

**2 Assignment**

**Prepared by:** Amirkhan Serikbay

**ID:** 21B030716

**Faculty: SITE**

**Major:** Computer Systems and Software

Almaty, 2024

**Exercise 1: Feature Selection with SelectKBest**

**Objectives:** Use SelectKBest from scikit-learn to select the top k features from a dataset.

1. Load the Iris dataset from scikit-learn.

2. Split the dataset into features and target variable.

3. Use SelectKBest with the chi2 score function to select the top 2 features.

4. Print the selected feature names.

**Изображение выглядит как текст, снимок экрана, Шрифт

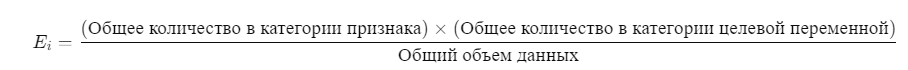
Автоматически созданное описание**

**Изображение выглядит как Шрифт, белый, зарисовка, диаграмма

Автоматически созданное описание**

**Изображение выглядит как текст, снимок экрана, Шрифт, документ

Автоматически созданное описание**

****

**Exercise 2: Feature Importance with Random Forest**

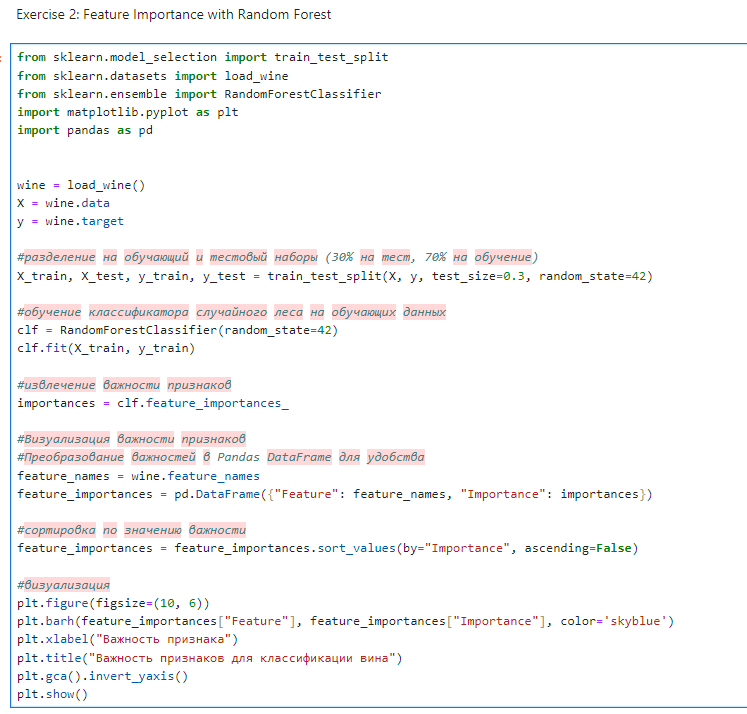
**Objectives:** Use a Random Forest classifier to determine feature importance.

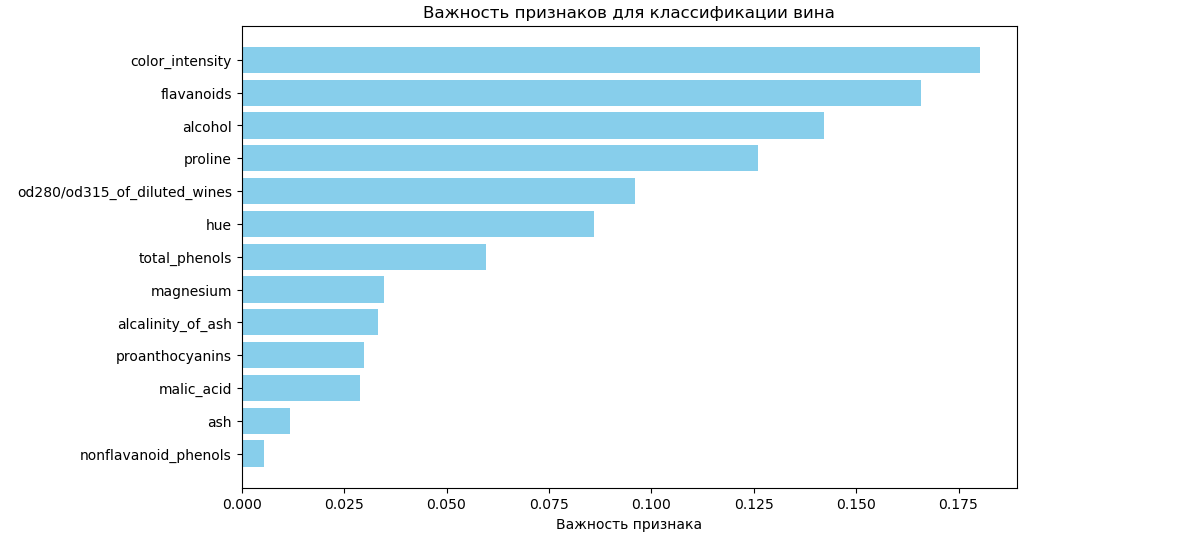
1. Load the Wine dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train a Random Forest classifier on the training data.

