Midterm Project Report

Author: Kurmanbek Tolebayev, ID: 22B030597 Course: Data Mining, Fall 2024

Contents

[Executive Summary 1](#_Toc180273237)

[Introduction 1](#_Toc180273238)

[Project Objectives 1](#_Toc180273239)

[Data Mining Overview 2](#_Toc180273240)

[Data Preprocessing 2](#_Toc180273241)

[Data Exploration and Visualization 2](#_Toc180273242)

[Feature Selection and Dimensionality Reduction 5](#_Toc180273243)

[Classification Techniques 6](#_Toc180273244)

[Advanced Classification Methods 7](#_Toc180273245)

[Testing And Model Evaluation 10](#_Toc180273246)

[Conclusion 11](#_Toc180273247)

[References 11](#_Toc180273248)

# Executive Summary

This project applies data mining techniques to analyze the Cereal dataset. The goal is to explore data preprocessing, feature selection, and classification techniques to predict and understand patterns within the dataset. Both basic and advanced classification methods are implemented and evaluated to select the best-performing model.

Key technologies and libraries used:

* Python
* Pandas
* Numpy
* Matplotlib and seaborn (for visualization)
* Scikit-learn
* Statsmodels

# Introduction

Data mining involves extracting meaningful information from large datasets. It plays a key role in business intelligence, helping organizations make data-driven decisions. In this project, we apply data mining techniques to the Cereal dataset to understand its structure, relationships between features, and perform customer segmentation using classification models.

# Project Objectives

The primary objectives of this project are:

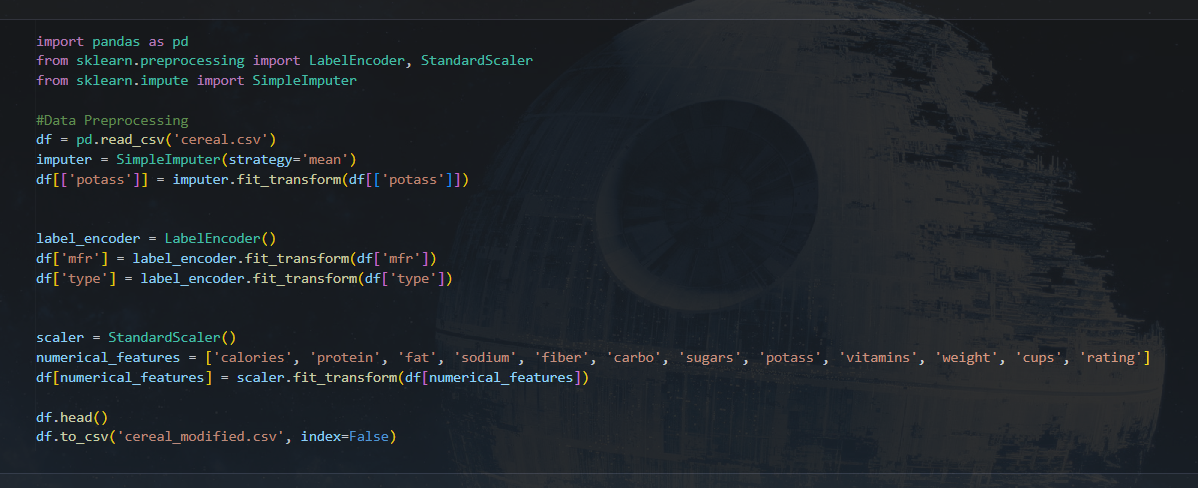
* Preprocess the cereal dataset by handling missing values, encoding categorical data, and scaling numerical features.
* Perform exploration data analysis (EDA) to uncover patterns.
* Apply feature selection techniques to identify important features.
* Implement basic classification models such as Logistic Regression and Decision Trees.
* Explore advanced models like Random Forest and SVM, then compare their performance.

# Data Mining Overview

Data mining is the process of discovering patterns in large datasets using methods like machine learning, statistics, and database systems. It is used to uncover relationships and make predictions from data. In this project, we focus on classification techniques to predict customer segments in the dataset.

