassified as edible, the model was retrained. Though, False positive increased from 65 to 68 and accuracy reduced to 95.8%, it was pertinent to err on the side of caution. The Decision Tree and Random Forest Classifiers achieved 100% accuracy on the dataset with no misclassification error.

Correlation matrix was checked to figure out the reason Naïve Baye's classifier could not achieve 100% accuracy. This can only be possible when there are no violations of its assumption of independent features. The Correlation matrix showed that there are instances of this violation which can be attributed to the classifier performing at 96% accuracy. Few of correlation table as shown below.

Pairs of features with high correlation:

```
stalk-color-below-ring_c     ring-number_n                  1.00000
O
odor_m                       stalk-color-below-ring_c       1.00000
O
                             ring-number_n                  1.00000
O
                             ring-type_n                    1.00000
O
stalk-color-above-ring_y     veil-color_y                   1.00000
O
stalk-color-above-ring_o     stalk-color-below-ring_o       1.00000
O
```

In conclusion, while each model has its advantages and disadvantages, Accuracy metrics should be the det ermining factor in choosing a particular classifier based on data characteristics. The ability of the Decisio n Tree and Random Forest models to perfectly classify the mushroom Datasets , make them the best choic e for this particular dataset. Nonetheless, Naive Bayes classifier, remains a viable option when cost and si mplicity into consideration

References: Machine Learning with Python Cookbook, 2nd Edition (oreilly.com) Chp 18