МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Вычислительной техники

ОТЧЕТ

по лабораторной работе № 2

по дисциплине «Машинное обучение»

Тема: «Подготовка и нормализация данных»

Студент гр. 0308 Сабурова Е.А.

Преподаватель Гатауллин Р.И.

Санкт-Петербург2023

Оглавление

1. Цель работы

Получение и закрепление навыков предобработки данных и применения методов машинного обучения для решения задач классификации.

2. Задание

1) Обучение моделей и подбор параметров (желательно с использованием Grid Search):
    a. K-ближайших соседей (KNN)
    b. Машина опорных векторов (SVM)
    c. Дерево решений ИЛИ Случайный лес
2) Оценка моделей
    a. Визуализация предсказанных значений
    b. Оценка качества прогноза (precision/recall/f1-score/ROC-AUC)
    c. Визуализация дерева решений ИЛИ Визуализация Feature Importance для случайного леса

3. Краткая теоретическая информация по теме работы

В машинном обучении существует задача разделения множества наблюдений (объектов) на группы, называемые классами, на основе анализа их формального описания. При классификации каждая единица наблюдения относится определенной группе или номинальной категории на основе некоторого качественного свойства.

В машинном обучении задача классификации решается с использованием обучения с учителем, поскольку классы определяются заранее и для примеров обучающего множества метки классов заданы. Аналитические модели, решающие задачу классификации, называются классификаторами.

Задача классификации представляет собой одну из базовых задач прикладной статистики и машинного обучения, а также искусственного интеллекта в целом. Это связано с тем, что классификация является одной из наиболее понятных и простых для интерпретации технологий анализа данных, а классифицирующие правила могут быть сформулированы на естественном языке.

Задача классификации применяется во многих областях:

- в торговле — классификация клиентов и товаров позволяет оптимизировать маркетинговые стратегии, стимулировать продажи, сокращать издержки;
- в сфере телекоммуникаций — классификация абонентов позволяет определять уровень лояльности, разрабатывать программы лояльности;
- в медицине и здравоохранении — диагностика заболеваний, классификация населения по группам риска;
- в банковской сфере — кредитный скоринг.

Алгоритмы моделей машинного обучения, использованные в работе

Дерево принятия решений

Это метод поддержки принятия решений, основанный на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия (с расчётом вероятности наступления того или иного события), эффективность, ресурсозатратность.

Это дерево складывается из минимального числа вопросов, предполагающих однозначный ответ — «да» или «нет». Последовательно дав ответы на все эти вопросы, мы приходим к правильному выбору. Методологические преимущества дерева принятия решений – в том, что оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов.

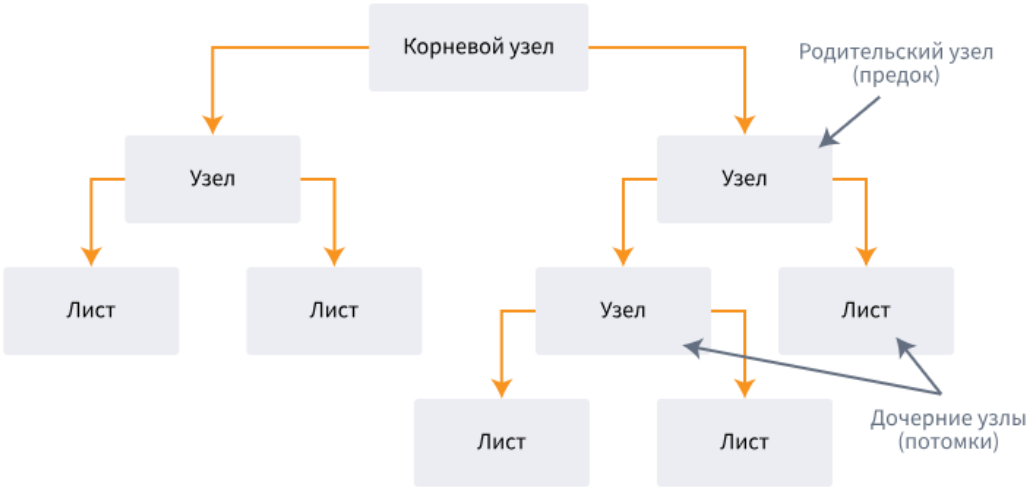

Рисунок 1 – изображение работы дерева решений

Метод опорных векторов (SVM)

Это целый набор алгоритмов, необходимых для решения задач на классификацию и регрессионный анализ. Исходя из того, что объект, находящийся в N-мерном пространстве, относится к одному из двух классов, метод опорных векторов строит гиперплоскость с мерностью (N – 1), чтобы все объекты оказались в одной из двух групп. На бумаге это можно изобразить так (рисунок 2): есть точки двух разных видов, и их можно линейно разделить. Кроме сепарации точек, данный метод генерирует гиперплоскость таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы.

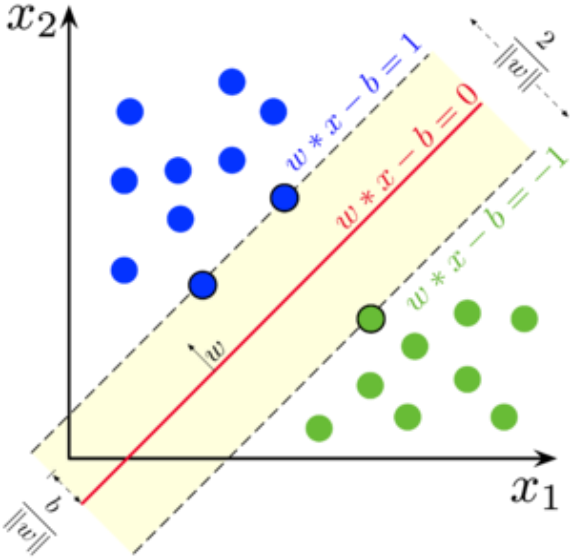

Рисунок 2 – изображение SVM

Метод k-ближайших соседей (K-nearest neighbor)

Метод k-ближайших соседей относит объекты к классу, которому

принадлежит большинство из k его ближайших соседей в многомерном пространстве признаков. Это один из простейших алгоритмов обучения классификационных моделей.

Число k — это количество соседних объектов в пространстве признаков, которые сравниваются с классифицируемым объектом. Иными словами, если k=10, то каждый объект сравнивается с 10-ю соседями.

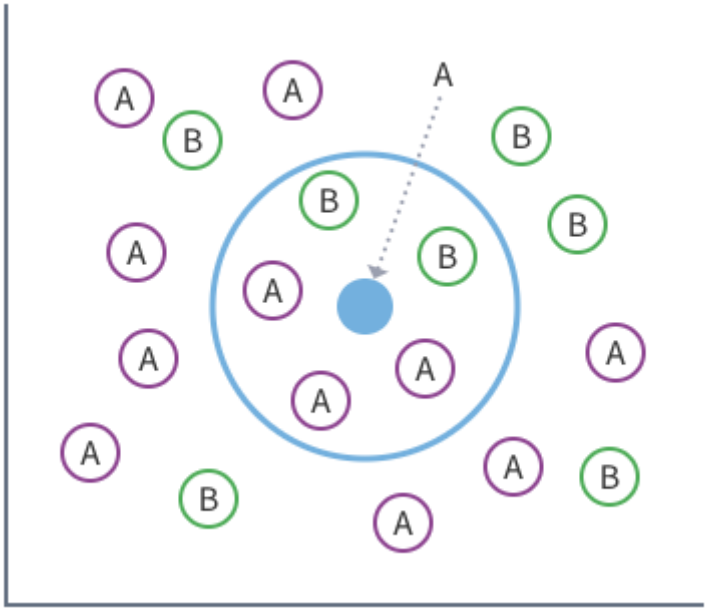

Рисунок 3 – изображение kNN

4. Ход работы

Обучение моделей и подбор параметров с помощью Grid Search

Для обучения модели необходимо разбить выборку на 2 части: тренировочные данные и тестовые.

Данные были разбиты в пропорции 80:20

Количество данных для тренировки: 14444

Количество данных для теста: 3611

К ближайших соседей

В качестве параметра для перекрестной проверки использовался параметр количества ближайших 5 соседей.

