Министерство образования Республики Беларусь

Учреждение образования

«Брестский Государственный технический университет»

Кафедра ИИТ

Лабораторная работа №1

По дисциплине: «Основы машинного обучения»

Тема: «Знакомство с анализом данных: предварительная обработка и визуализация»

Выполнил:

Студент 3 курса Группы АС-65

Лопато А.В.

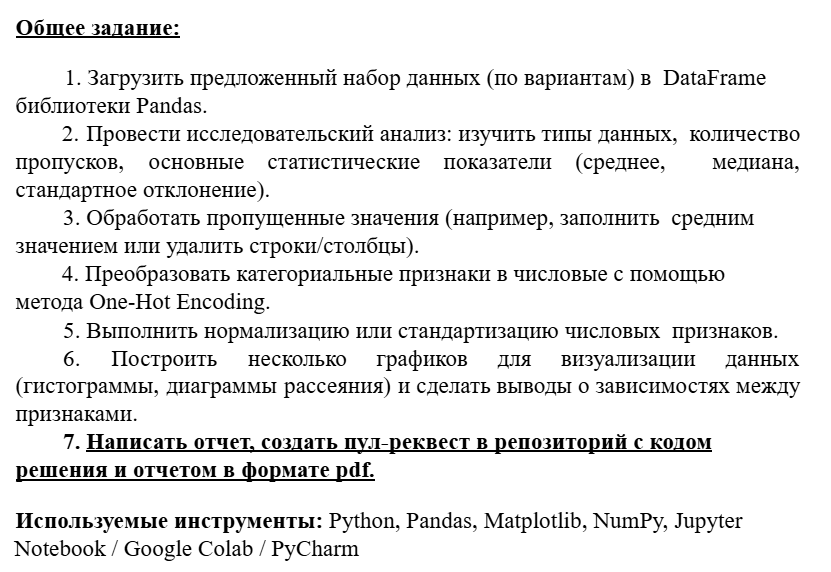
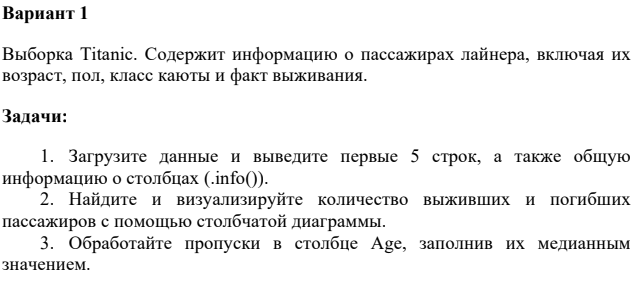
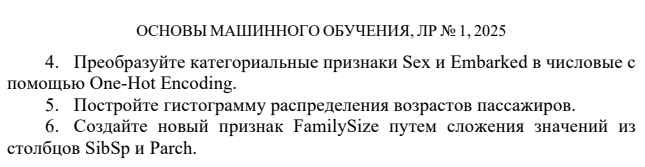
Проверил:

Крощенко А. А.

Брест 2025

Цель работы: получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Ход работы

Код программы:

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

# Загрузка данных

df = pd.read\_csv("Titanic-Dataset.csv")

# Просмотр данных

print("Первые 5 записей:")

print(df.head())

print("\nСведения о данных:")

print(df.info())

print("\nРаспределение по выживаемости:")

survival\_counts = df['Survived'].value\_counts()

print(survival\_counts)

# Визуализация выживаемости

plt.figure(figsize=(8, 5))

survival\_counts.plot(kind='bar', color=['lightcoral', 'lightgreen'])

plt.title("Распределение пассажиров по выживаемости")

plt.xlabel("Выжил (1) / Не выжил (0)")

plt.ylabel("Число пассажиров")

plt.xticks(rotation=0)

plt.show()

# Обработка пропущенных значений в возрасте

print(f"\nПропущенных значений в возрасте до обработки: {df['Age'].isna().sum()}")

median\_age = df['Age'].median()

df['Age'] = df['Age'].fillna(median\_age)

print(f"Пропущенных значений в возрасте после обработки: {df['Age'].isna().sum()}")

# Преобразование категориальных переменных

categorical\_cols = ['Sex', 'Embarked']

df = pd.get\_dummies(df, columns=categorical\_cols, drop\_first=True)

print("\nДанные после преобразования категориальных переменных:")

print(df.head())

# Распределение возраста

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

plt.hist(df['Age'], bins=15, color='lightblue', edgecolor='navy', alpha=0.7)

plt.title("Распределение возраста пассажиров")

plt.xlabel("Возраст")

plt.ylabel("Частота")

plt.grid(axis='y', alpha=0.3)

plt.show()

# Создание нового признака

df['FamilySize'] = df['SibSp'] + df['Parch'] + 1  # +1 для учета самого пассажира

print("\nПример данных с новым признаком размера семьи:")

print(df[['SibSp', 'Parch', 'FamilySize']].head(8))

Графики:

Изображение выглядит как текст, снимок экрана, дисплей, Прямоугольник

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как снимок экрана, текст, Шрифт

Содержимое, созданное искусственным интеллектом, может быть неверным. Изображение выглядит как текст, меню, снимок экрана, Шрифт

Содержимое, созданное искусственным интеллектом, может быть неверным. Изображение выглядит как текст, снимок экрана, Шрифт

Содержимое, созданное искусственным интеллектом, может быть неверным.  
Изображение выглядит как снимок экрана, текст

Содержимое, созданное искусственным интеллектом, может быть неверным.  
  
Вывод: в результате выполнения данной лабораторной работы получили практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации.