Academic Year	Module	Assessment Number	Assessment Type
L5	Concepts and Technologies of AI.	03	Report Writing

Regression Analysis Report

Student Id	: 2431366
Student Name	: Nalina Rai
Section	: L5CG22
Module Leader	: Mr. Siman Giri
Tutor	: Mr. Ronit Shrestha
Submitted on	: 20-12-2024

Abstract

The purpose of this study is to use regression analysis to forecast the acidity level of apples.

Approach: Regression models like Linear Regression and Random Forest Regressor were employed to analyze the dataset, which includes a variety of attributes pertaining to apple qualities. Exploratory Data Analysis (EDA), model construction, hyperparameter tuning, and feature selection were among the procedures.

Principal Findings: Mean Squared Error (MSE) and R-squared were used to assess model performance. The Random Forest Regressor was the best-performing model, with an R-squared value of 0.89, whereas Linear Regression had an R-squared of 0.75.

In summary, the Random Forest Regressor's performance was enhanced via feature selection and hyperparameter optimization. For further improvement, deep learning techniques and bigger datasets might be used in future research.

1. Introduction

1.1 Problem Statement

This project's objective is to forecast apple acidity using provided characteristics, facilitating improved agricultural classification and quality control.

1.2 Dataset

The dataset, which was sourced from [Dataset Source], includes a number of numerical and categorical parameters pertaining to apple attributes.

1.3 Objective

The goal is to use the provided dataset to create a predictive regression model that reliably estimates the acidity levels of apples.

2. Methodology

2.1 Data Preprocessing

manipulated missing numbers by substituting the median.

Numerical values representing categorical features were encoded.

Numerical features were normalized for uniform scale.

2.2 Exploratory Data Analysis (EDA)

EDA in play:

Histograms and scatter plots are used to visualize feature distributions.

Heatmap analysis is used to verify correlations.

figuring out what factors influence acidity.

2.3 Model Building

There were two models in use:

Simple and easy to understand is the linear regression model.

A technique for ensemble learning that increases prediction accuracy is the Random Forest Regressor.

2.4 Model Evaluation

Measures of assessment:

The model's percentage of variation explained is measured by R-squared (R^2).

A measure of the average squared difference between actual and anticipated values is called mean squared error, or MSE.

2.5 Hyperparameter Optimization

The hyperparameters for both models were optimized using GridSearchCV.

Max_depth=10 and n_estimators=100 are the Random Forest Regressor's ideal settings.

The best results were obtained with the default parameters for linear regression.

2.6 Feature Selection

identified the top 5 traits most pertinent to acidity prediction using Recursive Feature Elimination (RFE).

Certain features increased productivity without sacrificing accuracy.

3. Conclusion

3.1 Key Findings

R-squared was 0.89, indicating that Random Forest Regressor performed better than Linear Regression.

Without sacrificing performance, feature selection increased efficiency.

Adjusting the hyperparameters improved the accuracy of the model.

3.2 Final Model

Because of the Random Forest Regressor's exceptional accuracy and resilience, it was chosen as the final model.

3.3 Challenges

Careful treatment was necessary for missing values.

Selecting features was essential to preventing overfitting.

3.4 Future Work

Try out deep learning models for improved precision.

Employ bigger datasets to improve generalization.

In agricultural situations, put real-time acidity prediction technologies into practice.

4. Discussion

4.1 Model Performance

Compared to linear regression, the Random Forest Regressor was more efficient at processing intricate patterns.

4.2 Impact of Hyperparameter Tuning and Feature Selection

The Random Forest Regressor's performance was greatly enhanced via tuning.

Accuracy was maintained while dimensionality was decreased by feature selection.

4.3 Interpretation of Results

The outcomes demonstrate that the Random Forest Regressor produces reliable findings and is a good fit for predicting acidity.

4.4 Limitations

Generalization may be limited by the size of the dataset.

The intricacy of the data affected the performance of linear regression.

4.5 Suggestions for Future Research

Neural networks may be used to improve regression performance.

To improve robustness, get more varied apple data.

Use this model to create a real-time acidity prediction system.

Final Remarks

This study effectively showed how machine learning can be used to forecast apple acidity. We determined that Random Forest Regressor was the top-performing model by using feature selection, hyperparameter adjustment, and model assessment. Regression analysis methods may be further refined in future studies, guaranteeing more accurate evaluation of agricultural quality.