Метод опорных векторов (SVM)

В качестве параметра для перекрестной проверки рассматривались параметры:

– Тип ядра: rbf

– Параметр регуляризации: 1

Дерево решений

В качестве параметра для перекрестной проверки рассматривались параметры:

– Критерии разделения: entropy

– Делитель: best

– Глубина дерева: (4, 16, 64, 256)

Оценка моделей

Визуализация предсказаний

Для визуализации предсказаний использовались confusion matrix. По оси y – реальные данные, по оси x – предсказанные значения. На пересечении указано количество пересечений.

Чем больше числа на главной диагонали, тем модель точнее.

Результаты для всех моделей представлены на рисунке:

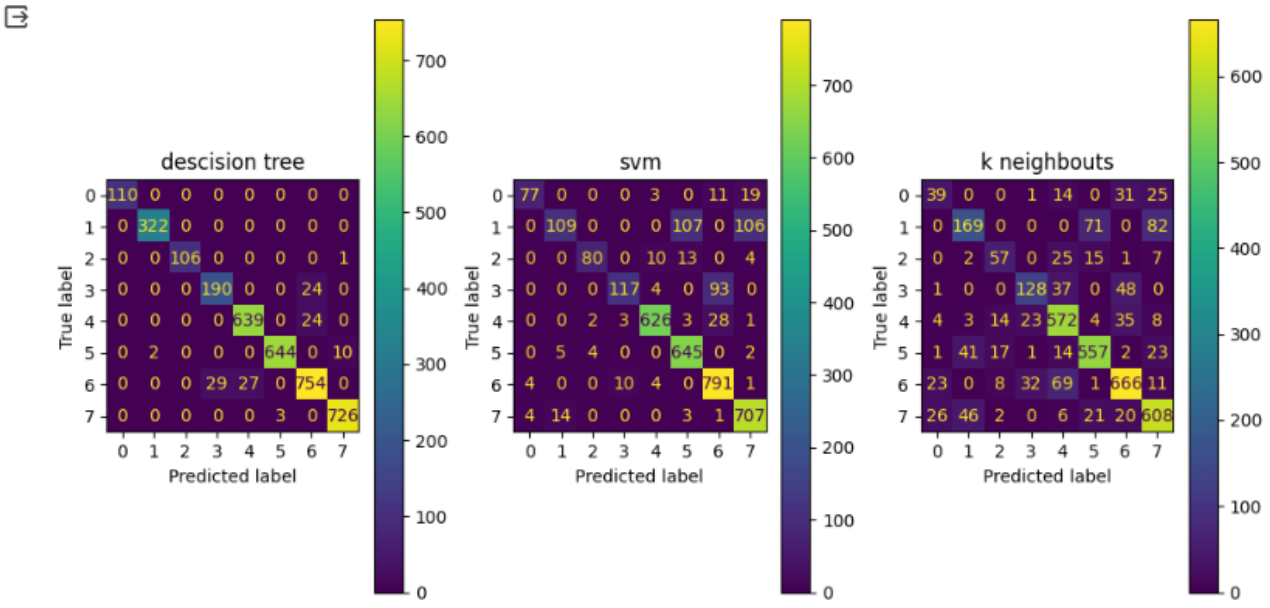

Рисунок 5 – confusion matrix

Лучшей моделью является Дерево решений, худшей – k ближайших соседей.

Оценка по метрикам

Результаты сравнения моделей по всем метрикам представлены на рисунке.

	model	precision_score	recall_score	f1_score	accuracy
0	knn (scaled)	0.696229	0.672288	0.682246	0.774301
1	svm (scaled)	0.886190	0.775836	0.807089	0.872888
2	descision tree (scaled)	0.967682	0.968845	0.968232	0.966768

Рисунок 6 – оценка моделей по основным метрикам

Лучшей моделью является Дерево решений, худшей – k ближайших соседей. Подтвердились выводы, сделанные на предыдущем шаге. На втором месте по качеству – метод опорных векторов.

Ниже на графиках представлены визуализации метрик для каждой модели.

