МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное   
учреждение высшего образования

**"Южно-Уральский государственный университет**

**(национальный исследовательский университет)"**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

**ОТЧЕТ**о выполнении лабораторной работы № 4

по дисциплине  
«Технологии аналитической обработки информации»

|  |  |
| --- | --- |
|  | Выполнил: студент группы КЭ-403  Разуев Г.А.  Проверил:  Преподаватель кафедры СП  Гоглачев А.И. |

Челябинск-2025

ОГЛАВЛЕНИЕ

[1. ЗАДАНИЕ 2](#_Toc191729743)

[2. РЕАЛИЗАЦИЯ ПРОГРАММЫ 3](#_Toc191729744)

[3. ЭКСПЕРИМЕНТЫ 4](#_Toc191729745)

# ЗАДАНИЕ

1. Разработайте программу, которая выполняет классификацию заданного набора данных с помощью одной из техник ансамблевой классификации. Параметрами программы являются набор данных, ансамблевая техника (бэггинг, случайный лес или бустинг), количество участников ансамбля, а также параметры в соответствии с выбранной техникой ансамблевой классификации.
2. Проведите эксперименты на наборе данных из задания Классификация с помощью дерева решений, варьируя количество участников ансамбля (от 50 до 100 с шагом 10).
3. Выполните визуализацию полученных результатов в виде следующих диаграмм:
   1. показатели качества классификации в зависимости от количества участников ансамбля для заданного набора данных; нанесите на диаграмму соответствующие значения, полученные в задании Классификация с помощью дерева решений.
4. Подготовьте отчет о выполнении задания и загрузите отчет в формате PDF в систему. Отчет должен представлять собой связный и структурированный документ.

# РЕАЛИЗАЦИЯ ПРОГРАММЫ

Реализация программы представлена в классе «EnsebleModel», который инкапсулирует логику по обучению и использованию модели, а также по предобработке данных и их разбиении при обучении. Один из ключевых атрибутов класса – техника ансамблирования. Данный класс реализует стандартный интерфейс Scikit-Learn (имеет методы «fit» и «predict»), однако помимо данных метод «fit» имеет параметр размера тестовой выборки, что позволяет как просто обучать модель на всех данных, так и экспериментировать, сразу вычисляя метрики качества на тренировочных и тестовых данных. Для предобработки сырых данных реализован метод «\_preprocess\_data», который является приватным. Он создает объекты препроцессоров («StandardScaler» для числовых данных и «OneHotEncoder» для категориальных), а потом в зависимости от параметров обучения модели применяет преобразования к данным.

Код реализованной программы и всех проведенных экспериментов находится в репозитории по ссылке https://github.com/ke103rga/analytics.

Программа была протестирована полном наборе данных «Census Income», который содержал как числовые, так и категориальные данные.

# ЭКСПЕРИМЕНТЫ

Были проведены эксперименты на наборе данных, с фиксацией техники ансамблирования и варьированием количеством базовых классификаторов от 50 до 100 с шагом 10. Код проведения эксперимента представлен в листинге 1.

Листинг 1 – Код применения реализованной программы

for technique in techniques:

for n\_estimator in tqdm(n\_estimators, desc=f'Estimators for {technique}', leave=True):

model = EnsembleModel(technique, n\_estimators=n\_estimator)

# \_data = full\_df.head(1000)

# model.fit(\_data.drop(columns=['target']), \_data['target'], test\_size=0.3)

model.fit(full\_df.drop(columns=['target']), full\_df['target'], test\_size=0.3)

# Сохранение метрик

results[technique]["accuracy"].append(model.accuracy)

results[technique]["precision"].append(model.precision)

results[technique]["recall"].append(model.recall)

results[technique]["f1\_score"].append(model.f1\_score)

На основании проведенных экспериментов была выполнена визуализация результатов с использованием линейной диаграммы, на которой ось X – значение размера тестовой выборки, ось Y – метрика качества классификации. Для каждой метрики качества были построены свои сравнительные диаграммы, которые представлены на рисунках 1 – 4.

