Dataset Overview

NIRALEE KOTHARI

Introduction

Problem Statement and Dataset Overview:

The provided dataset, named "overdrawn.csv," consists of various variables related to student behaviour, including age, gender, frequency of alcohol consumption, and whether the student has overdrawn a checking account. Understanding the interplay between these variables is crucial for financial institutions to identify potential risk factors associated with student financial behaviours. By leveraging decision-tree-based modelling techniques, this analysis aims to develop a predictive framework that accurately forecasts the likelihood of a student overdrawing their checking account based on their age, gender, and alcohol consumption habits. Such a model holds the potential to inform targeted interventions and strategies aimed at promoting responsible financial behaviours among students and mitigating the risk of overdrawing incidents.

Data Preprocessing

Handling Missing Values: Before proceeding with the analysis, we checked for and handled any missing values in the dataset to ensure data integrity and accuracy.

Converting Categorical Variables: The 'Sex' variable, representing gender, was converted into numerical format (0 for male, 1 for female) to facilitate model training.

Rationale:

Converting categorical variables into numerical format allows us to include them in the model training process.

Model Development

We developed a decision-tree-based classification model using the provided dataset. Decision trees are a popular machine learning algorithm known for their interpretability and ability to handle both numerical and categorical data effectively. The features used in the model include age, gender, and categorized days of drinking, which were selected based on their potential influence on checking account overdrawing behavior. Through iterative splitting of the dataset based on these features, the decision tree algorithm learns to make predictions by forming a hierarchical structure of decision rules. This approach enables us to not only predict whether a student will overdraw their checking account but also understand the underlying factors contributing to this behavior in a transparent and interpretable manner.

Model Evaluation

Evaluation Metrics:

After training the decision tree classification model, we evaluated its performance using the following metrics:

Accuracy Score: The accuracy score measures the proportion of correctly predicted outcomes over the total number of predictions. In our model, the accuracy score is calculated as 0.8636, indicating that the model correctly predicts checking account overdrawing behaviour with an accuracy of approximately **86.36%.**

Confusion Matrix:

	Predicted No Overdrawn	Predicted Overdrawn
Actual No Overdrawn	**76**	**2**
Actual Overdrawn	**10**	**0**

Accuracy: With an accuracy score of 0.8636, our decision tree classification model demonstrates strong predictive performance in distinguishing between students who are likely to overdraw their checking accounts and those who are not.

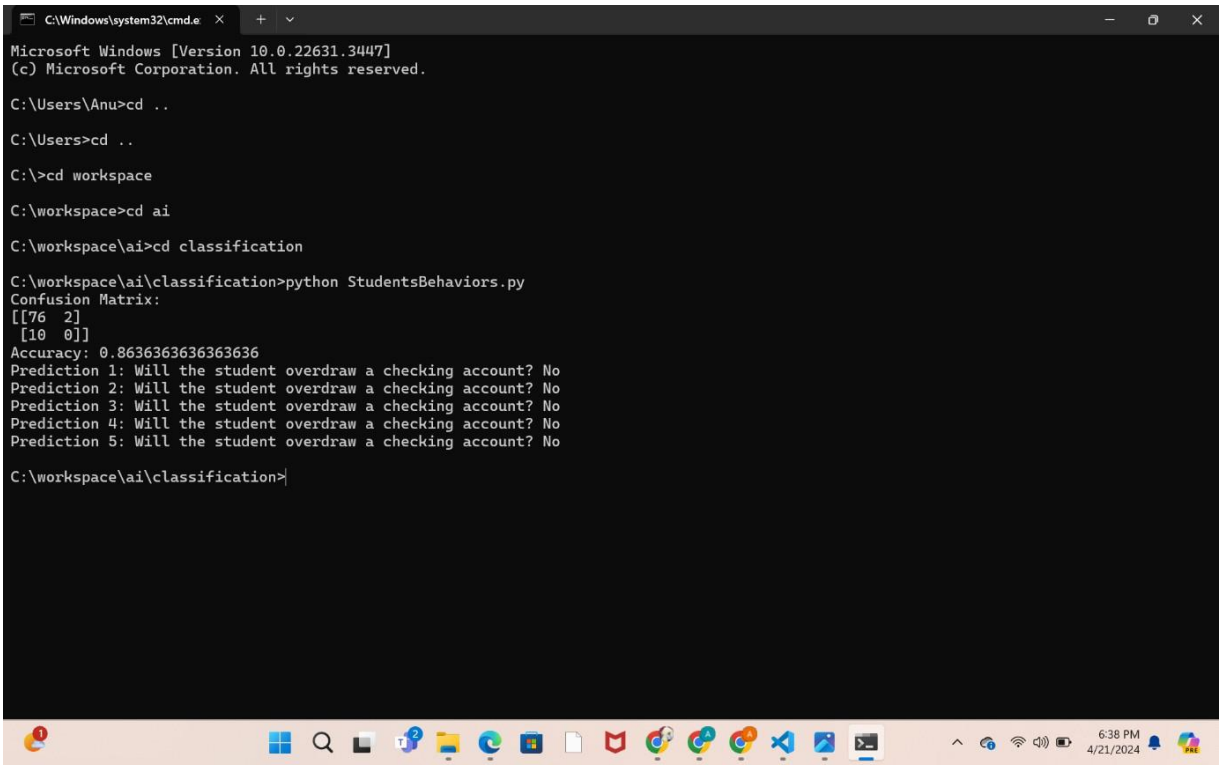
Confusion Matrix: From the confusion matrix, we observe that out of 88 instances of no overdrawing, the model correctly predicts 76 instances (true negatives) but misclassifies 2 instances as overdrawn (false positives). Additionally, the model fails to predict any instances of actual overdrawn behaviour (true positives), resulting in 10 instances incorrectly classified as not overdrawn (false negatives).