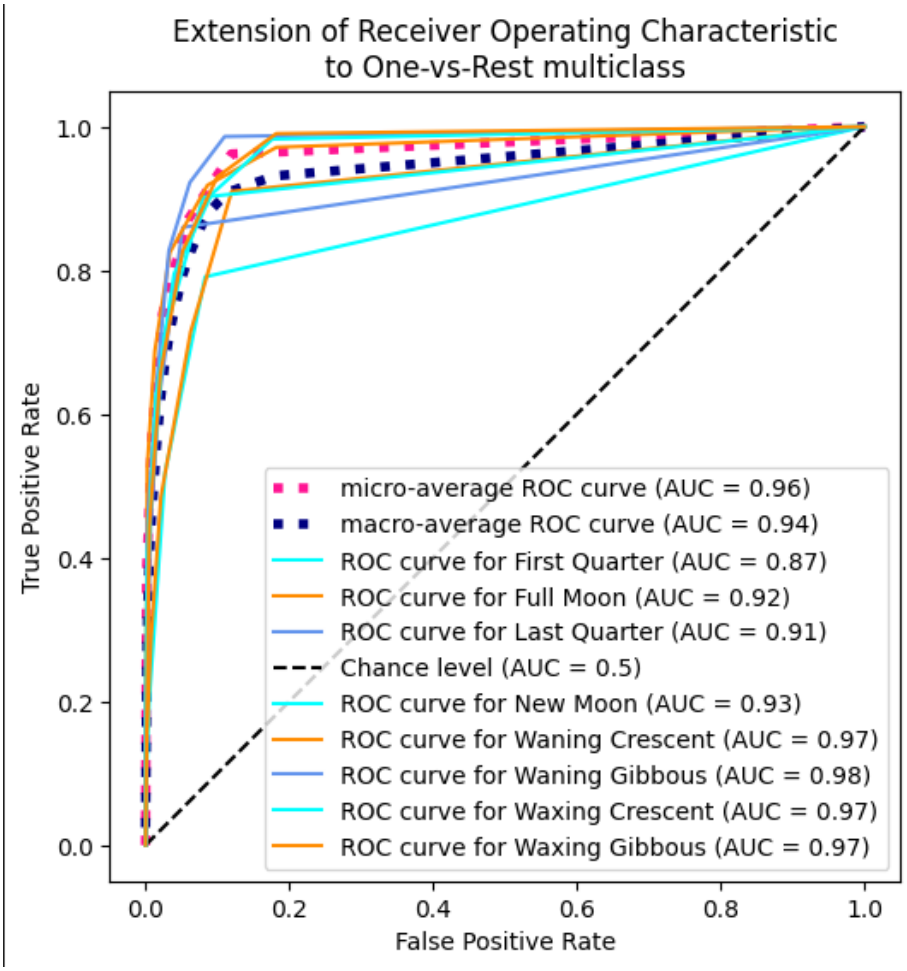

Рисунок 7 – график ROC-AUC для kNN

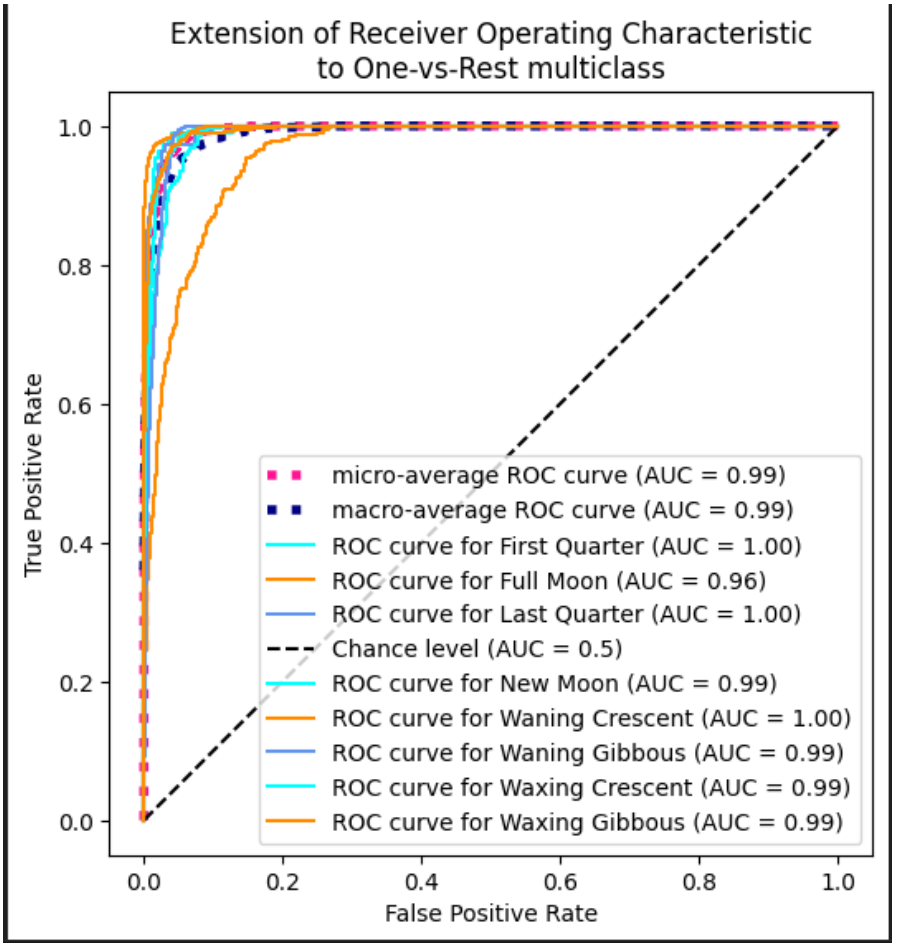

Рисунок 8 – график ROC-AUC для SVM

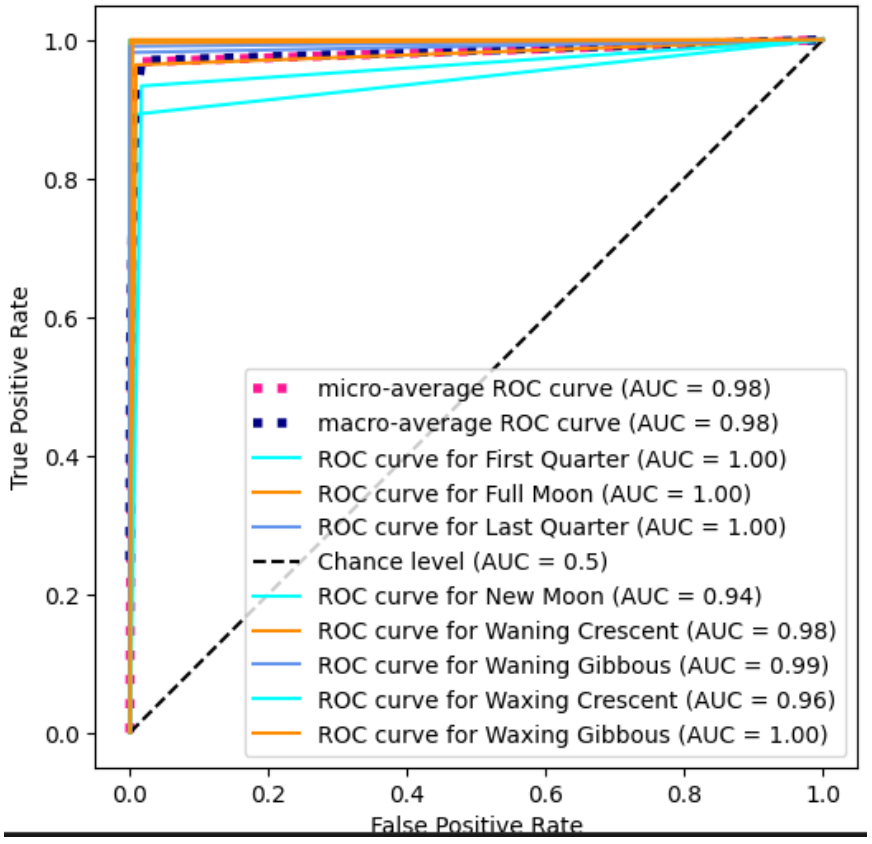

Рисунок 9 – график ROC-AUC для DT

Визуализация дерева решений

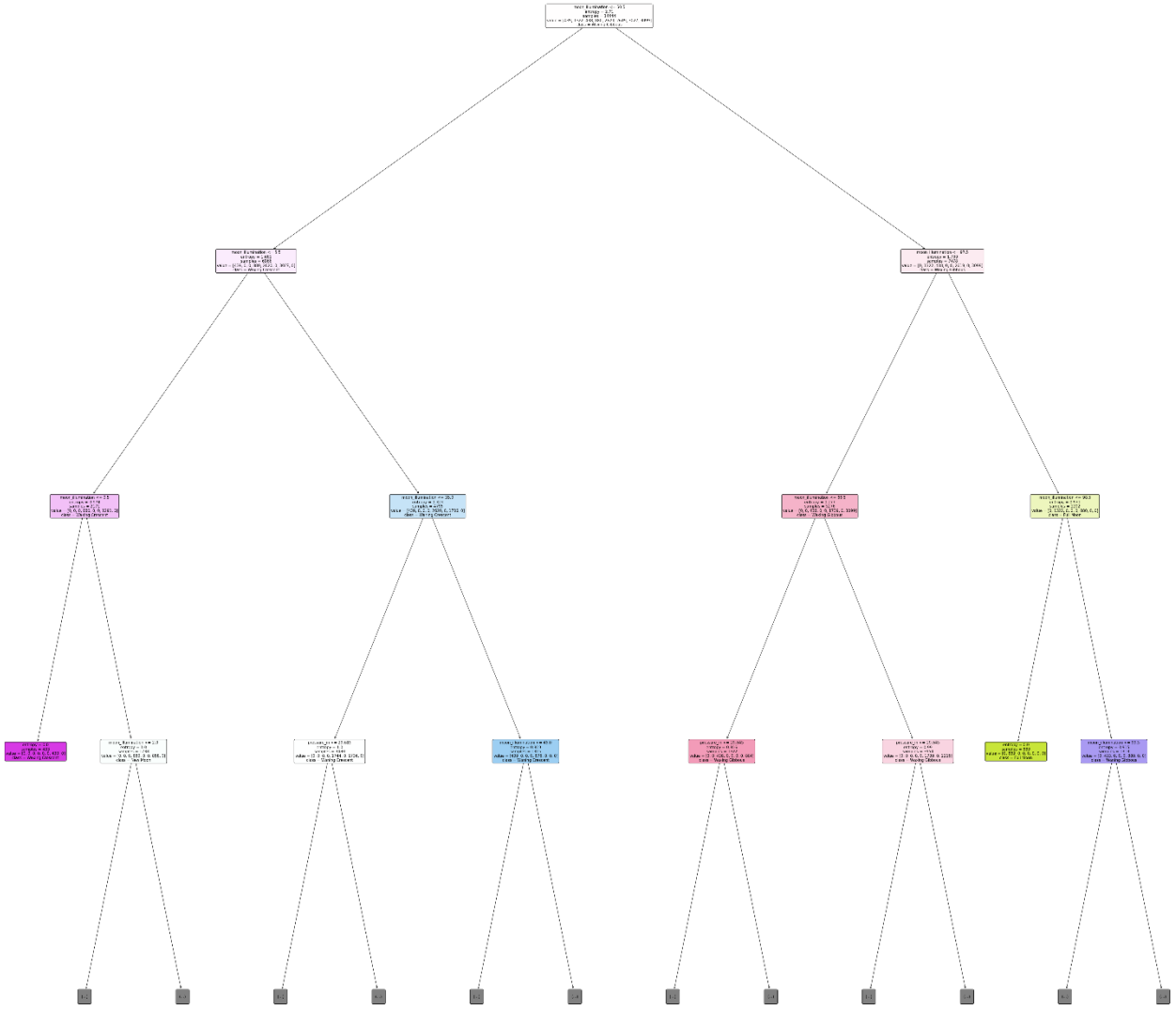

Рисунок 10 – Визуализация дерева решений

5. Выводы

В ходе лабораторной работы оптимальным классифкатором оказался классификатор на основе дерева решений.

Метод опорных векторов также хорош в задачах классификации.

Для подбора параметров можно использовать, встроенный в sklearn метод поиска по сетке. С помощью него можно найти оптимальные параметры модели.