Data Mining Assignment Report

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

GitHub Repository: https://github.com/justkurama/data-mining-fall-2024

# Introduction

This report documents the implementation of various feature selection, classification, and regression techniques using Python and scikit-learn. The assignment aimed to explore different methods for selecting important features and building machine learning models to solve classification and regression problems.

# Deliverables

The code for all the exercises described below is available in the provided GitHub repository.

# Feature Selection Techniques

## Feature Selection with SelectKBest

Dataset: Iris Dataset

Objective: Select the top 2 features using SelectKBest and the chi2 score function. Method:

* The dataset was split into features and target.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

* SelectKBest was applied to select the most relevant features.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

* The names of the top 2 features were printed.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

## Feature Importance with Random Forest

Dataset: Wine Dataset

Objective: Determine feature importance using a Random Forest classifier. Method:

* The dataset was split into training and testing sets.
* Изображение выглядит как текст, снимок экрана, 3D-моделирование, Программное обеспечение для видеоигр

  Автоматически созданное описание
* A Random Forest classifier was trained, and the feature importances were extracted.

Изображение выглядит как текст, снимок экрана, Компьютерная игра, Программное обеспечение для видеоигр

Автоматически созданное описание

* The feature importances were visualized using a bar plot.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

## Recursive Feature Elimination (RFE)

Dataset: Breast Cancer Dataset

Objective: Use RFE with a Support Vector Machine (SVM) classifier to select features. Method:

* The dataset was split into training and testing sets.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

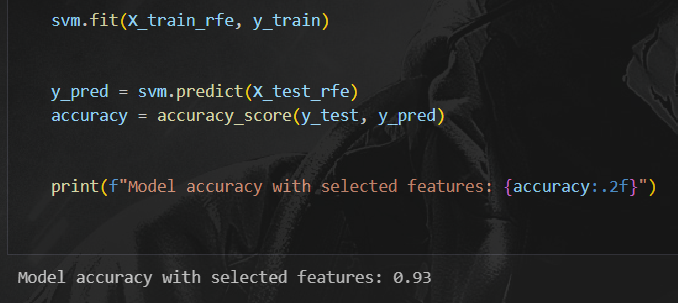
* RFE was used to select the most important features.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

* The SVM model was trained with the selected features, and its performance was evaluated.





## L1 Regularization (Lasso) for Feature Selection

Dataset: Diabetes Dataset

Objective: Apply Lasso regression for feature selection. Method:

* The dataset was split into training and testing sets.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

* Lasso regression was applied to select relevant features.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

* A model was trained using the selected features, and its performance was evaluated.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

# Classification Techniques

## Logistic Regression

Dataset: Iris Dataset

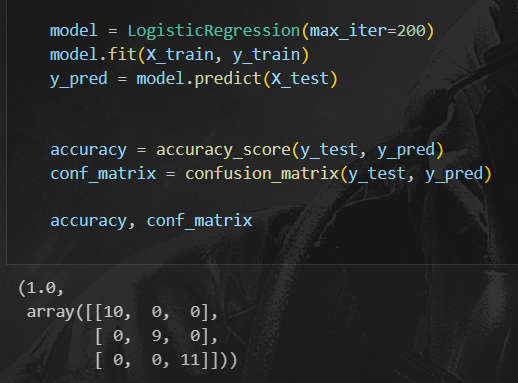
Objective: Build a logistic regression model for classification. Method:

* The dataset was split into training and testing sets.



* A logistic regression model was trained, and its performance was evaluated using accuracy and a confusion

matrix.

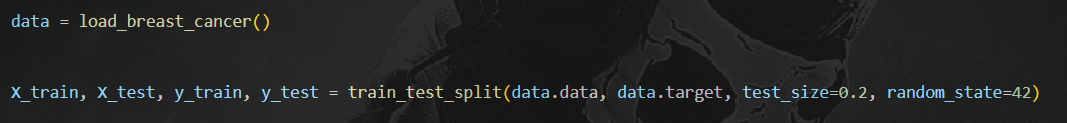


## Support Vector Machine (SVM)

Dataset: Breast Cancer Dataset

Objective: Build an SVM classifier for data classification. Method:

* The dataset was split into training and testing sets.



* An SVM model was trained and evaluated using accuracy and a confusion matrix.