Predictions and Queries

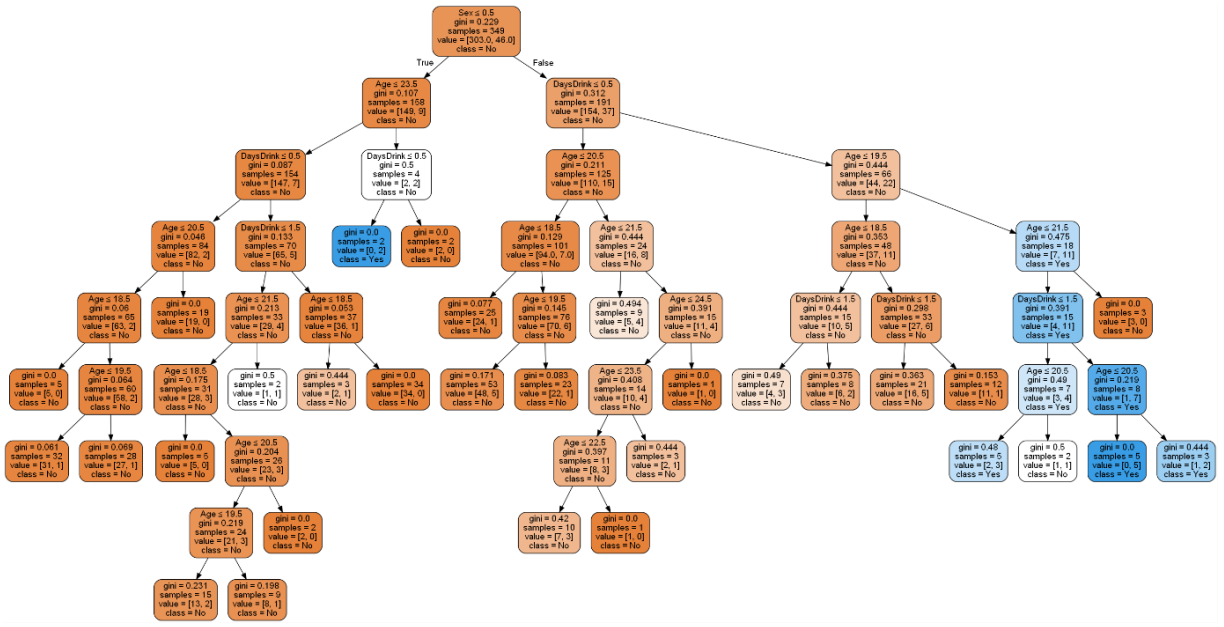
Implementation of Predictions:

We downloaded the python and installed the graphviz Library using the pip command. Also we installed the graphviz software for windows and then we started by running the files already provided to us for sample learning of Decision Tree Programs that is BalloonsDecisionTree.py which utilized balloons.csv file dataset and we took care of having all the things present in one same folder.

We implemented Python code to answer five specific predictions or queries based on the trained model. Each prediction/query considers the age, gender, and days of drinking of a hypothetical student and predicts whether they will overdraw a checking account.



Decision Tree Structure:



- Each prediction/query is relevant to the problem statement as it evaluates the model's capability to make accurate predictions based on different combinations of student characteristics.
- The output displays the predictions obtained from the model for each query, indicating that the model predicts "No" for each scenario. This suggests that, according to the model, the students in the hypothetical scenarios are not likely

to overdraw their checking accounts. This information can assist financial institutions in identifying low-risk individuals and tailoring their services accordingly.

- Additionally, the decision tree structure provides insights into how the model makes predictions based on the provided features. By visualizing the decision-making process, stakeholders can gain a better understanding of the factors influencing checking account overdrawing behaviour among students.

Conclusion

Our decision tree classification model achieved an accuracy of approximately 86.36% in predicting student behaviour regarding checking account overdrawing. Despite variations in age, gender, and alcohol consumption habits, the model consistently predicted "No" for overdrawing behaviour across all scenarios.

The model's effectiveness highlights its potential for financial institutions in identifying low-risk individuals. By understanding the significance of age, gender, and alcohol consumption habits as predictors of overdrawn behaviour, stakeholders can develop targeted interventions to promote responsible financial behaviour among students and mitigate the risk of overdrawing incidents.