Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Анализ и классификация данных переписи населения с использованием моделей машинного обучения

Студент ИУ5-34М _____ **Д.Д. Ваганов**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Ю.Е.Гапанюк**
(Подпись, дата) (И.О.Фамилия)

2024г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ____ИУ5____
(Индекс)
_____ ___В.И. Терехов___
(И.О.Фамилия)
« _____ » _____ 2024 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме __Анализ и классификация данных переписи населения с использованием моделей машинного обучения_____
_____

Студент группы ____ИУ5-34М_____

_____Ваганов Даниил Дмитриевич_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

_____исследовательская_____

Источник тематики (кафедра, предприятие, НИР) _____

График выполнения НИР: 25% к _5_ нед., 50% к _9_ нед., 75% к 13 нед., 100% к 16 нед.

Техническое задание _Разработать и реализовать систему анализа и классификации данных переписи населения на основе различных моделей машинного обучения с целью предсказания уровня дохода._____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _15_ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

_____
_____
_____

Дата выдачи задания «__» _____ 2024 г.

Руководитель НИР _____ ___Ю.Е. Гапанюк___
(Подпись, дата) (И.О.Фамилия)

Студент _____ **Д.Д. Ваганов**
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Введение

В настоящее время анализ больших объемов данных становится неотъемлемой частью принятия решений в различных сферах деятельности. Данные переписи населения представляют собой ценный источник информации для изучения демографических, социальных и экономических характеристик общества. Однако эффективное использование этих данных требует применения современных методов обработки и анализа.

Машинное обучение предоставляет мощные инструменты для классификации и прогнозирования, позволяя выявлять скрытые закономерности и тенденции в данных. Применение различных моделей машинного обучения к данным переписи населения может значительно повысить точность прогнозов и упростить процесс принятия решений в социальной политике и экономическом планировании.

В данном проекте рассматривается задача классификации данных переписи населения с использованием пяти различных моделей машинного обучения и анализа корреляции между признаками. Основная цель работы заключается в сравнении эффективности этих моделей, выявлении наиболее подходящей для решения поставленной задачи и исследовании взаимосвязей между различными характеристиками данных.

Описание исследуемых данных

В данной работе будет использован набор данных, извлеченный из базы данных Бюро переписи населения США за 1994 год. Датасет содержит 32561 записей, включающие такие данные как возраст, пол, тип трудоустройства, уровень образования, семейное положение и прочие. Основная задача заключается в предсказании уровня дохода человека – превышает ли он $50,000 в год.

Описание признаков:

Age – Возраст (числовой).

Workclass – Тип трудоустройства (категориальный): частный сектор, государственный сектор, работа на себя и др.

Final Weight (fnlwgt) – Итоговый вес (числовой). Представляет собой вес, присвоенный каждому респонденту для обеспечения репрезентативности выборки. Используется для учета демографических характеристик и приведения данных к национальным показателям.

Education – Уровень образования (категориальный): бакалавр, среднее образование, начальное профессиональное и др.

Education-Num – Числовое представление уровня образования.

Marital Status – Семейное положение (категориальный): женат, разведен, никогда не был женат, вдовец и др.

Occupation – Профессия (категориальный): техническая поддержка, продажи, управление, профессиональные специальности и др.

Relationship – Родственные отношения (категориальный): жена, муж, ребенок, отсутствует в семье и др.

Race – Расовая принадлежность (категориальный).

Sex – Пол (категориальный): мужчина, женщина.

Capital Gain – Прирост капитала (числовой).

Capital Loss – Потери капитала (числовой).

Hours per Week – Количество рабочих часов в неделю (числовой).

Native Country – Страна рождения (категориальный): США, Германия, Япония, Китай, Индия и др.

Разведочный анализ данных

Проведем разведочный анализ данных для получения общего представление о структуре и содержании данных, выявим закономерности и аномалии, а также подготовим данные для последующего моделирования.

Перед началом анализа данные были предобработаны с целью обеспечения их качества и пригодности для машинного обучения. В рамках предобработки были выполнены следующие шаги:

- Удаление пропусков: записи с отсутствующими значениями были исключены или заполнены с использованием подходящих методов.
- Кодирование категориальных признаков: применен метод One-Hot Encoding для преобразования категориальных переменных в числовой формат.
- Нормализация данных: числовые признаки были масштабированы для приведения их к единому диапазону.
- Удаление выбросов: из выборки были исключены аномальные значения, влияющие на статистику данных.

