

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

**"МИРЭА - Российский технологический университет"**

**РТУ МИРЭА**

**Институт** Информационных технологий

**Кафедра** Вычислительной техники

**ОТЧЁТ ПО ЛАБОРАТОРНЫМ РАБОТАМ**

**ПО ДИСЦИПЛИНЕ**

**«ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»**

Выполнил студент группы:

ИКБО-05-16 Артемьев Виталий Андреевич

Принял лабораторные работы

Старший преподаватель Елизавета Александровна Зикеева

|  |  |  |
| --- | --- | --- |
|  |  |  |

Работа выполнена 12.02.2020 г.

«Зачтено»

Москва 2020

# Лабораторная работа 1

## Задание:

Выполнить набор базовых задач для ознакомления с языком программирования Python, вариант №1.

## Выполнение:

Программную реализацию с демонстрацией работы можно найти по адресу https://github.com/VitalyArtemiev/Datamine-labs/blob/master/Python%20showcase.ipynb

# Лабораторная работа 2

## Задание:

Ознакомиться с библиотеками Pandas и Numpy для ввода данных и операций над ними соответственно.

## Задание по Numpy:

1. Сгенерируйте матрицу, состоящую из 1000 строк и 50 столбцов, элементы которой являются случайными из нормального распределения N(1,100).

2. Произведите нормировку матрицы из предыдущего задания: вычтите из каждого столбца его среднее значение, а затем поделите на его стандартное отклонение.

3. Выведите для заданной матрицы номера строк, сумма элементов в которых превосходит 10.

4. Сгенерируйте две единичные матрицы (т.е. с единицами на диагонали) размера 3x3. Соедините две матрицы в одну размера 6x3.

## Выполнение:

Программную реализацию с демонстрацией работы можно найти по адресу https://github.com/VitalyArtemiev/Datamine-labs/blob/master/Numpy%20showcase.ipynb

## Задание по Pandas:

Загрузите датасет titanic.csv и, используя описанные выше способы работы с данными, найдите ответы на вопросы:

1. Какое количество мужчин и женщин ехало на корабле? В качестве ответа приведите два числа через пробел.

2. Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров. Ответ приведите в процентах (число в интервале от 0 до 100, знак процента не нужен), округлив до двух знаков.

3. Какую долю пассажиры первого класса составляли среди всех пассажиров? Ответ приведите в процентах (число в интервале от 0 до 100, знак процента не нужен), округлив до двух знаков.

4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров. Посчитайте среднее и медиану возраста пассажиров. В качестве ответа приведите два числа через пробел.

5. Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.

6. Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name). Это задание — типичный пример того, с чем сталкивается специалист по анализу данных. Данные очень разнородные и шумные, но из них требуется извлечь необходимую информацию. Попробуйте вручную разобрать несколько значений столбца Name и выработать правило для извлечения имен, а также разделения их на женские и мужские.

## Выполнение:

Программную реализацию с демонстрацией работы можно найти по адресу https://github.com/VitalyArtemiev/Datamine-labs/blob/master/Pandas%20showcase.ipynb

# Лабораторная работа 3

## Задание:

Ознакомиться с работой решающих деревьев с использованием набора инструментов sklearn.

## Инструкция по выполнению:

1. Загрузите выборку из файла titanic.csv с помощью пакета Pandas.

2. Оставьте в выборке четыре признака: класс пассажира (Pclass), цену билета (Fare), возраст пассажира (Age) и его пол (Sex).

3. Обратите внимание, что признак Sex имеет строковые значения.

4. Выделите целевую переменную — она записана в столбце Survived.

5. В данных есть пропущенные значения — например, для некоторых пассажиров неизвестен их возраст. Такие записи при чтении их в pandas принимают значение nan. Найдите все объекты, у которых есть пропущенные признаки, и удалите их из выборки.

6. Обучите решающее дерево с параметром random\_state=241 и остальными параметрами по умолчанию.

7. Вычислите важности признаков и найдите два признака с наибольшей важностью. Их названия будут ответами для данной задачи (в качестве ответа укажите названия признаков через запятую без пробелов).

## Выполнение:

Программную реализацию с демонстрацией работы можно найти по адресу https://github.com/VitalyArtemiev/Datamine-labs/blob/master/Trees.ipynb

## Вывод

Были выполнены 3 лабораторные работы, в ходе которых получены базовые навыки по анализу данных, в том числе опыт использования библиотек numpy, pandas и sklearn, а так же применения решающих деревьев.