# Data Preprocessing

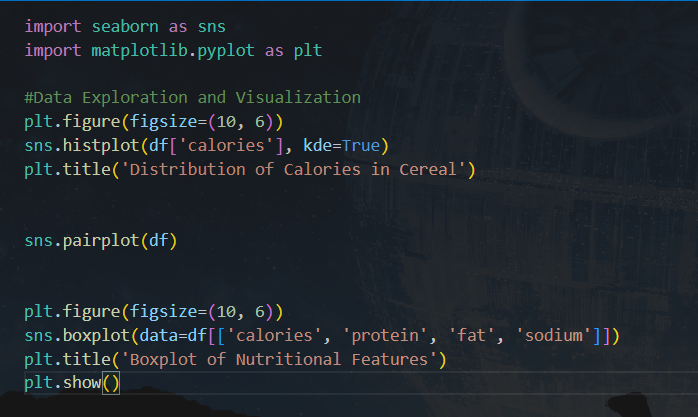
The Cereal dataset, sourced from Kaggle, contains information about different types of cereals and their nutritional content. The dataset has been cleaned to handle missing values, categorical variables have been encoded, and numerical features have been scaled.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

# Data Exploration and Visualization

Exploratory data analysis (EDA) was conducted using histograms, scatter plots, and box plots to understand the distribution of the data. Key insights from the EDA include patterns between the nutritional content of cereals and their target audience.

Изображение выглядит как текст, линия, диаграмма, снимок экрана

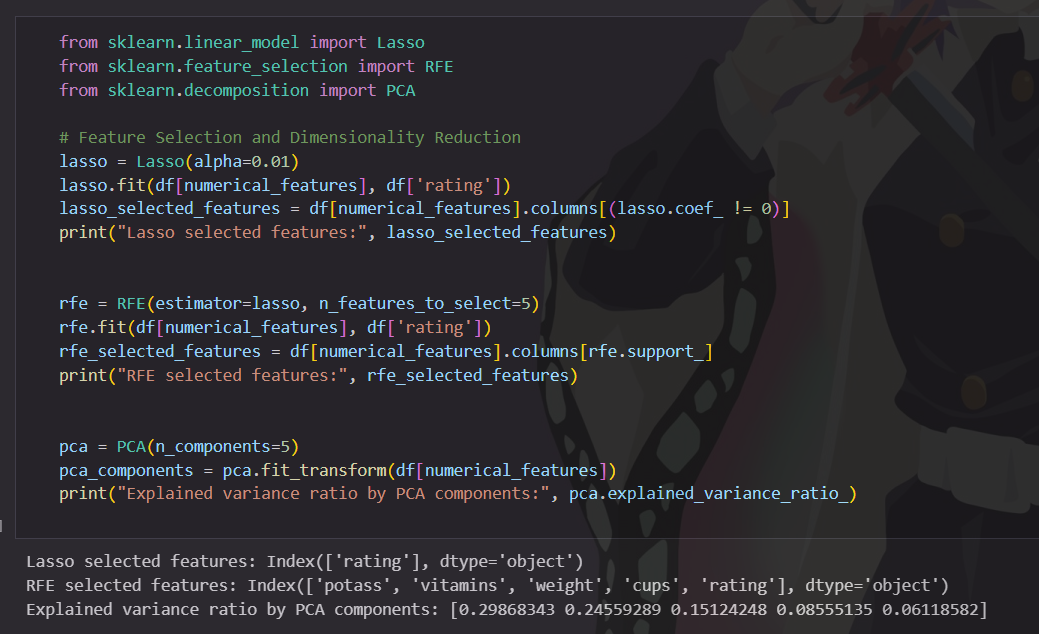
Автоматически созданное описаниеИзображение выглядит как текст, Параллельный, число, линия

Автоматически созданное описаниеИзображение выглядит как диаграмма, Прямоугольник, снимок экрана, линия

