|  |  |  |
| --- | --- | --- |
| Лабораторная работа № 2.  Разведочный анализ данных | Студент | Бокова О. Д. |
| Группа | ИВТ-363 |
| Преподаватель | Фокин Р.О. |
| Оценка |  |
| Дата | 6.03.2023 |

Для выбранного или произвольного набора данных из репозитория

UC Irvine Machine Learning Repository необходимо выполнить следующие

вычисления:

1. найти среднее значение, медиану, экстремумы любого параметра;

2. произвести предобработку данных (удалить шумы, дубликаты)

3. найти дисперсию и стандартное отклонение любого параметра;

4. найти линейную зависимость между двумя любыми параметрами;

5. построить диаграммы рассеяния;

6. построить гистограмму, определить тип распределения одного из

параметров;

7. подвести итог (summary).

Код:

# Загрузка необходимых библиотек и датасета

import pandas as pd

import numpy as np

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

df = pd.read\_csv("biofam.csv", delimiter = ';')

# 1. Найти среднее значение, медиану, экстремумы любого параметра

mean\_age = df['birthyr'].mean()

median\_age = df['birthyr'].median()

min\_age = df['birthyr'].min()

max\_age = df['birthyr'].max()

print('Средний возраст:', mean\_age)

print('Медиана возраст:', median\_age)

print('Минимальный возраст:', min\_age)

print('Максимальный возраст:', max\_age)

# 2. Произвести предобработку данных (удалить шумы, дубликаты)

# проверка на наличие дубликатов

print('Количество дубликатов:', df.duplicated().sum())

# проверка выбросов на графике boxplot

plt.boxplot(df['birthyr'], vert=False)

plt.show()

# 3. Найти дисперсию и стандартное отклонение любого параметра

var\_age = df['birthyr'].var()

std\_age = df['birthyr'].std()

print('Дисперсия:', var\_age)

print('Стандартное отклонение:', std\_age)

# 4. Найти линейную зависимость между двумя любыми параметрами

corr\_coef = df['a16'].corr(df['a30'])

print('Коэффициент корреляции:', corr\_coef)

# 5. Построить диаграммы рассеяния

plt.scatter(df['a16'], df['a30'])

plt.show()

# 6. Построить гистограмму, определить тип распределения одного из параметров

plt.hist(df['a30'])

plt.show()

Результаты:

