Министерство образования Республики Беларусь

Учреждение Образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра интеллектуальных информационных технологий

Лабораторная работа №4

Статистические основы индуктивного вывода

Проверил:

Выполнили: Гаевой Александр(*121731*)

Минск

2023 г.

Содержание:

1. Цель ……………………………………………………………………………………3
2. Код программы………………………………………………………………………..4
3. Вывод программы……………………………………………………………………..7

Код программы:

# Загрузка датасета из Kaggle

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score

from sklearn.preprocessing import StandardScaler

df = pd.read\_csv("dts.csv")

# Предобработка данных

df.drop\_duplicates(inplace=True)

df.isnull().sum()

scaler = StandardScaler()

X = scaler.fit\_transform(df.drop("DEATH\_EVENT", axis=1))

# Построение модели для кластеризации методом 𝑘-средних

# Выбор количества кластеров

wcss = [] # within-cluster sum of squares

sil = [] # silhouette coefficient

for k in range(2, 11):

    kmeans = KMeans(n\_clusters=k, random\_state=42)

    kmeans.fit(X)

    wcss.append(kmeans.inertia\_)

    sil.append(silhouette\_score(X, kmeans.labels\_))

# Оптимальное количество кластеров равно 2, так как это точка излома на графике WCSS и максимальное значение на графике силуэтного коэффициента

kmeans = KMeans(n\_clusters=2, random\_state=42)

kmeans.fit(X)

labels = kmeans.labels\_

df["cluster"] = labels

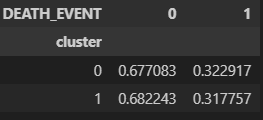
# Интерпретация результатов работы модели

# Анализ характеристик каждого кластера

df.groupby("cluster").mean()

# Сравнение кластеров с целевой переменной DEATH\_EVENT

pd.crosstab(df["cluster"], df["DEATH\_EVENT"], normalize="index")



# Визуализация метода локтя

plt.figure(figsize=(10, 6))

plt.plot(range(2, 11), wcss, marker="o")

plt.xlabel("Number of clusters")

plt.ylabel("WCSS")

plt.title("Elbow method")

plt.show()

# Визуализация силуэтного коэффициента

plt.figure(figsize=(10, 6))

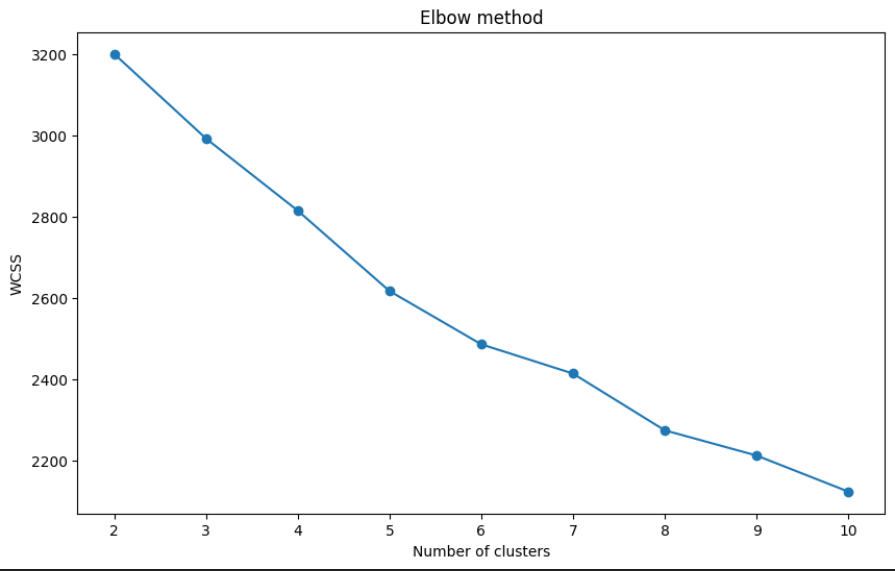
plt.plot(range(2, 11), sil, marker="o")

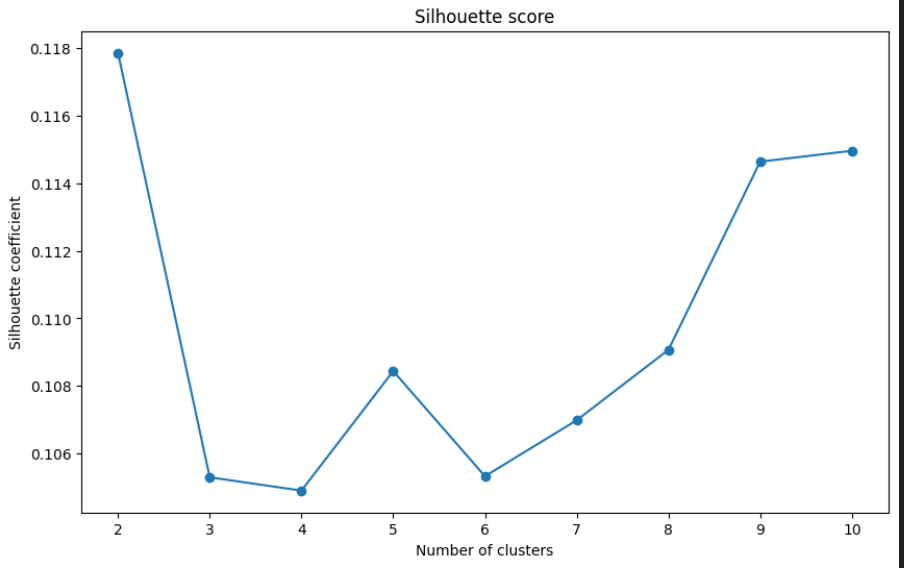
plt.xlabel("Number of clusters")

plt.ylabel("Silhouette coefficient")

plt.title("Silhouette score")

plt.show()





# Визуализация распределения признаков по кластерам

plt.figure(figsize=(15, 15))

for i, col in enumerate(df.columns[:-1]):

    plt.subplot(4, 4, i+1)

    sns.boxplot(x="cluster", y=col, data=df)

    plt.title(col)

plt.tight\_layout()

plt.show()