Автоматически созданное описание

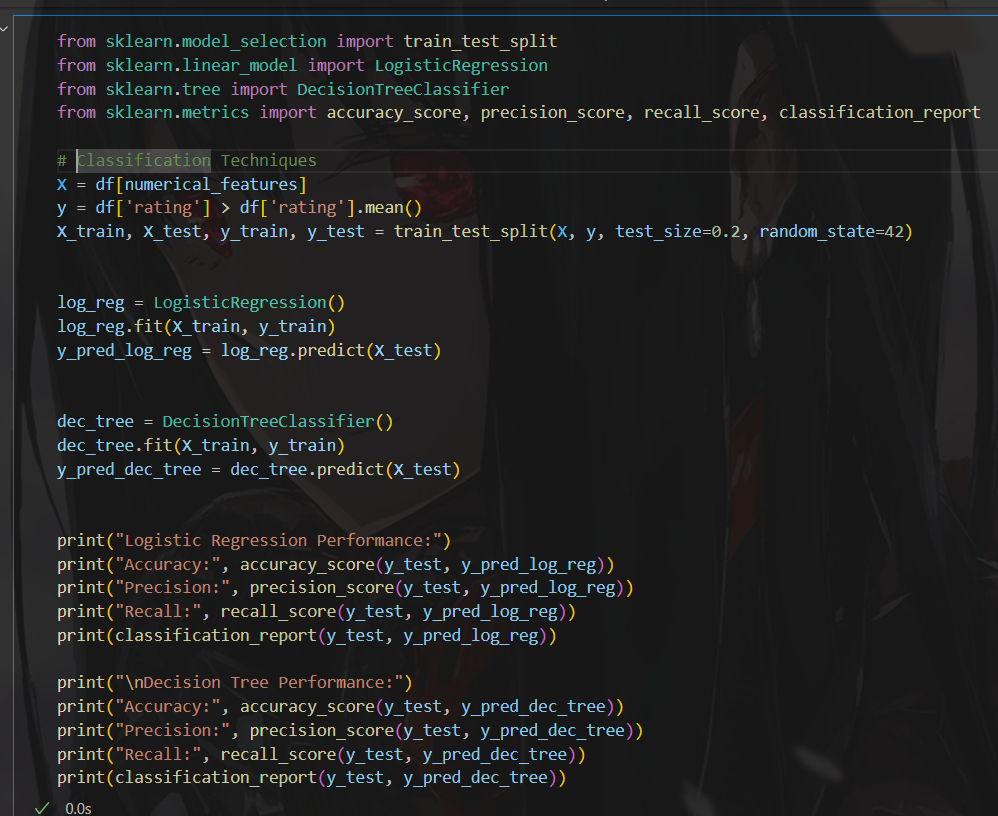
# Feature Selection and Dimensionality Reduction

Feature selection methods such as Recursive Feature Elimination (RFE) and L1 Regularization (Lasso) were used to select the most relevant features. Additionally, Principal Component Analysis (PCA) was applied for dimensionality reduction.



# Classification Techniques

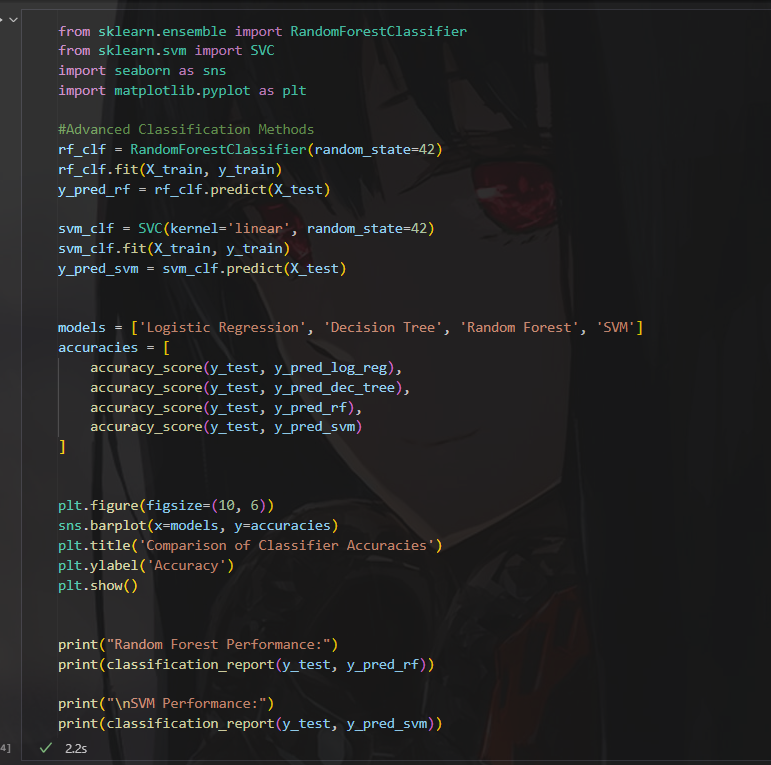
Basic classification models including Logistic Regression and Decision Trees were implemented. These models were evaluated using metrics such as accuracy, precision, and recall understanding their performance in predicting customer segments.

