**Московский государственный технический**

**университет им. Н.Э. Баумана**

Факультет "Информатика и системы управления"

Кафедра ИУ5 "Системы обработки информации и управления"

Технологии машинного обучения

Отчет по лабораторной работе №2

Выполнила: Пасатюк Александра, группа ИУ5-63Б

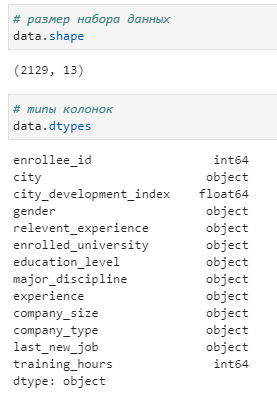
Москва, 2020

**Задание**

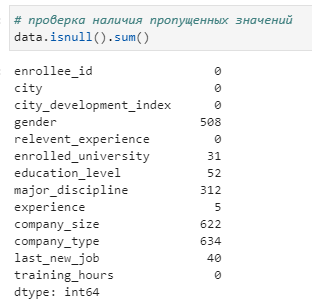
1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
   * обработку пропусков в данных;
   * кодирование категориальных признаков;
   * масштабирование данных.

**Выполнение работы**

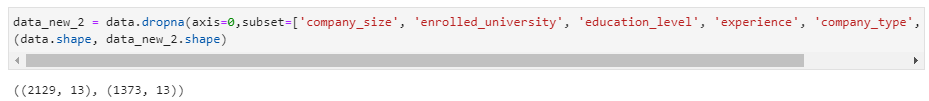
Информация о наборе данных:



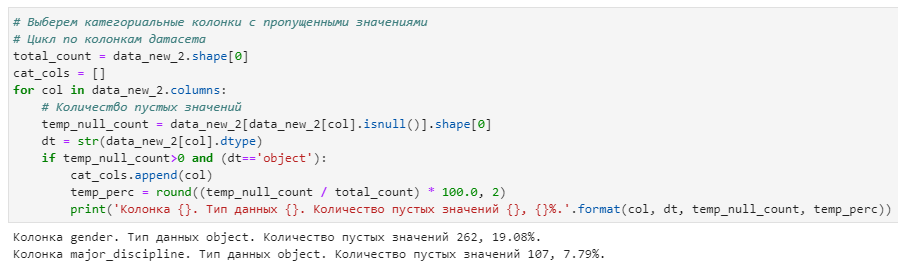
**Обработка пропусков данных**



Удалим строки, где отсутствуют значения в столбцах "company\_size", "enrolled\_university", "education\_level", "experience", "company\_type", "last\_new\_job".

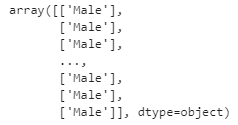
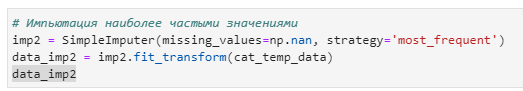


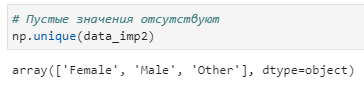
**Обработка пропусков в категориальных данных**



