|  |
| --- |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ |
| ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ  высшего образования |
| **«Национальный исследовательский ядерный университет «МИФИ»**  **(НИЯУ МИФИ)** |

**Отчет**

**по результатам выполнения задания**

**демонстрационного экзамена**

|  |  |
| --- | --- |
| Студент | Кулик Михаил Ростиславович |
| Группа | Б20-201 |
|  |  |
| Дата | 30.08.2023 |

Москва 2023

**Оглавление**

[**I Исходные данные** 3](#_Toc143075002)

[**II Предобработка данных** 4](#_Toc143075003)

[**III Построение и исследование модели машинного обучения** 5](#_Toc143075004)

[**Заключение** 6](#_Toc143075005)

# **I Исходные данные**

Данный набор данных состоит из 4-х признаков: “age”, “sex”, “profession” и “income”. Общий размер датасета 10000 строк. Входные данные содержат 479 пропущенных поля в признаках возраст (age) и пол (sex). Два признака категориальных: “sex” и “profession” (объектный тип данных object), а два числовых (вещественных float64). Категориальный признак “sex”: имеет два уникальных значения: “male” и “female”. Наиболее часто встречаемое значение это “female” (freq. 5413). Категориальный признак “profession” имеет 3 уникальных значения: “worker”, “unemployed” и “homewife”. Наиболее часто встречаемое значение это “worker” (freq. 4298). Средний возраст 35 лет, минимальный 18 лет, максимальный 55, среднеквадратичное отклонение 5. Средняя зарплата 77494.4 , минимальная 46087, максимальная 20000, среднеквадратичное отклонение 28965,8. Пропуски удаляем. По этим данным решается задача кластеризации. Ниже представлены графики распределений и различные статистические визуализации данных (требуемые по заданию).

Изображение выглядит как линия, График, диаграмма, снимок экрана

Автоматически созданное описание

Рисунок 1 – Гистограмма вместе с распределением признака"age"

Изображение выглядит как текст, число, линия, снимок экрана

Автоматически созданное описание

Рисунок 2 - Box and Whisker для признака "age" сгруппированного с признаком “sex”

Изображение выглядит как текст, диаграмма, линия, число

Автоматически созданное описание

Рисунок 3 - Box and Whisker для признака "age" сгруппированного с признаком “profession”

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Рисунок 4 - Гистограмма распределения признака “sex”.

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Рисунок 5 - Гистограмма распределения признака “profession”.

Как итог: в исходном дата сете классы являются не равномощными, есть пропуски вы данных и выбросы, а также дубликаты.

# **II Предобработка данных**

2.1. Очистка данных:

В данных были обнаружены выбросы, дубликаты и пропуски. Согласно тексту задания пропуски были устранены путем удаления соответствующей строки, а дубликаты и выбросы были оставлены (по совету организатора и наилучшего соответствия с ответом теста). Возможные варианты устранения дубликатов это: удаление или замены их средним или медианным значением. Возможные варианты устранения выбросов: удаление выбросов (робастный анализ) или выбрасывание только заквантильных значений. Ниже приведен пример кода:

Изображение выглядит как текст, Мультимедийное программное обеспечение, программное обеспечение, снимок экрана

Автоматически созданное описание

Рисунок 6 - Пример кода квантильного теста.

Ниже приведены корреляционная матрица признаков. Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Рисунок 7 - Корреляционная матрица признаков

Видна корреляциям между признаками profession и income.

# **III Построение и исследование модели машинного обучения**

Для данной задачи использовалась модель kmeans на 3 кластера. Произведено разделение данных на тест и трайн (в пропорции согласно тексту задания). Использовался One-hot-encoding.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 8 - Точечная диаграмма основных компонент.

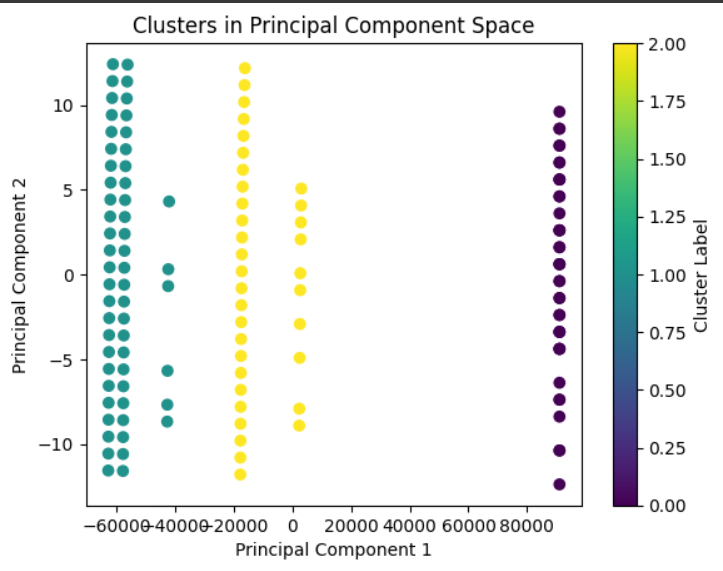
**

Рисунок 9 - Разделение на кластеры

Варианты улучшения базовой модели. Было произведена замена выбросов на среднее значение, а пропуски на случайное значение (в признаке “sex” случайный выбор между male и female). Также был использован автоматический подпор параметров и реализован DBSCAN подход.

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 10 - Подбор гиперпараметров

# ***Заключение***

*Полученное значение коэффициента силуэта 0,86. После улучшения модели удалось получить значение силуэта в 0.89 и 0.99 при различных случайных наборах male-female.*