4. Extract and visualize feature importance.





RandomForestClassifier строится на случайном выборе подмножеств из исходных данных, то есть каждое дерево обучается на случайных данных и на случайных признаках.

В RandomForestClassifier важность признаков определяется на основе того, насколько часто и насколько существенно каждый признак улучшает критерий разделения в каждом дереве случайного леса. Важность признака отражает его вклад в классификацию или регрессию.

**Exercise 3: Recursive Feature Elimination (RFE)**

**Objectives:** Use Recursive Feature Elimination (RFE) to select features and evaluate model performance.

1. Load the Breast Cancer dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Use RFE with a Support Vector Machine (SVM) classifier to select features.

4. Train an SVM model with the selected features and evaluate its performance.

Изображение выглядит как текст, снимок экрана, веб-страница, Шрифт

Автоматически созданное описание



**Рекурсивное исключение признаков (Recursive Feature Elimination, RFE)** — это метод выбора признаков, который применяется для выявления наиболее значимых признаков в наборе данных. Он используется для уменьшения размерности и повышения производительности моделей машинного обучения, отбрасывая менее важные признаки и сохраняя только те, которые оказывают наибольшее влияние на целевую переменную.

**Формула для Support Vector Machine (SVM)** основывается на поиске оптимальной гиперплоскости, которая максимально разделяет два класса. Цель SVM — найти такую гиперплоскость, которая будет иметь максимальный зазор (margin) между точками (объектами) двух классов.

Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Автоматически созданное описание

**Exercise 4:** L1 Regularization for Feature Selection

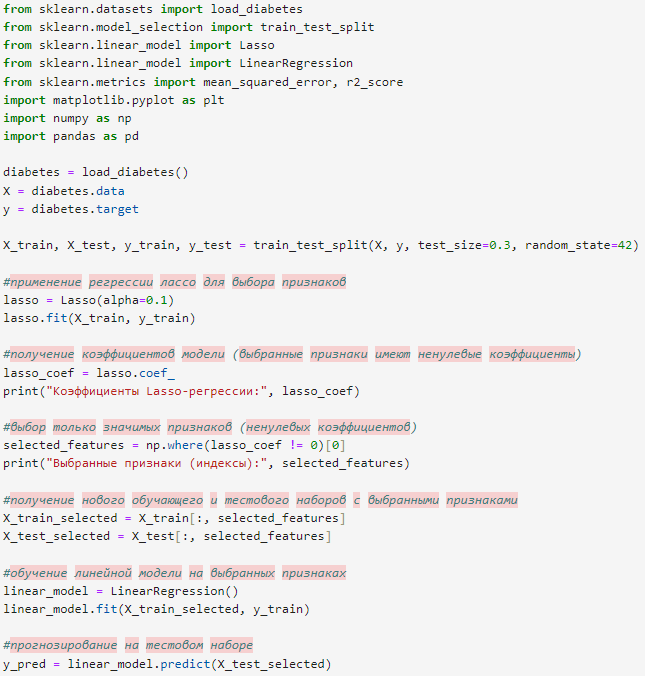
**Objective:** Use L1 regularization (Lasso) for feature selection.

1. Load the Diabetes dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Apply Lasso regression for feature selection.

4. Train a model using selected features and evaluate its performance.



Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, Шрифт, рукописный текст, снимок экрана

Автоматически созданное описание

**Classification Exercises**

**Exercise 1: Logistic Regression**

**Objective:** Build a logistic regression model to classify data.

1. Load the Iris dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train a logistic regression model on the training set.

4. Evaluate the model's performance on the test set using accuracy and a confusion matrix.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

**Логистическая регрессия** — это алгоритм машинного обучения, используемый для решения задач **классификации**. Её основная цель — предсказать, к какому классу принадлежит объект, на основе его признаков. В отличие от линейной регрессии, которая предсказывает числовое значение, логистическая регрессия **оценивает вероятность** того, что объект принадлежит к определенному классу, и переводит это значение в конечный класс (например, "0" или "1").

Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Автоматически созданное описание

**Exercise 2: Support Vector Machine (SVM)**

**Objective:** Use an SVM classifier to classify data.

1. Load the Breast Cancer dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train an SVM model on the training data.

4. Evaluate the model's performance on the test data using accuracy and a confusion matrix.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

**Exercise 3: Decision Tree Classifier**

**Objective:** Build a decision tree classifier and visualize it.

1. Load the Wine dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train a decision tree classifier on the training set.

4. Visualize the decision tree.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание 

**Изображение выглядит как текст, снимок экрана, Самоклеющийся листок, Шрифт

Автоматически созданное описание**

**Изображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание**

**Дерево решений** — это алгоритм машинного обучения, который **разделяет пространство признаков на области** путем последовательного применения **условий** на каждом шаге.Алгоритм строит **деревовидную структуру**, где каждый узел представляет собой **условие** на каком-либо признаке, а листья **— итоговую классификацию** или **предсказание.**

**Regression Exercises**

**Exercise 1: Linear Regression**

**Objective:** Build a linear regression model to predict a continuous target variable.

1. Load the Boston Housing dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train a linear regression model on the training set.

4. Evaluate the model's performance using mean squared error (MSE) and R-squared score.

**Изображение выглядит как текст, снимок экрана, документ

Автоматически созданное описание**

**Изображение выглядит как снимок экрана, текст, линия, График

Автоматически созданное описание**

**Линейная регрессия** — это алгоритм машинного обучения, который используется для **предсказания значения** непрерывной целевой переменной на основе одного или нескольких признаков. Основная идея линейной регрессии заключается в нахождении **линейной зависимости** между входными признаками и целевой переменной. Линейная модель описывается следующей формулой:

**Изображение выглядит как текст, снимок экрана, чек, Шрифт

Автоматически созданное описание**

**Exercise 2: Ridge Regression**

**Objective:** Use Ridge regression to perform regularized linear regression.

1. Load the Diabetes dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train a Ridge regression model on the training set.

4. Evaluate the model's performance using mean squared error (MSE) and R-squared score.

Изображение выглядит как снимок экрана, текст, линия, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

**Регрессия гребня** (Ridge Regression) — это разновидность линейной регрессии, которая включает **регуляризацию**. Регуляризация помогает **уменьшить переобучение** и **стабилизировать модель** при наличии мультиколлинеарности (когда признаки коррелируют друг с другом) или при использовании большого количества признаков. Это достигается путем добавления **штрафа** на величину коэффициентов модели.

Изображение выглядит как текст, снимок экрана, Шрифт, чек

Автоматически созданное описание

**Exercise 3: Decision Tree Regression**

**Objective:** Build a decision tree regression model and visualize it.

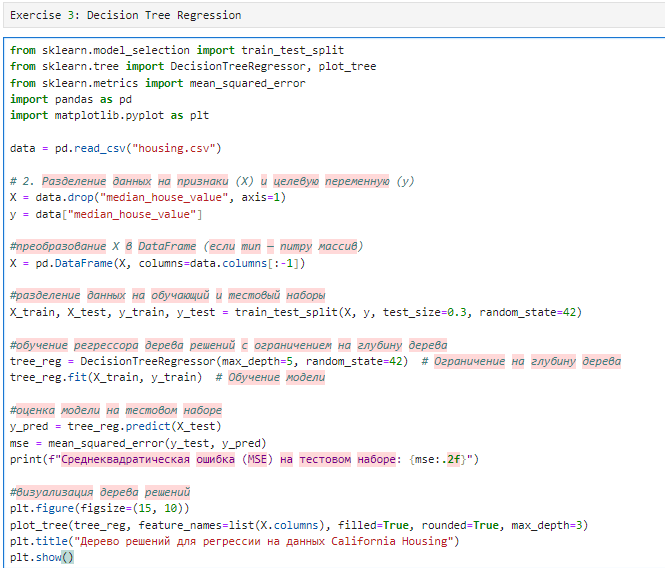
1. Load the Boston Housing dataset from scikit-learn.

2. Split the dataset into training and testing sets.

3. Train a decision tree regressor on the training set.

4. Evaluate the model's performance using mean squared error (MSE).

5. Visualize the decision tree.



Изображение выглядит как диаграмма, линия, штатив

Автоматически созданное описание

**Регрессия дерева решений** — это метод машинного обучения, который используется для **предсказания числовой целевой переменной** (регрессии) путем создания структуры дерева решений. Дерево решений делит пространство признаков на области, чтобы минимизировать ошибку предсказания.