Decision Tree Classifier for Wines

Shreyas Patil
G01382371
Department of Computer Science
George Mason University
Fairfax, Virginia, United States of America
spatil28@gmu.edu

***Abstract*—Machine learning is a technology that enables computers to automatically learn from past data. It uses various algorithms for building mathematical models and making predictions using historical data or information. The decision Tree classifier is one of these algorithms. In this project, we implemented a decision tree classifier for a dataset of wines with 3 different types. The dataset has been explored and applied feature selection techniques have been applied to identify the most important features for classification. A Decision tree classifier was trained based on the selected features and its performance was evaluated using accuracy metrics.**

I. Introduction

In recent years, machine learning has emerged as a powerful tool for data analysis and classification. Decision tree classifiers are a popular choice for classification problems as they are easy to interpret and can handle both numerical and categorical data. In this project, we explore the use of decision tree classifiers for the classification of wines based on their features. The goal of this project is to develop a classifier that can accurately categorize different types of wines based on their attributes.

A Decision Tree is a classifier that is represented in the form of a tree structure where each node is either a decision node or a leaf node. The leaf node represents the value of the target attribute of examples. The decision node represents some test to be carried out on a single attribute value, with one branch and one subtree for each possible outcome of the test. They are commonly used for tasks in which attributes are represented by key-value pairs. The attributes can be continuous or categorical. [2]

II. Background

The dataset used in this project consists of 3 different types of wines designated as Type 1, Type 2, and Type 3 and contains several features for each wine, such as Alcohol content, Malic acid content, ash, Alcalinity, Magnesium, Phenols, Flavonoids, Nonflavonoid, Proanthocyanins, Color intensity, Hue, Diluted Wines, and Proline. The dataset used for training contains 28 data values. Before training our classifier, we performed data cleaning and exploratory data analysis to understand the distribution of features in the dataset. We also applied feature selection techniques to identify the most important features for classification.

A. Attribute selection measures

A heuristic for choosing the splitting criterion that splits data in the best way feasible is the attribute selection measure. Because it enables us to identify the breakpoints for tuples on a specific node, it is also referred to as splitting rules.

Gini Index: GINI is used in Classification and Regression Tree(CART). All the attributes are continuous-valued and it is assumed that several possible split values exist for each attribute. [2]

If a dataset T contains an example from n class, the Gini index is:

$$gini(T) = 1 - \sum_{j}^{n} = 1(P_j)^2 \tag{1}$$

It represents the relative frequency of class j in T. After splitting T into 2 subsets T1 and T2 with sizes N1 and N2 respectively, the Gini index is:

$$gini_(split)(T) = N_1/N * gini(T_1) + N_2/N * gini(T_2) \tag{2}$$

Information Gain: Assume there are 2 classes P and N. Let the set of records S contain p records of class P and n records of class N. The amount of information required to decide if a random record belongs in P or N is [2]

$$I(p,n) = -p/(p+n)log_2(p/(p+n)) - n/(p+n)log_2(n/(p+n)) \tag{3}$$

Entropy: It is the expected amount of information needed to assign a class to a randomly drawn object in S under the shortest optimal length code. It's also known as the randomness or impurity in the examples. [2]

$$E(A) = \sum_{i=1}^{v} p_i + n_i/(p+n) * I(p_i, n_i) \tag{4}$$

Gain: It measures the entropy achieved because of the split. Choose the split that results in a maximum gain. [2]

$$Gain(A) = I(p,n) - E(A) \tag{5}$$

III. Proposed Approach

The python modules used in this project were:

- numpy
- pandas
- matplotlib
- seaborn

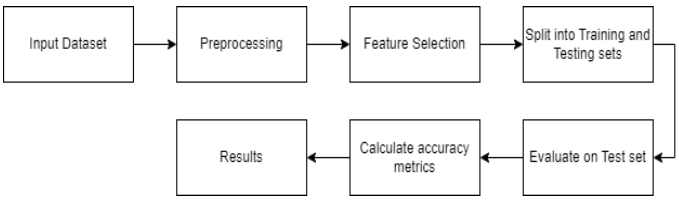

Fig. 1. Flowchart for Classifier.