В данном разделе будут представлены графики и визуализации, иллюстрирующие распределение признаков, взаимосвязи между переменными, а также результаты анализа корреляций.

На рисунке 1 представлен график распределение возраста респондентов в зависимости от уровня дохода (<=50K и >50K).

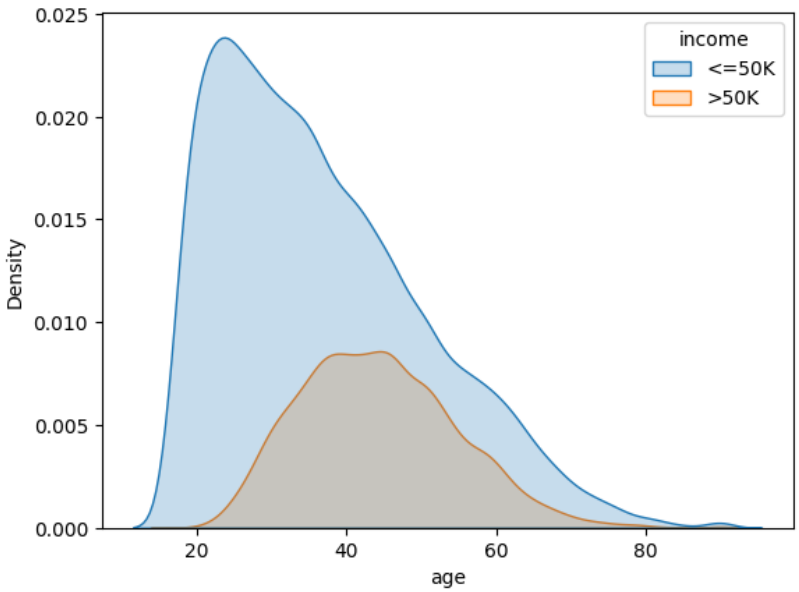

Рисунок 1. График распределение возраста в зависимости от дохода

Можно сделать следующие выводы, исходя из представленного графика:

1. Большинство респондентов, чей доход не превышает 50k, находятся в возрасте от 20 до 50 лет, с пиком около 30 лет.
2. Респонденты с доходом выше 50k встречаются реже и, как правило, имеют более высокий возрастной диапазон, с пиком около 40-50 лет.
3. С возрастом количество людей с низким доходом постепенно уменьшается, тогда как доля респондентов с доходом выше 50k возрастает до среднего возраста и снижается в пожилом возрасте.

На рисунке 2 показано распределение уровня образования (education.num) среди респондентов с разным уровнем дохода.

