# Лабораторная работа 2: Реализация моделей машинного обучения с использованием Scikit-learn

**Цель**
Цель данной лабораторной работы — реализовать простые модели машинного обучения, такие как линейная регрессия и классификация, используя библиотеку Scikit-learn. Студенты должны научиться строить, обучать и оценивать модели машинного обучения, а также визуализировать результаты их работы.

**Описание**
В рамках лабораторной работы студенты будут работать с различными задачами машинного обучения, включая линейную регрессию и классификацию. Каждому студенту будет назначено индивидуальное задание, которое поможет освоить конкретные аспекты реализации и оценки моделей. Лабораторная работа включает в себя построение моделей, оценку их точности, а также визуализацию результатов с использованием библиотек Python, таких как Matplotlib и Seaborn.

**Общие задачи**
1. **Построение модели**: Реализовать модель машинного обучения в соответствии с заданием (линейная регрессия или классификация).
2. **Обучение модели**: Обучить модель на предоставленных данных.
3. **Оценка модели**: Оценить производительность модели, используя соответствующие метрики точности.
4. **Визуализация результатов**: Построить графики и визуализации для анализа результатов работы модели.
5. **Отчёт**: Подготовить отчёт о выполненной работе, включая описание метода, результаты и визуализации.

**Индивидуальные задания**
Каждому студенту предоставляется уникальная задача в соответствии с его номером в списке группы (смотреть в SSO).
**1. Линейная регрессия на наборе данных о недвижимости:**
- **Задача:** Загрузить набор данных с информацией о недвижимости (например, площадь, количество комнат, год постройки) и предсказать цену дома, используя линейную регрессию. Оценить точность модели с помощью MSE (Mean Squared Error).
- **Визуализация:** Построить график зависимости реальной и предсказанной цены дома. Визуализировать линейную регрессионную прямую на фоне данных.

**2. Классификация видов ирисов:**
- **Задача:** Использовать набор данных Iris для построения модели классификации, которая будет предсказывать вид ириса на основе длины и ширины лепестков и чашелистиков. Оценить точность модели с использованием метрики F1-score.
- **Визуализация:** Построить диаграммы рассеяния (scatter plot) для различных пар признаков (например, длина и ширина лепестков), раскрасив точки в зависимости от предсказанного класса.

3. Многофакторный анализ производительности студентов:
Задача: Построить модель логистической регрессии для предсказания, сдаст ли студент экзамен, на основе данных о его посещаемости, оценках по контрольным работам и времени, затраченном на подготовку.

Визуализация: Визуализировать распределение студентов по метке класса (сдал/не сдал экзамен) с учетом различных факторов (посещаемость, оценки).

4. Классификация данных о кредитоспособности:
Задача: Загрузить набор данных о кредитах (например, возраст, доход, кредитная история) и построить модель классификации, предсказывающую, выдадут ли кредит. Использовать ROC-кривую для оценки качества модели.
Визуализация: Построить ROC-кривую для оценки модели и отобразить AUC (Area Under Curve). Также можно визуализировать матрицу ошибок (confusion matrix).

5. Регрессия на данных о расходе топлива:
Задача: Использовать набор данных о характеристиках автомобилей (объем двигателя, вес, тип топлива) для предсказания расхода топлива на 100 км, применяя полиномиальную регрессию. Оценить точность модели с помощью коэффициента детерминации ($R^2$).
Визуализация: Построить график зависимости реального и предсказанного расхода топлива. Визуализировать полиномиальную регрессионную кривую на фоне данных.

6. Классификация писем как спам или не спам:
Задача: Загрузить набор данных с письмами и построить модель классификации для распознавания спама. Применить методы работы с текстом (TF-IDF) и оценить модель с использованием Precision и Recall.
Визуализация: Визуализировать метрики Precision и Recall в виде столбчатых диаграмм, а также построить confusion matrix.

7. Классификация видов вина:
Задача: Использовать набор данных Wine и построить модель классификации для предсказания типа вина на основе его химических характеристик. Оценить точность модели с использованием confusion matrix.
Визуализация: Построить диаграммы рассеяния для пар признаков и визуализировать разделяющую поверхность, построенную моделью классификации.