Выберем колонку gender для импьютации наиболее частыми значениями

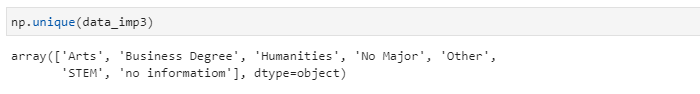




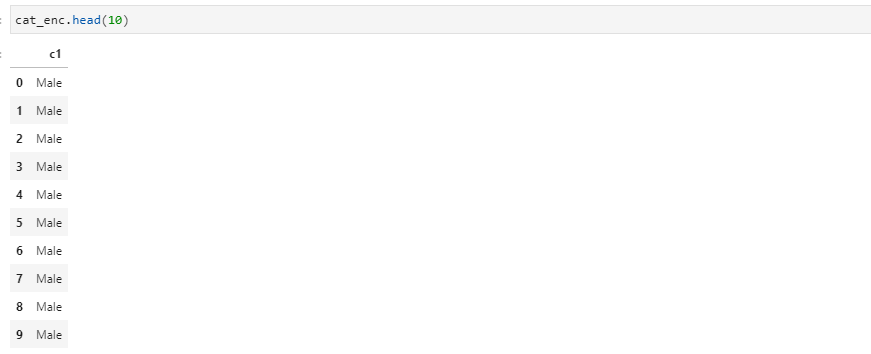
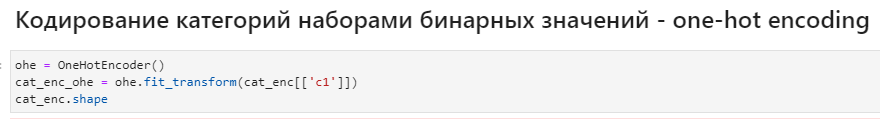
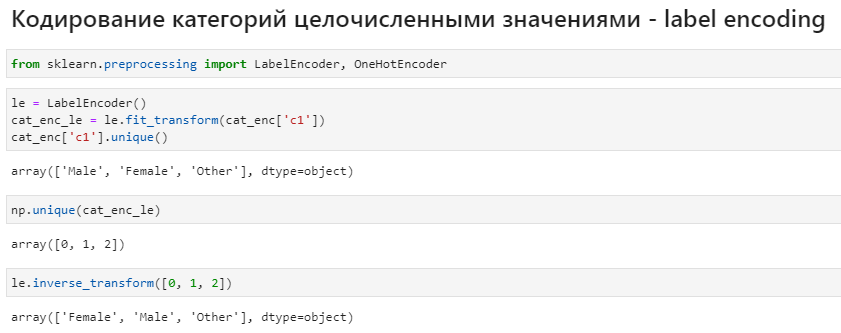
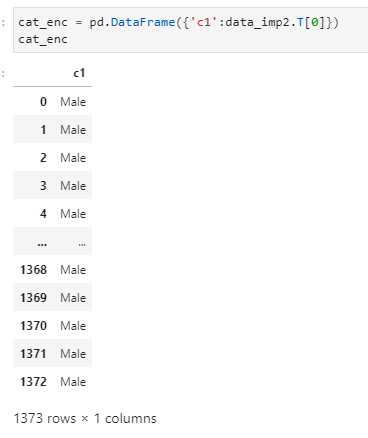


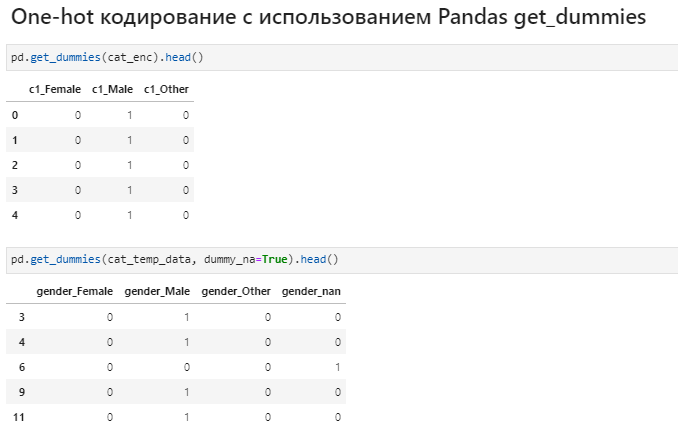
Выбираем колонку major\_discipline для импьютации константой





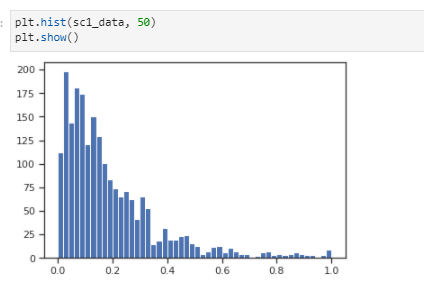
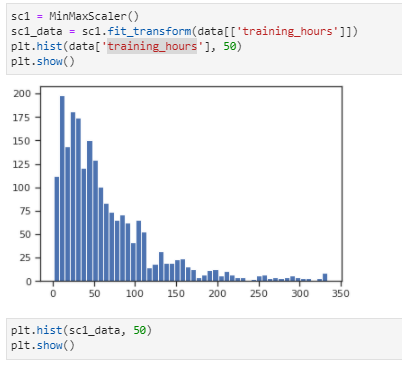
**Кодирование категориальных признаков**





**Масштабирование данных**

MinMax – масштабирование



Масштабирование данных на основе Z-оценки - StandardScaler