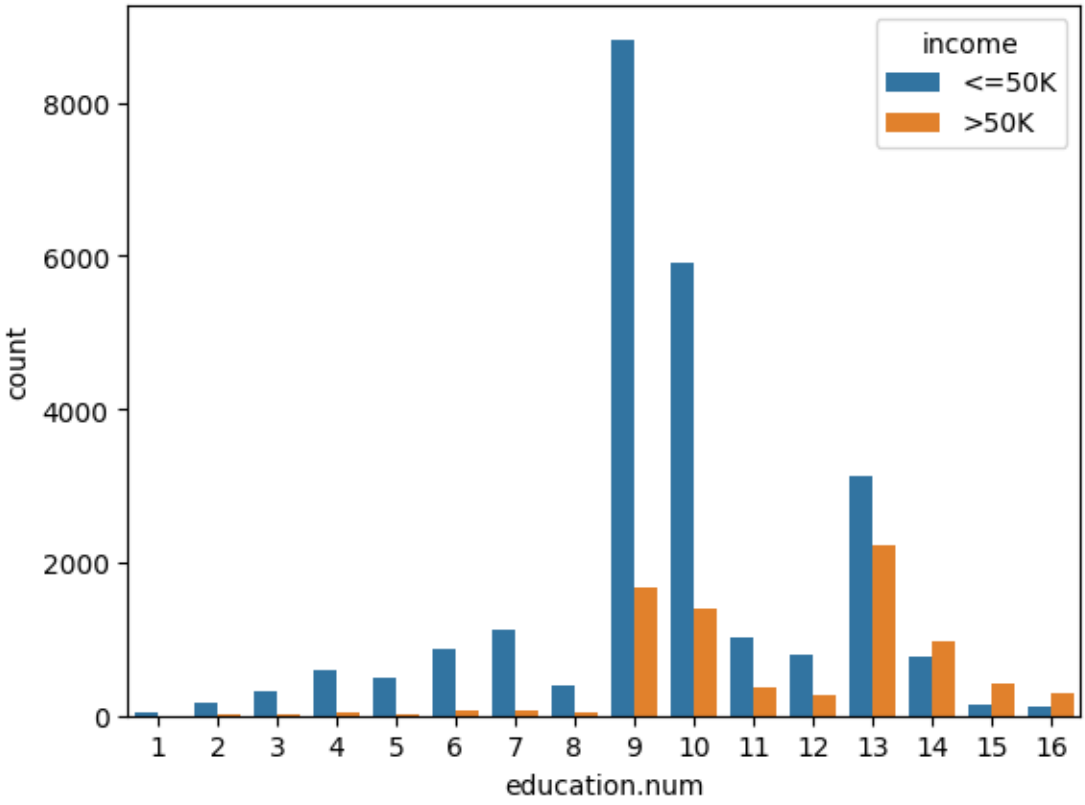

Рисунок 2. График распределения уровня образования по доходам

Анализ графика:

- Большинство респондентов имеют уровень образования, соответствующий значениям 9 и 10, что соответствует окончанию

средней школы (HS-grad) и некоторому обучению в колледже (Some-college).

- Среди людей с низким доходом (<=50k) преобладают респонденты с уровнями образования 9 и 10, что соответствует окончанию средней школы (HS-grad) и некоторому обучению в колледже (Some-college).
- Люди с более высоким доходом (>50k) чаще встречаются среди тех, кто имеет уровни образования 13 (степень бакалавра) и более высокие академические степени.
- Наблюдается явная тенденция увеличения доли респондентов с доходом >50K по мере повышения уровня образования.

На рисунке 3 показана матрица корреляции.

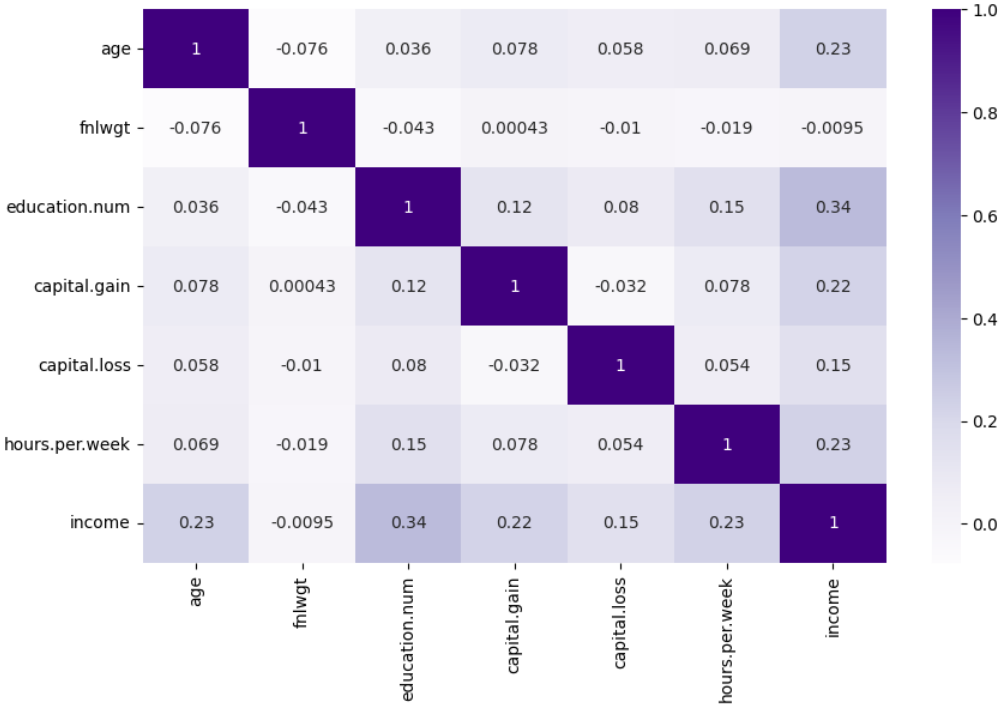

Рисунок 3. Матрица корреляции

Из матрицы можно выделить сильные положительные корреляции:

- Между образованием (Education-Num) и возрастом. Это логично, так как с возрастом люди, как правило, получают больше образования.