8. Регрессия для предсказания уровня арендной платы:
Задача: Использовать данные о недвижимости для предсказания ежемесячной арендной платы, применяя Ridge-регрессию. Оценить модель с помощью RMSE (Root Mean Squared Error).
Визуализация: Визуализировать распределение ошибок модели (остатки) на графике, чтобы увидеть, насколько точно модель предсказывает арендную плату.

9. Классификация сердечных заболеваний:
Задача: Построить модель классификации для предсказания наличия сердечных заболеваний на основе медицинских данных пациента. Оценить качество модели с использованием AUC (Area Under Curve).
Визуализация: Построить ROC-кривую и визуализировать матрицу ошибок для оценки работы модели.

10. Регрессия на данных о погоде:
Задача: Использовать линейную регрессию для предсказания температуры на основе данных о влажности, скорости ветра и давления. Оценить качество модели с помощью $R^2$.

Визуализация: Построить график зависимости реальной и предсказанной температуры. Визуализировать линейную регрессионную линию.

11. Классификация изображений цифр:
Задача: Использовать набор данных MNIST для построения модели классификации цифр с использованием метода k-ближайших соседей (k-NN). Оценить точность модели с помощью accuracy score.
Визуализация: Построить матрицу ошибок (confusion matrix), чтобы увидеть, какие цифры чаще всего ошибочно классифицируются. Также можно визуализировать несколько примеров предсказанных изображений.

12. Регрессия для предсказания уровня заработной платы:
Задача: Построить модель линейной регрессии для предсказания заработной платы на основе возраста, уровня образования и стажа работы. Оценить качество модели с использованием MAE (Mean Absolute Error).
Визуализация: Визуализировать предсказанную и реальную заработную плату на графике, добавив регрессионную линию.

13. Классификация отзывов как положительных или отрицательных:
Задача: Загрузить набор данных с отзывами пользователей и построить модель классификации для предсказания, положительный или отрицательный отзыв. Применить методы работы с текстом и оценить модель с использованием Precision и Recall.
Визуализация: Построить Word Cloud для положительных и отрицательных отзывов. Визуализировать матрицу ошибок для модели классификации.

14. Классификация заболеваний на основе медицинских данных:
Задача: Использовать набор данных о пациентах и построить модель классификации для предсказания диагноза. Оценить точность модели с использованием F1-score.
Визуализация: Визуализировать ROC-кривую, матрицу ошибок и столбчатые диаграммы по метрикам Precision, Recall, F1-score.

15. Регрессия на данных о продажах:
Задача: Использовать полиномиальную регрессию для предсказания объема продаж на основе данных о рекламных затратах в различных каналах (телевидение, радио, интернет). Оценить модель с помощью MSE.
Визуализация: Построить график зависимости реальных и предсказанных продаж. Визуализировать полиномиальную регрессионную кривую на фоне данных.

**Вопросы для подготовки**

1. Какие основные этапы включает процесс реализации модели машинного обучения?
2. Что такое линейная регрессия и как она работает? Какие типы линейной регрессии существуют?
3. Каковы основные метрики для оценки качества модели машинного обучения?
4. Как работает метод опорных векторов (SVM) и какие его основные параметры?
5. Что такое регуляризация в контексте линейной регрессии, и как она помогает улучшить модель?

6. Каковы преимущества и недостатки методов классификации, таких как k-ближайших соседей (k-NN) и деревья решений?
7. Что такое перекрестная проверка и как она помогает в оценке модели?
8. Как интерпретировать матрицу ошибок и какие метрики точности можно извлечь из неё?
9. Какие методы существуют для обработки пропущенных данных в наборах данных?
10. Как можно визуализировать результаты работы модели машинного обучения, и какие графики наиболее полезны?

**Критерии Оценки:**
- Написание кода для решения индивидуальной задачи: 2 балл
- Объяснение и понимание написанного кода во время защиты: 1 балла
- Ответ на один из теоретических вопросов, выбранный преподавателем: 1 балл
***Обязательно предоставьте отчет в формате Word, соответствующий всем стандартам.***

**Дополнительное оценивание:**
- Выполнение основных задач лабораторной работы (построение, обучение и оценка моделей) без визуализации: **4 балла**.
- Дополнительный балл: Если вы добавите визуализацию результатов вашей работы (например, графики зависимости реальных и предсказанных значений, распределение классов, матрицы ошибок и т.д.), вы получите **дополнительный 1 балл**. Этот дополнительный балл может быть добавлен к вашему итоговому баллу, если не хватает баллов для достижения максимального результата.