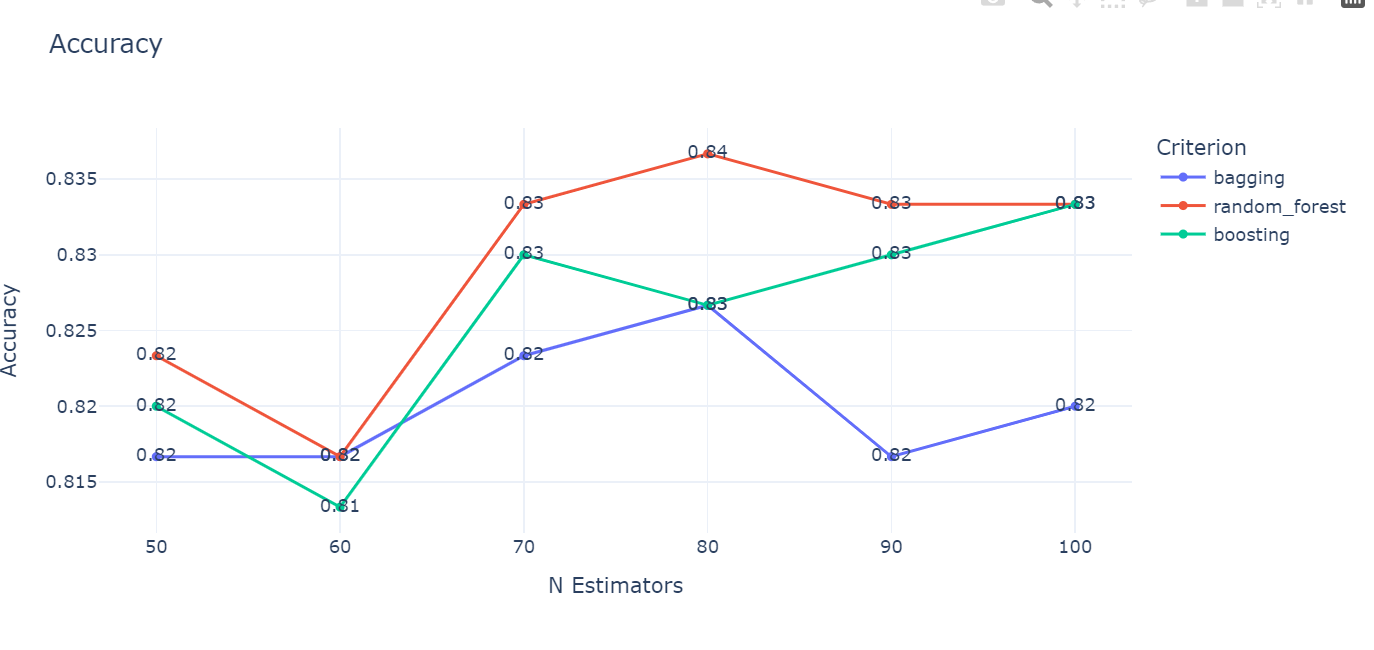


Рисунок 1 – Диаграмма сравнения Accuracy



Рисунок 2 – Диаграмма сравнения Precision



Рисунок 3 – Диаграмма сравнения Recall



Рисунок 4 – Диаграмма сравнения F1 Score

Из анализа данных графиков можно сделать ряд выводов:

1. По совокупности метрик правильнее всего для решения данной задачи будет использовать технику случайного леса.;
2. С ростом числа базовых классификаторов почти всегда наблюдается рост метрик;
3. Технику бэггинга уместно использовать, когда есть потребность найти как можно больше положительных значений класса, технику бустинга уместно использовать, когда необходимо сделать наиболее точные предсказания.

Также была построена визуализация, позволяющая сравнить лучшие показатели точности кластеризации каждой техники ансамблирования и обычного дерева решений. График сравнения приведен на рисунке 5.

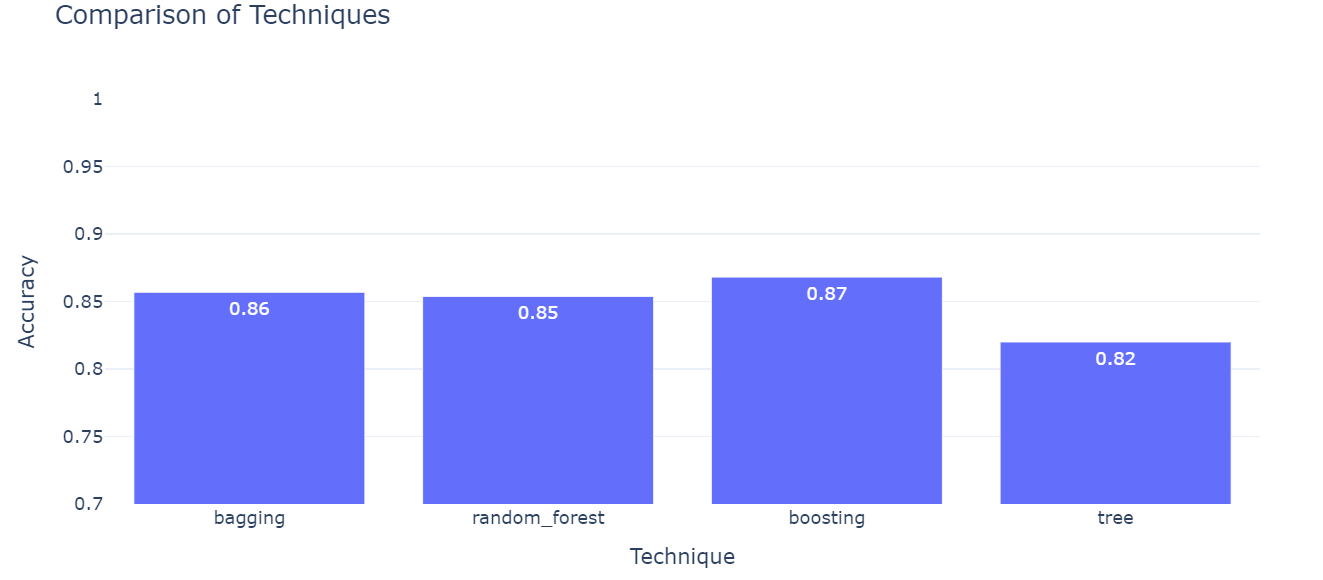


Рисунок 5 – Диаграмма сравнения лучших показателей качества