- Между образованием и часами работы в неделю. Это может быть связано с тем, что люди с более высоким образованием чаще занимают позиции, требующие большего рабочего времени.

Между большинством других признаков наблюдаются слабые корреляции. Это говорит о том, что эти признаки могут быть не сильно связаны между собой или их взаимосвязь более сложная и не линейная.

На рисунке 4 представлена диаграмма размаха возраста.

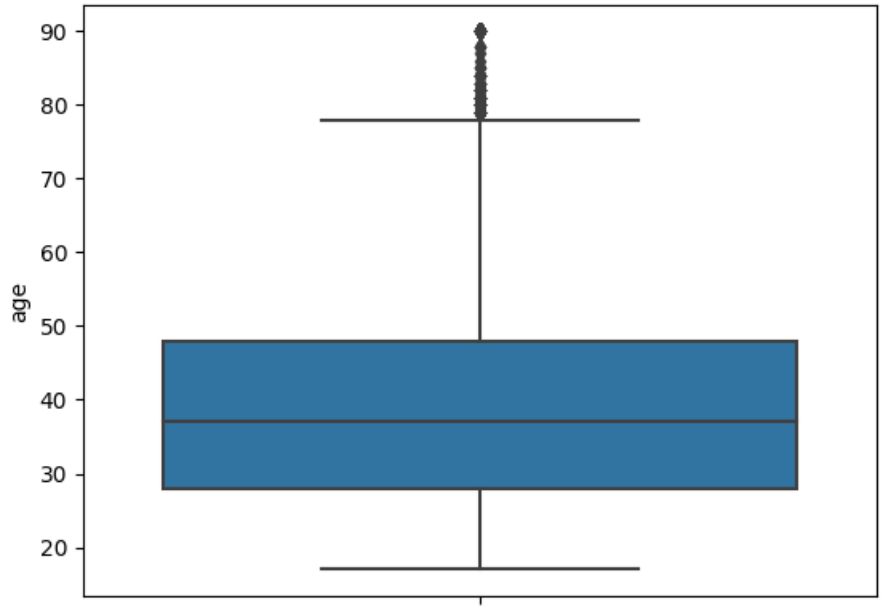

Рисунок 4. Диаграмма размаха возраста

Исходя из диаграммы, большинство респондентов находятся в возрасте от 30 до 50 лет. Это отражает возрастной диапазон наиболее экономически активной части населения.

С помощью классификатора Random Forest определим 20 наиболее значимых признаков для задачи классификации и построим матрицу корреляции по выделенным признакам. Результат представлен на рисунке 5.