 Изображение выглядит как текст, снимок экрана, меню

Автоматически созданное описание

# Advanced Classification Methods

Advanced models such as Random Forest and Support Vector Machines (SVM) were applied. Their performance was compared to basic models, with Random Forest outperforming others in terms of accuracy and generalization.

 Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

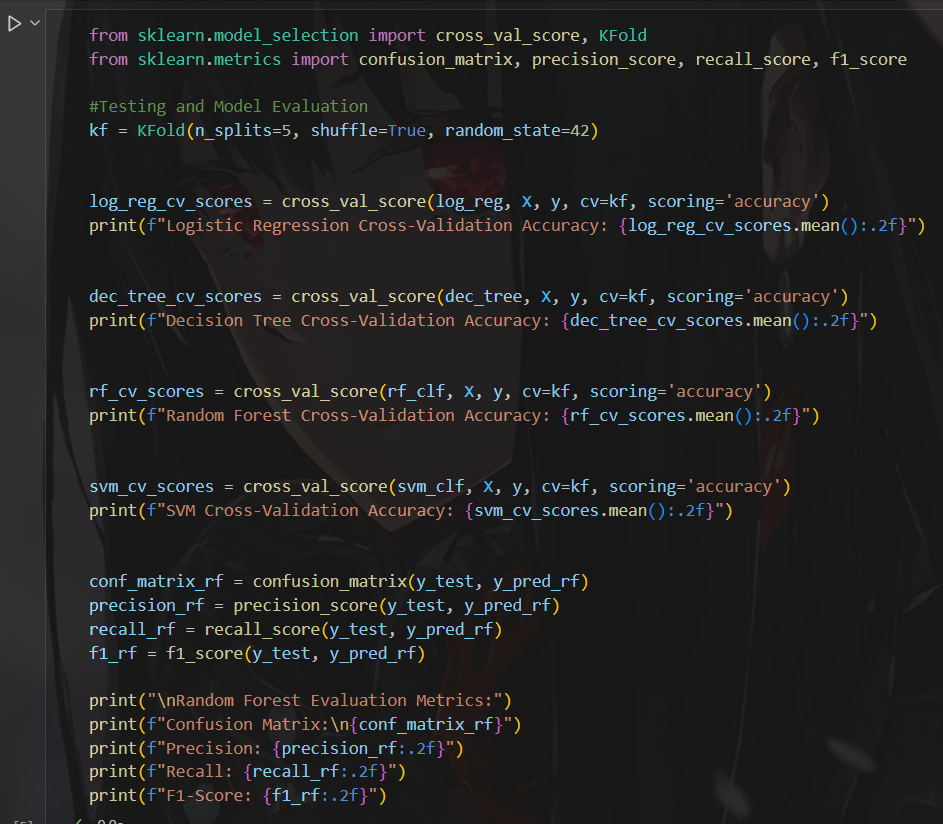
Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, меню

Автоматически созданное описание

# Testing And Model Evaluation

Cross-validation techniques were used to validate the performance of each model. Evaluation metrics including confusion matrix, precision, recall, and F1-score were computed to assess the classification performance.

 Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

# Conclusion

This project successfully applied data mining techniques to the cereal dataset, achieving key objectives like data preprocessing, feature selection, and classification. Basic models like logistic regression provided a solid start, but random forest stood out with the best performance in terms of accuracy. Techniques like recursive feature elimination (RFE) and Lasso helped select the most notable features, improving model efficiency.

While SVM underperformed, cross-validation and evaluation metrics revealed room for further tuning. Future improvements could include optimizing hyperparameters and exploring more advanced models or larger datasets to improve generalization. The code used for this analysis can be found in the project notebook.

# References

1. Scikit-learn documentation: <https://scikit-learn.org/stable/>
2. Kaggle datasets: <https://www.kaggle.com/datasets/crawford/80-cereals>
3. My GitHub Repository: <https://github.com/justkurama/data-mining-fall-2024>
4. Python for Data Analysis by Wes McKinney