Изображение выглядит как текст, снимок экрана, Шрифт

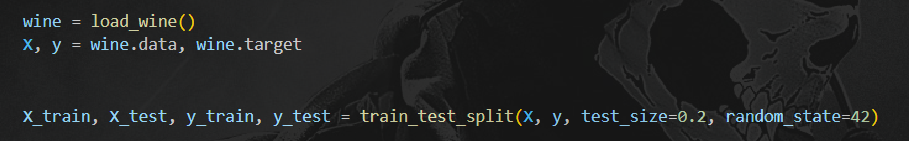
Автоматически созданное описание

## Decision Tree Classifier

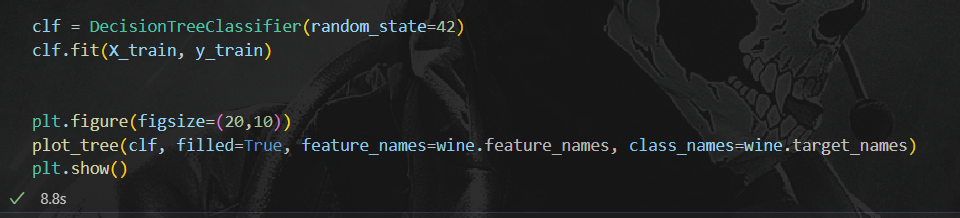
Dataset: Wine Dataset

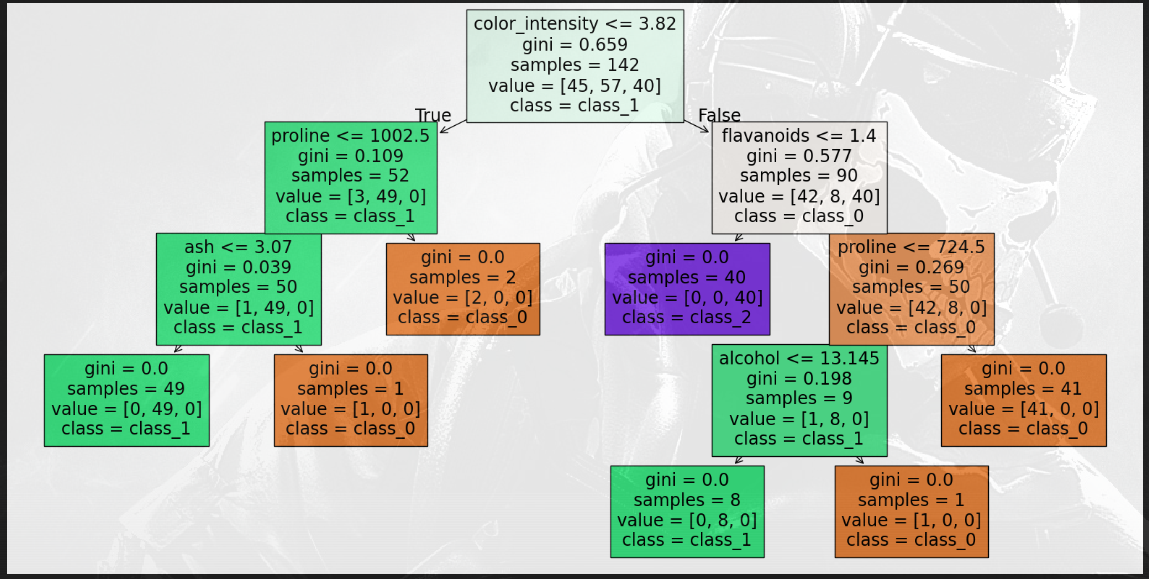
Objective: Build and visualize a decision tree classifier. Method:

* The dataset was split into training and testing sets.



* A decision tree classifier was trained, and the resulting tree was visualized.





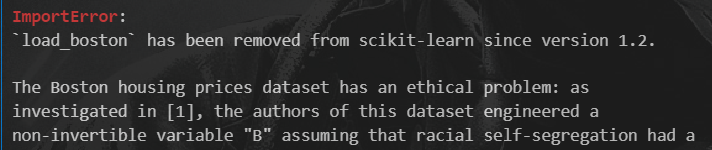
# Regression Techniques

## Linear Regression

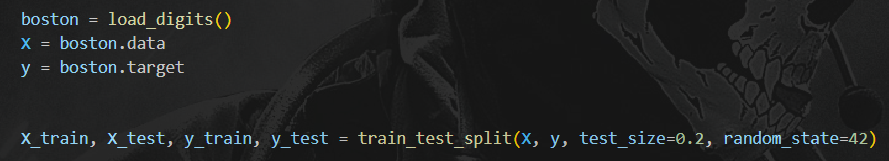
Dataset: Boston Housing Dataset (digits dataset)

Objective: Build a linear regression model for predicting a continuous target variable. Method:

* The Boston housing prices dataset has been removed from scikit-learn since version 1.2.



* The dataset was split into training and testing sets.



* A linear regression model was trained, and its performance was evaluated using Mean Squared Error (MSE) and R-squared score.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

## Ridge Regression

Dataset: Diabetes Dataset

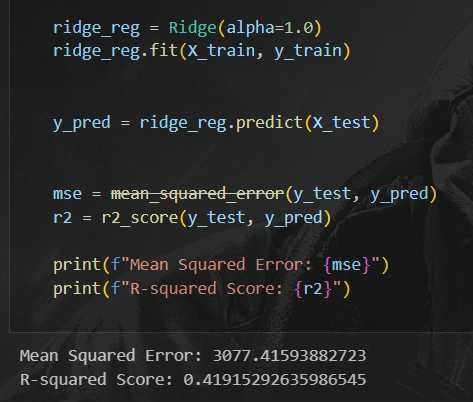
Objective: Use Ridge regression for regularized linear regression. Method:

* The dataset was split into training and testing sets.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

* A Ridge regression model was trained and evaluated using MSE and R-squared score.

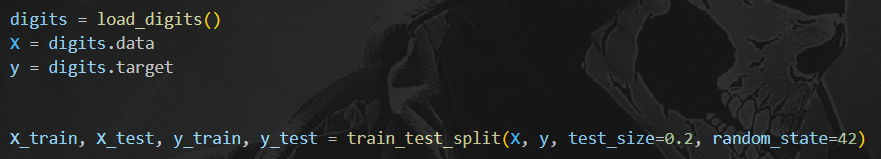


## Decision Tree Regression

Dataset: Boston Housing Dataset

Objective: Build and visualize a decision tree regression model. Method:

* The dataset was split into training and testing sets.



* A decision tree regressor was trained, and its performance was evaluated using MSE.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

* The decision tree was visualized.