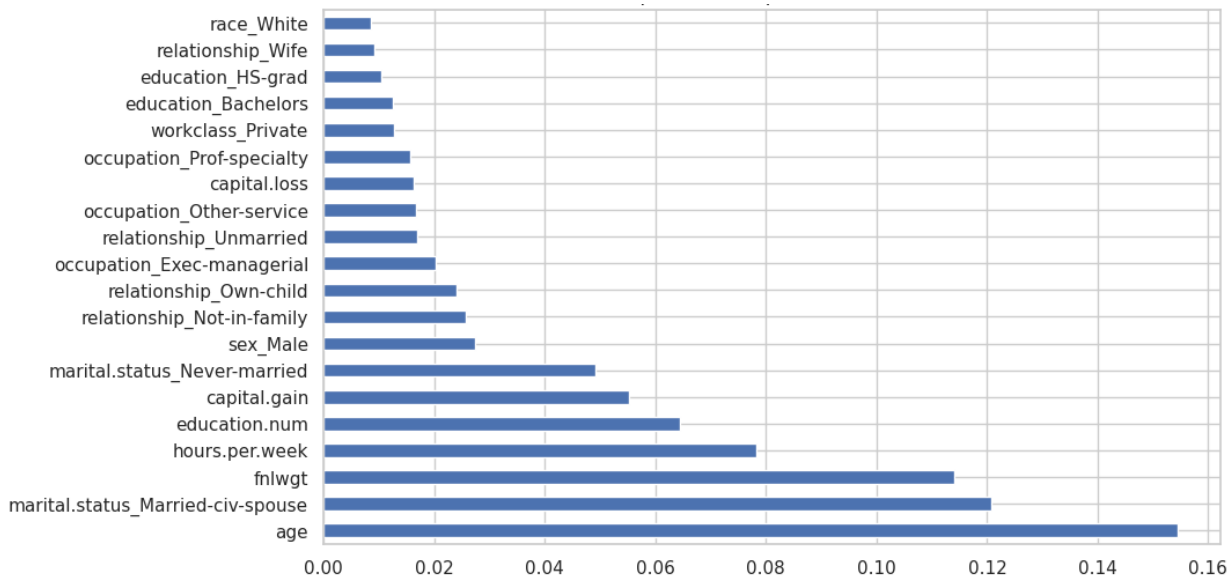

Рисунок 5. 20 наиболее значимых признаков

Матрица корреляции для выделенных признаков представлена на рисунке 6.

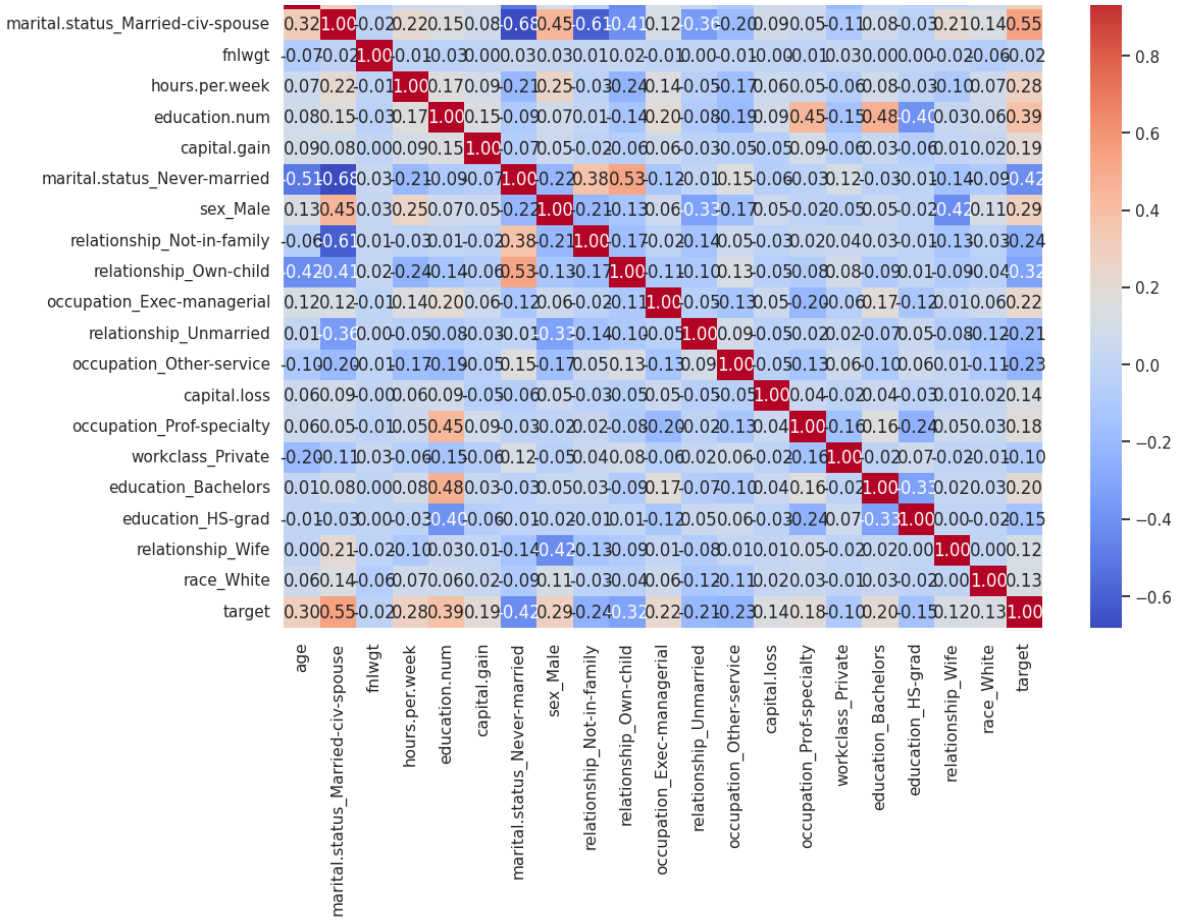

Рисунок 6. Матрица корреляции для 20 наиболее значимых признаков

Корреляционный анализ:

Семейное положение:

- marital-status_Married-civ-spouse имеет высокую положительную корреляцию с целевой переменной (0.54), что указывает на то, что состоящие в браке люди чаще относятся к группе с доходом больше >50k.
- Напротив, marital-status_Never-married показывает отрицательную корреляцию (−0.41), что свидетельствует о меньшей вероятности высокой заработной платы у людей, не состоящих в браке.