- scikit-learn
- pydotplus
- six
- IPython

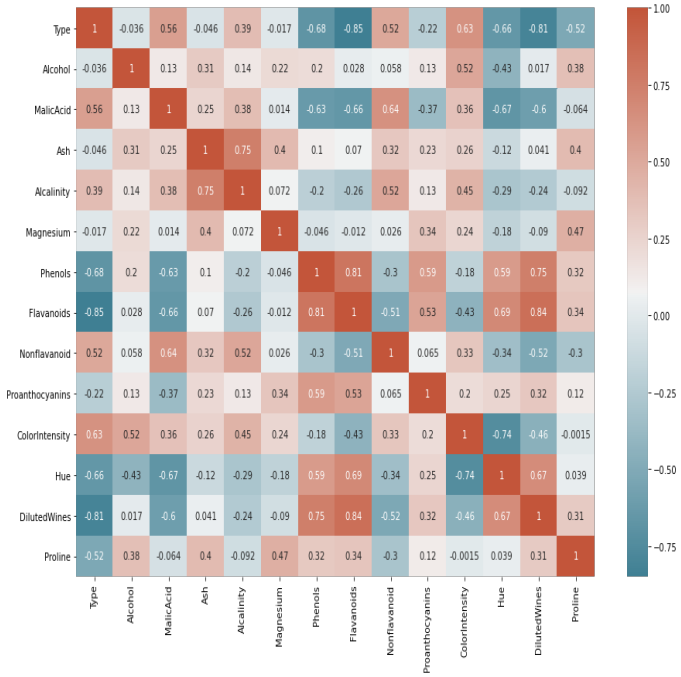

Fig. 2. Correlation Matrix for Wines.

We began by performing data cleaning on the wine dataset. After removing the duplicate values and the rows with empty values, exploratory data analysis was carried out on the wine dataset to understand its distribution and characteristics. We then applied feature selection techniques to identify the most important features for classification. Specifically, we calculated the correlation coefficients between each feature and the wine types and selected the top 5 features with the highest correlation coefficients. These features selected are Flavanoids, Diluted Wines, Phenols, Hue, and Color Intensity. The other features were not used for training or testing. We then split the dataset into training and testing sets, with 70% of the data used for training and 30% used for testing. Save the testing dataset in a file called testingdata.csv.

We then trained a decision tree classifier on the selected features using the training dataset. The decision tree classifier was implemented using scikit-learn, a popular machine-learning library in Python on the testing set. The maximum depth of the testing set was set to 3 and the Information Gain(entropy) criterion was used with a random state of 0.

IV. Experimental Results

After training the decision tree classifier, we evaluated its performance on the testing dataset using accuracy metrics. Our experimental results showed that the decision tree classifier achieved an accuracy of 1.0 on the training dataset and 0.888 on the testing dataset. While this accuracy score is lower than our desired threshold of 0.9, we believe it is still a good result given the small size of the dataset.

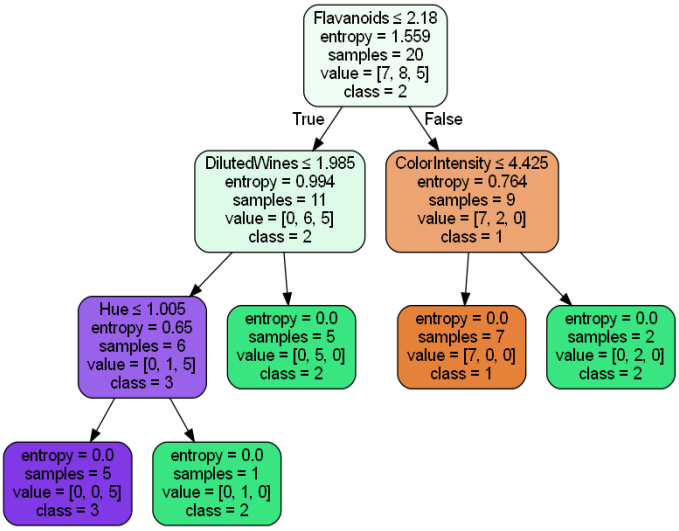

Fig. 3. Decision Tree For Wines.

V. Conclusion

In this project, we developed a decision tree classifier for the classification of wines based on their features. We performed exploratory data analysis and applied feature selection techniques to identify the most important features for classification. Our experimental results showed that the decision tree classifier was able to classify the wines with high accuracy. However, there is still room for improvement in the classification accuracy, and future work could explore using more sophisticated classifiers or feature selection techniques. Our testing accuracy is 88.8% which could be higher if we had a larger dataset. The current dataset only contains 28 values of wine. The testing accuracy can be improved to greater than 95% for a larger dataset containing more than 150 values so there us scope for improvement in the future.

References

[1] Russell. S, and Norvig. P, Artificial Intelligence: A Modern Approach, 4th US edition, Pearson: Boston, 2020.
[2] Dr Shubhangi Vaikole, Dr. Savarkar S.D., Machine Learning, Tech-Neo Publications, 2021.
[3] https://www.datacamp.com/tutorial/decision-tree-classification-python, 2023.