**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ**

**Московский государственный технический университет**

**им. Н.Э. Баумана**

**(МГТУ им. Н.Э. Баумана)**

**Кафедра «Информационная безопасность» (ИУ8)**

Лабораторная работа № 4

По дисциплине: «Машинное обучение»

Тема: «Предобработка данных»

Выполнил: Веденеев А.А.,

Студент группы ИУ8-92

Проверила: Коннова Н.С.,

Преподаватель каф. ИУ8

г. Москва, 2024 г.

# Практическая часть лабораторной работы 4

**Цель работы:** познакомиться с основными задачами и приемами предварительного анализа и обработки данных для целей машинного обучения.

Предварительная обработка данных является неотъемлемым этапом машинного обучения, поскольку качество данных и полезная информация, которую можно извлечь из них, напрямую влияют на способность нашей модели к обучению; поэтому чрезвычайно важно, чтобы мы предварительно обработали наши данные, прежде чем вводить их в нашу модель.

**Ход работы**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph with blue squares

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph with blue and orange squares

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph with blue and black lines

Description automatically generated**

**A graph with blue squares

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph of different sizes and colors

Description automatically generated with medium confidence**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A group of graphs with different colors

Description automatically generated with medium confidence**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

A screenshot of a graph

Description automatically generated

## Вывод

## В ходе анализа датасета Titanic было выявлено несколько ключевых закономерностей, связанных с вероятностью выживания пассажиров.

## Женщины имели значительно больше шансов на выживание по сравнению с мужчинами, что объясняется приоритетной эвакуацией женщин и детей на шлюпки. Также было замечено, что пассажиры с более дорогими билетами имели выше шансы на выживание, вероятно, из-за лучших условий обслуживания и преимуществ при эвакуации. Кроме того, существовала связь с портом отправления Саутгемптон: пассажиры, отправлявшиеся из этого порта, обладали немного более высокой вероятностью выживания по сравнению с другими портами, такими как Куинстаун и Йорк.

**Контрольные вопросы**

1. Какие основные виды визуализации вы знаете? Какие у них области применения?

* Гистограммы (Histograms). Позволяют анализировать распределение переменных и частоты значений. Применяются для изучения распределения числовых данных и выявления выбросов.
* Диаграммы разброса (Scatter Plots). Используются для выявления зависимости между двумя переменными и определения их корреляции.
* Линейные графики (Line Charts). Применяются для анализа изменений данных во времени, например, тренды и динамику во временных рядах.
* Круговые диаграммы (Pie Charts). Наиболее удобны для отображения пропорций или долей данных. Часто используются в задачах, где нужно показать процентное соотношение категорий.
* Столбчатые диаграммы (Bar Charts). Позволяют сравнивать категории данных по величине. Применяются для сравнения значений между группами данных.
* Ящико-боксы (Box Plots). Отображают распределение данных, медиану, межквартильный диапазон и выбросы. Применяются для статистического анализа и проверки гипотез.
* Тепловые карты (Heatmaps). Используются для отображения матричных данных и выявления закономерностей между переменными. Например, в анализе корреляции или визуализации таблиц данных.

1. Какие типы визуализации больше всего подходят для анализа совместного распределения двух непрерывных переменных?

* Scatter Plot показывает корреляцию и тренды между переменными.
* Density Plots демонстрируют вероятность распределения данных.
* Contour Plots визуализируют зоны высокой плотности данных.

1. Какие типы визуализации больше всего подходят для анализа совместного распределения двух дискретных переменных?

* Столбчатые диаграммы сравнивают частоты категорий переменных.
* Кросс-таблицы (Cross-tabulations): отображают частотное распределение двух переменных.
* Heatmaps визуализируют частоты пар категорий с помощью цветовой шкалы.
* Диаграммы точек показывают распределение пар значений и их частоты.

1. Как лучше всего построить совместное распределение дискретной и непрерывной переменной?

* Box Plot позволяет показать распределение непрерывной переменной для каждой категории дискретной переменной.
* Bar + Density Plot позволяют сравнить средние значения и распределения.
* График точек (Violin Plot) комбинирует элементы ящика с усами и плотности данных, показывая распределение непрерывной переменной по группам дискретной переменной.
* Круговые диаграммы и разреженные плотности визуализируют группы дискретных категорий и их соответствующие распределения.

1. Как лучше всего построить совместное распределение двух непрерывных и одной дискретной переменной?

* Графики “Violin Plot” с разбивкой по категориям позволяют сравнить распределение двух непрерывных переменных по группам, определяемым дискретной переменной.
* Параллельные координаты (Parallel Coordinates) позволяют визуализировать многомерные зависимости между переменными.

1. Как лучше всего построить совместное распределение двух дискретных и одной непрерывной переменной?

* Box Plot позволяет показать распределение непрерывной переменной по каждой комбинации категорий двух дискретных переменных.
* Графики с наложением плотности (Density/Violin Plots) позволяют визуализировать, как непрерывная переменная распределяется по группам, определяемым двумя дискретными переменными.
* Столбчатые диаграммы с наложением позволяют визуально сравнить распределения непрерывной переменной между группами двух категориальных переменных.