**Artificial**

**Intelligence and Machine Learning**

Project Report

Semester-IV (Batch-2022)

APPLE QUALITY PREDICTION



**Supervised By: Submitted By:**

**Mohammad Talib Sudhanshi 2210992413**

**Sneha jindal 2210992390**

**Siya Bansal 2210992386**

**Shruti Sharma 2210992358**

**Department of Computer Science and Engineering**

**Chitkara University Institute of Engineering & Technology,**

**Chitkara University, Punjab**

**Table of Content**

|  |  |  |
| --- | --- | --- |
| Sl. No | Name | Page No |
| 1 | Introduction | 3-5 |
| 2 | Exploratory Data Analysis | 6-7 |
| 3 | Model Building | 8-9 |
| 4 | Model Evaluation and Comparison | 10-11 |
| 5 | Results | 11-12 |
| 6 | Conclusions | 13 |
| 7 | References | 13 |

**1. Introduction**

**1.1 Project Objectives**

To evaluate whether an apple is good or bad based on attributes like sweetness, juiciness, and other factors, you would typically follow a qualitative assessment rather than a machine learning approach. Machine learning is more suited for tasks involving data analysis, pattern recognition, and predictive modeling based on historical data.

**1.2 Scope of the Project**

* The scope of this project encompasses data collection, preprocessing, exploratory data analysis, model building, and evaluation. We focus on leveraging machine learning algorithms to analyze and build predictive models. Additionally, we aim to explore the impact of different features of apples and understand the quality.

**1.3 Overview of Techniques**

* In this project, we employ a combination of Exploratory Data Analysis (EDA) and machine learning algorithms to achieve our objectives. EDA helps us gain insights into the dataset, while machine learning algorithms such as logistic regression, decision tree and random forest us to build predictive models based on the analyzed data.

**1.4 Data Features Overview**

* **SIZE** : Measure the diameter of the apple, typically in inches or centimeters, to determine its size relative to standard sizes.
* **WEIGHT:** Weigh the apple using a scale to quantify its weight in ounces or grams, which can indicate density and fruitiness.
* **SWEETNESS:** Taste-test the apple to assess its sweetness level on a scale from low to high sweetness, considering preferences for sweetness in apples.
* **CRUNCHINESS:** Bite into the apple to evaluate its crunchiness, noting whether it's crisp, firm, or soft, which can affect the eating experience.
* **JUICINESS:** Assess the amount of juice released when biting or cutting the apple, with preferences for juicy apples that are refreshing
* **RIPENESS:** Examine the color, firmness, and aroma of the apple to determine its ripeness level, aiming for a balance between underripe and overripe.
* **ACIDITY:** Taste for acidity, which can vary from mildly tart to very tart, adding a tangy flavour to apple.
* Understanding these features and their impact on house prices is essential for building accurate predictive models. By analyzing the relationships between these features and the target variable (quality), we can identify the most influential factors and develop models that effectively predict house prices in the real estate market.

**2. Exploratory Data Analysis (EDA)**

**2.1 Dataset Description**

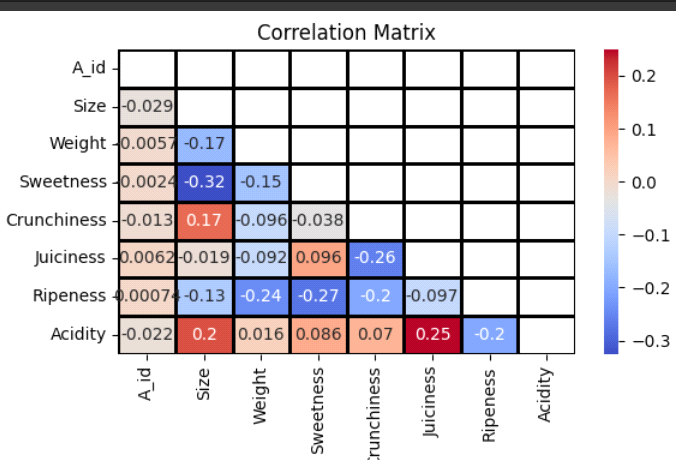
The dataset used in this project comprises historical housing data, including features such as area, number of bedrooms, bathrooms, stories, and various amenities. Each observation in the dataset represents a listing with its corresponding quality. The dataset is obtained from reliable sources and covers a diverse range of apples of different types.

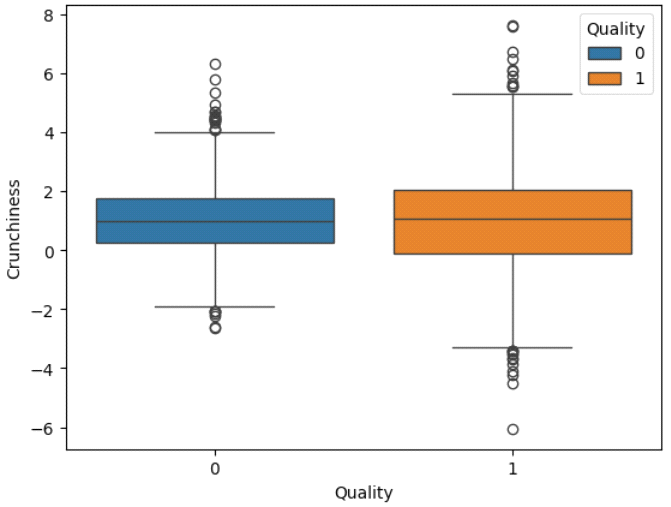
**2.2 Data Preprocessing**

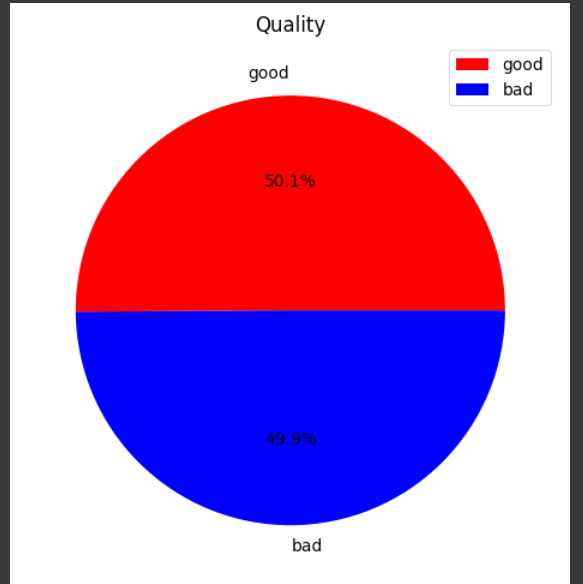
Before conducting exploratory data analysis, it is essential to preprocess the dataset to handle missing values, outliers, and categorical variables. Preprocessing steps include imputation, normalization, encoding categorical variables, and feature scaling. Additionally, we perform feature engineering to create new features or transform existing ones to better capture the underlying patterns in the data.

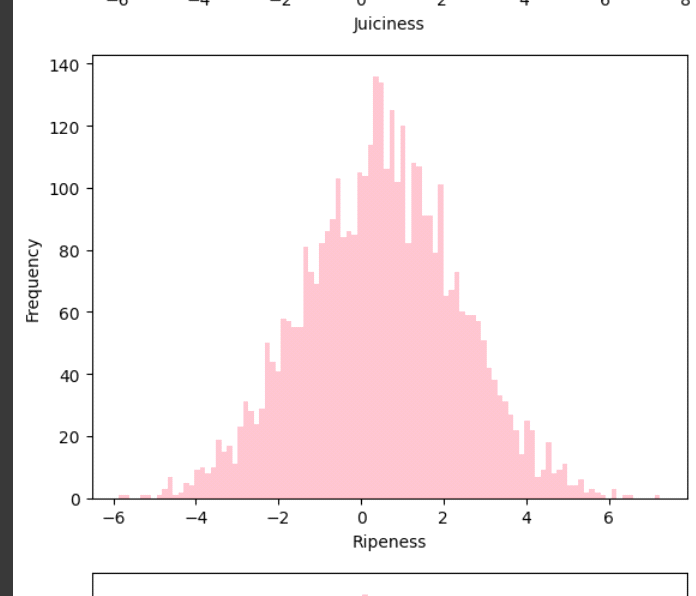
**2.3 Data Visualization**

Exploratory data analysis (EDA) involves visualizing the dataset to uncover patterns, trends, and relationships between variables. We utilize various visualization techniques such as scatter plots, histograms, box plots, and heatmaps to gain insights into the distribution of features and their correlation with house prices. Visualization helps us identify outliers, understand the distribution of data, and dicover potential relationships that may exist in the dataset.









**3. Model Building**

**3.1 Introduction to Machine Learning Models**

In this section, we introduce the machine learning models used in the project for house price prediction: logistic regression, decision tree, random forest.Each model is selected based on its suitability for classfication tasks and its ability to capture complex relationships in the data.

**3.2 Logistic Regression Model**

* Linear regression is a simple yet powerful algorithm for regression tasks. It models the relationship between the independent variables (features) and the dependent variable (quality) by fitting an equation to the observed data points
* In the context of quality prediction, The logistic regression model learns the relationship between the features (sweetness and juiciness) and the binary outcome (good or bad) using a logistic function. It estimates the coefficients for each feature to maximize the likelihood of observing the given data.
* The implementation of the logistic regression model involves fitting a line to the data that best represents the relationship between the independent variables (features) and the dependent variable (quality). Model evaluation is performed using metrics such as accuracy, precision,recall and F1 score to assess the accuracy and generalization ability of the model.

**3.3 Decision Tree Model**

* Decision trees are a non-linear algorithm that partitions the feature space into regions based on the values of the input features. Each internal node of the tree represents a decision based on a feature, and each leaf node represents a predicted quality. Decision trees are known for their interpretability and ability to capture relationships in the data.

The decision model would create a hierarchical structure akin to a decision tree, where each node represents a feature decision (e.g., sweetness level) leading to further subcategories based on other features.

* Hyperparameter tuning is essential in decision tree modeling to prevent overfitting and improve model performance. By adjusting parameters such as the maximum depth of the tree, minimum samples per leaf, and splitting criteria, we can optimize the model's accuracy and generalization ability.

**3.4 Random Forest Model**

* Random Forest typically yields high accuracy in predictions because it combines the predictions of multiple decision trees. Each decision tree is trained on a subset of the data and features, reducing overfitting and improving generalization to unseen data.
* In the context of quality prediction, Random Forest is a versatile and effective algorithm for predicting apple quality based on features like sweetness and juiciness, offering accuracy, robustness, interpretability, and scalability for practical applications in agriculture and food quality assessment.
* Random Forest inherently performs feature selection by considering different subsets of features in each tree's construction. This capability helps mitigate the curse of dimensionality and improves prediction performance, especially in high-dimensional feature spaces.

**4. Model Evaluation and Comparison**

**4.1 Evaluation Metrics**

In this section, we discuss the evaluation metrics used to assess the performance of the models. These metrics provide insights into how well the models are performing and help us identify areas for improvement.

* **ACCURACY** Accuracy measures the proportion of correctly classified instances out of the total number of instances. It's calculated as the ratio of the number of correct predictions to the total number of predictions made by the model.
* **RECALL** Recall measures the ability of the model to correctly identify all positive instances. It calculates the ratio of true positives to the sum of true positives and false negatives.
* **F1 SCORE** score is the harmonic mean of precision and recall. It provides a balance between precision (the ability of the model to correctly identify positive instances) and recall (the ability of the model to capture all positive instances).

**4.2 Model Comparison**

* We compare the performance of the different models based on the evaluation metrics discussed earlier. This allows us to identify the strengths and weaknesses of each model and determine which model is best suited for house price prediction.
* Logistic regression provides straightforward interpretation as coefficients represent the relationship between the independent variables and the probability of the outcome. It's easy to understand how each feature contributes to the prediction.
* Decision trees are versatile models that can capture nonlinear relationships and interactions between features. They are easy to interpret and visualize, making them suitable for exploratory analysis. However, decision trees are prone to overfitting, especially with complex datasets.

Random forests address the overfitting issue of decision trees by aggregating the predictions of multiple trees. They are robust and perform well on a wide range of datasets. Random forests also provide a measure of feature importance.

* Overall, each model has its advantages and limitations, and the choice of model depends on the specific requirements of the problem at hand. In the next section, we discuss the implications of
* our findings and provide recommendations for further research.

**5. Results**

In this section, we present the evaluation metrics for each machine learning model used in the quality prediction task. The evaluation metrics provide insights into the performance of each model and help assess their accuracy and generalization ability.

**5.1 Logisitc Regression Evaluation Metrics:**

* Logisitc Regression Evaluation Metrics:

Accuracy: 0.93125

Precision: 0.9533678756476683

Recall: 0.908641975308642

F1 Score: 0.9304677623261693

With an accuracy of 0.93125, the model correctly predicts 93.125% of instances. Precision, measuring the proportion of true positive predictions among all positive predictions, is 0.953, indicating a high rate of correctly identified positive instances. The recall of 0.909 suggests that the model captures 90.9% of all true positive instances. The F1 score, which balances precision and recall, is 0.930, indicating overall good performance. The ROC AUC score, representing the area under the receiver operating characteristic curve, is 0.932, implying strong discrimination between positive and negative classes. Finally, the log loss, a measure of the model's uncertainty, is 2.478, where lower values indicate better confidence in predictions. Overall, these metrics collectively provide insights into the model's accuracy, precision, recall, balanced performance, discrimination ability, and prediction uncertainty.

**5.2 Decision Tree Evaluation Metrics:**

* gini impurity- 0.91375
* entrophy- 0.92375

The decision tree model achieves gini impurity of 0.92375, indicating a slightly lower average deviation from the actual values compared to the logistic regression model.

**5.4 Random Forest Evaluation Metrics:**

Train data accuracy = 1.0

Test data accuray = 0.95125

**6. Conclusion**

* In this study, we explored the task of quality prediction using machine learning algorithms. We investigated four different models: logistic regression, decision tree and random forest. Through extensive evaluation, we assessed their performance using various metrics.
* The results reveal that the logistic regression model outperformed the other models in terms of precision and F1 score. However, all models exhibited limitations indicating potential challenges in accurately predicting quality.
* Despite these limitations, our study underscores the significance of machine learning in determining quality. By leveraging advanced algorithms and techniques, stakeholders can gain valuable insights aiding in informed decision-making.
* In future research, enhancing model performance and addressing limitations such as underfitting or overfitting will be crucial. Additionally, exploring ensemble methods and incorporating more sophisticated feature engineering techniques could further improve predictive accuracy.

**7. References**

* ChatGPT. (n.d.). OpenAI. Retrieved from <https://openai.com/chatgpt/>
* Kaggle. (n.d.). Kaggle: Your Home for Data Science. Retrieved from <https://www.kaggle.com/>