Возраст и рабочие часы:

- age и hours-per-week демонстрируют умеренную положительную корреляцию (0.28 и 0.27 соответственно). Это указывает на то, что с увеличением возраста и количества отработанных часов вероятность дохода >50k также возрастает.

Образование:

- education-num имеет значительную положительную корреляцию (0.39), что говорит о связи более высокого уровня образования с вероятностью заработка выше 50k.
- education_Bachelors также показывает положительную корреляцию (0.20), подтверждая влияние образования на уровень дохода.

Пол:

- sex_Male коррелирует положительно (0.29), указывая на более высокую вероятность принадлежности мужчин к группе с доходом >50k.

Решение задачи классификации

Для решения задачи классификации были использованы следующие модели:

1. Случайный лес (Random Forest Classifier):

Данный метод ансамблевого обучения строит несколько деревьев решений во время обучения и объединяет их результаты для повышения точности и контроля над переобучением. Он работает за счёт создания "леса" деревьев на основе случайных подвыборок данных и признаков, что делает его устойчивым к шумам и подходящим для работы с большим количеством признаков. Предоставленная моделью оценка важности признаков была использована для более точного корреляционного анализа.

2. Метод опорных векторов (C-Support Vector Classification):

C-SVC – разновидность метода опорных векторов (SVM), который находит оптимальную гиперплоскость для разделения классов в пространстве признаков. Параметр “C” управляет балансом между минимизацией ошибки на обучающих данных и предотвращением переобучения (регуляризация). Этот метод особенно эффективен для задач классификации в пространствах с высокой размерностью и может обрабатывать нелинейные границы решений с помощью ядерных функций.

3. Логистическая регрессия (Logistic Regression):

Логистическая регрессия – это статистическая модель для бинарной классификации, которая предсказывает вероятность принадлежности к классу на основе одного или нескольких предикторов. Она моделирует зависимость между целевой переменной (классом) и признаками с помощью логистической функции. Несмотря на название, это алгоритм классификации, а не регрессии. Логистическая регрессия часто используется из-за своей простоты, интерпретируемости и эффективности для линейно разделимых данных.

4. XGBoost (Extreme Gradient Boosting):

XGBoost – это оптимизированный алгоритм градиентного бустинга, известный своей высокой производительностью и эффективностью. Он строит ансамбль деревьев решений последовательно, где каждое новое дерево исправляет ошибки предыдущих деревьев. XGBoost включает механизмы регуляризации для предотвращения переобучения и может использоваться как для задач регрессии, так и для классификации. Благодаря своей скорости и точности он часто используется в соревнованиях по машинному обучению.

Полученные метрики точности (accuracy) для каждого метода представлены в таблице 1. График метрик представлен на рисунке 7.

Таблица 1. Метрики точности примененных методов

Метод	**Значение Accuracy**
Random Forest Classifier	0.8672
C-Support Vector Classification	0.6282
Logistic Regression	0.8223
Extreme Gradient Boosting	0.8725

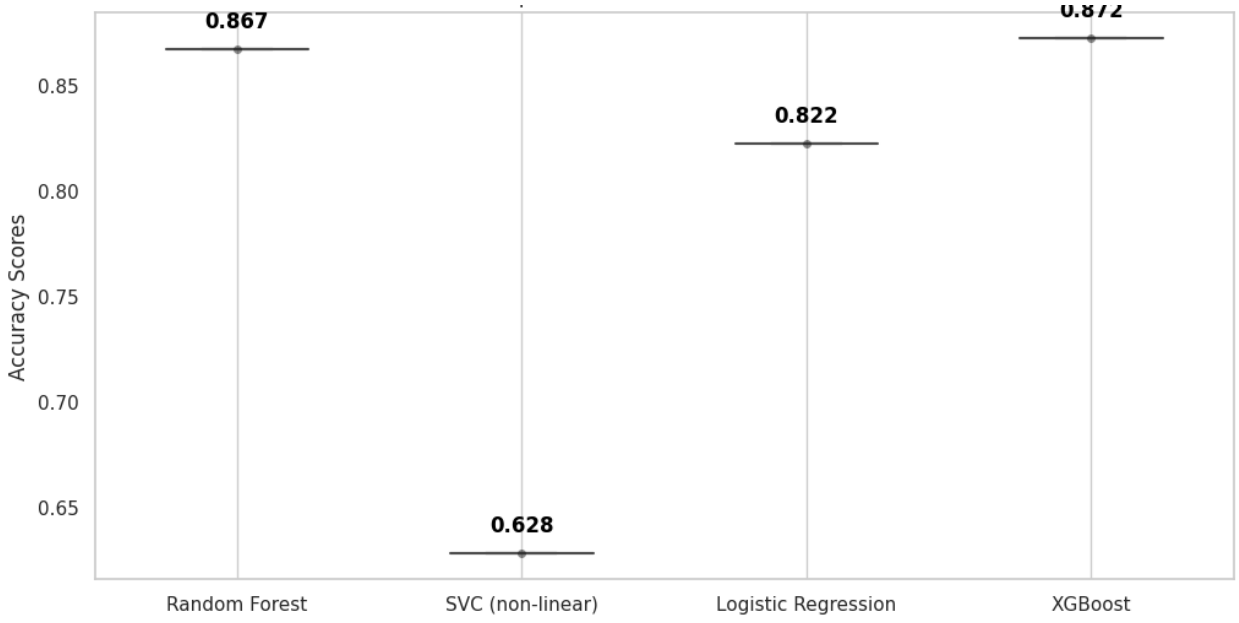

Рисунок 7. Визуализация значений accuracy

Анализ результатов классификации показал, что наибольшую точность (accuracy) продемонстрировала модель Extreme Gradient Boosting (XGBoost) с показателем 87.25%, что подтверждает её высокую эффективность и способность к решению задачи классификации. Random Forest Classifier также показал высокую точность в 86.72%, что также делает его надёжным и интерпретируемым инструментом для анализа данных. Хуже всех справился метод опорных векторов (accuracy = 62.82%), такой результат может быть связан с использованием ядра “rbf” и необходимостью дополнительной настройки гиперпараметров модели.

Заключение

В рамках научно-исследовательской работы был проведен анализ данных переписи населения и решена задача классификации с помощью 4 методов машинного обучения. На основе проведённого разведочного анализа данных и построения графических визуализаций были выявлены ключевые зависимости между признаками и целевой переменной.

Анализ данных показал, что уровень дохода тесно связан с такими факторами, как семейное положение, возраст, уровень образования, пол и количество отработанных часов.

Для решения задачи классификации были применены четыре алгоритма машинного обучения: Random Forest Classifier, C-Support Vector Classification (SVC), Logistic Regression и Extreme Gradient Boosting (XGBoost). Модели XGBoost и Random Forest являются наиболее эффективными для исследуемого набора данных (accuracy = 82.25% и 86.72%, соответственно), модель SVC продемонстрировала наименьшую точность (62.82%), что указывает на необходимость дальнейшей доработки данной модели для повышения её точности.

Список использованных источников

• Adult Census Income Data Set [Электронный ресурс] // Kaggle. URL: https://www.kaggle.com/datasets/uciml/adult-census-income/data (дата обращения: 25.12.2024).

• Scikit-learn: машинное обучение на Python [Электронный ресурс] // Scikit-learn. URL: https://scikit-learn.org/stable/ (дата обращения: 25.12.2024).

• XGBoost: Scalable Tree Boosting System [Электронный ресурс] // XGBoost. URL: https://xgboost.readthedocs.io/en/latest/ (дата обращения: 25.12.2024).

• A Foundational Library for Data Analysis and Statistics [Электронный ресурс] // Pandas. URL: https://pandas.pydata.org/ (дата обращения: 25.12.2024).

• Array programming with NumPy [Электронный ресурс] // NumPy. URL: https://numpy.org/ (дата обращения: 25.12.2024).

• A 2D Graphics Environment [Электронный ресурс] // Matplotlib. URL: https://matplotlib.org/stable/contents.html (дата обращения: 25.12.2024).