# Московский авиационный институт (национальный исследовательский университет)

## Факультет информационных технологий и прикладной математики

## Кафедра вычислительной математики и программирования


**Лабораторная работа №0 по курсу «Искусственный интеллект»**
**Тема: Анализ и подготовка данных**


Студент: А. А. Чернобаев
Преподаватель: Самир Ахмед
Группа: М8О-308Б-19

Дата:
Оценка:
Подпись:


**Москва, 2022**

# Задача

**Задача:** В данной лабораторной работе, вы выступаете в роли предприимчивого начинающего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба. Поэтому тщательно работайте:) И главное, день промедления и вас опередит ваш конкурент, да и сплагиаченная работа отразится на репутации По сути в данной лабораторной работе вы выполняете часть работы BI системы. Если вы заинтересовались этим направлением, то можно будет в дальнейшем что-то придумать)

# 1 Описание

В качестве датасета я выбрал «Titanic - Machine Learning from Disaster» с сайта kaggle.
Он находится по ссылке https://www.kaggle.com/competitions/titanic/data.
В данном датасете приведены следующие признаки:

1. survival : Выживет ли пассажир (0 - нет, 1 - да)
2. pclass: Класс билета (1 = 1st, 2 = 2nd, 3 = 3rd)
3. sex: Пол пассажира
4. Age: Возраст
5. sibsp: of siblings / spouses aboard the Titanic. The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)
6. parch: of parents / children aboard the Titanic. The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson
7. ticket: номер билета
8. fare: цена билета
9. cabin: номер каюты
10. embarked: порт погрузки

# 2 Ход работы

Сначала я считал данные с помощью библиотеки pandas. Я исключил из датафрейма колонки Name и Ticket, потому что имя пассажира и номер его билета не играют никакой роли в данной задаче. Проанализировав датафрейм, я заполнил пустые значения в колонке Age, удалил колонку Cabin, потому что она содержала только уникальные значения. Далее я заполнил пустые значения в колонке Embarked. Затем я стал квантифицировать поля Sex и Embarked, для этого воспользовался LabelEncoder из sklearn.preprocessing. Ниже привожу укороченный код из ipynb.

```
df = pd.read_csv("train.csv", index_col="PassengerId")
df = df.drop(["Name", "Ticket"], axis = 1)
df["Age"] = df["Age"].fillna(df["Age"].mean())
df = df.drop(["Cabin"], axis = 1)
df["Embarked"] = df["Embarked"].fillna("S")
LE = LabelEncoder()
df['Sex'] = LE.fit_transform(df['Sex'])
df['Embarked'] = LE.fit_transform(df['Embarked'])
```

В итог я получил следующие графики:

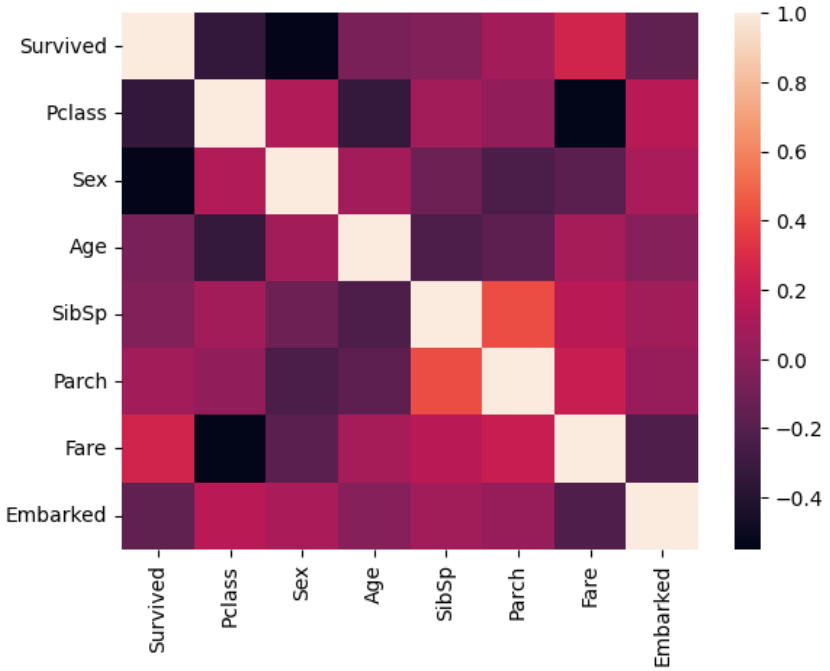


Рис. 1:

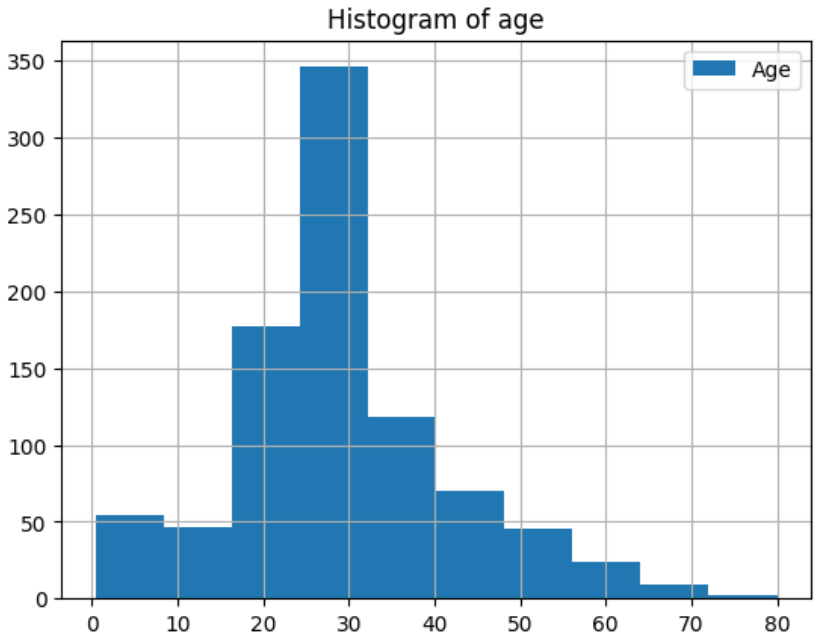
Histogram of age
Age
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80

Рис. 2:

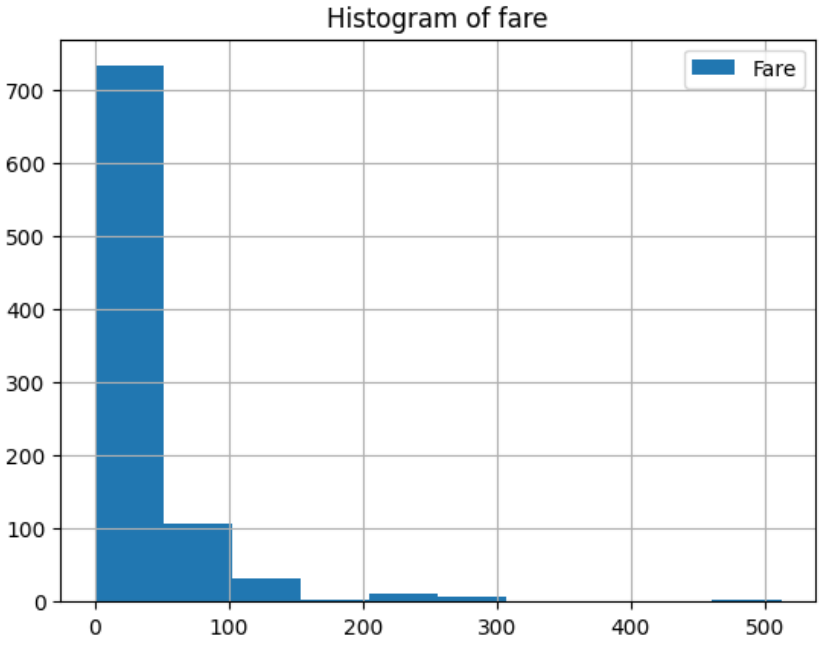
Histogram of fare
Fare
700
600
500
400
300
200
100
0
0
100
200
300
400
500

Рис. 3:

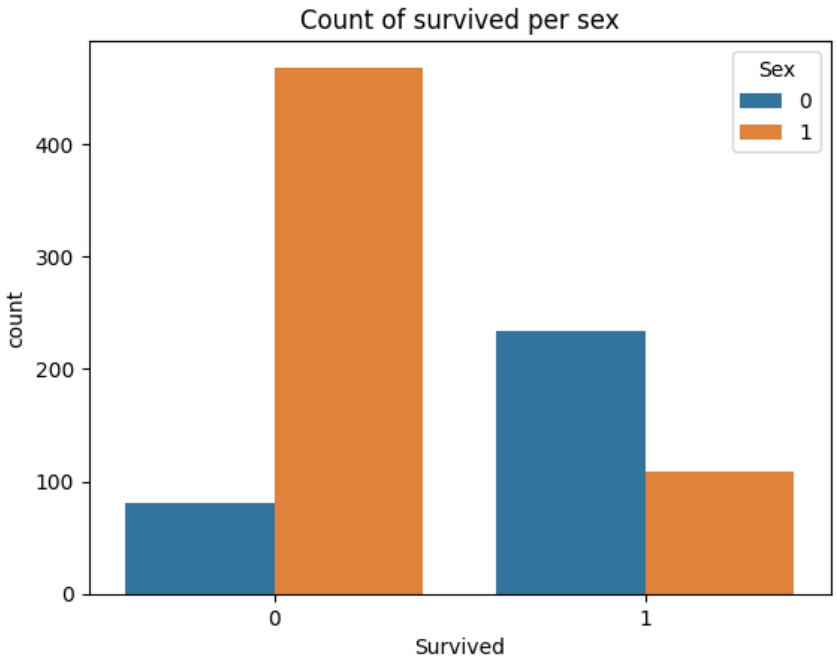
Count of survived per sex
Sex
0
1
400
300
200
100
0
count
0
1
Survived

Рис. 4:

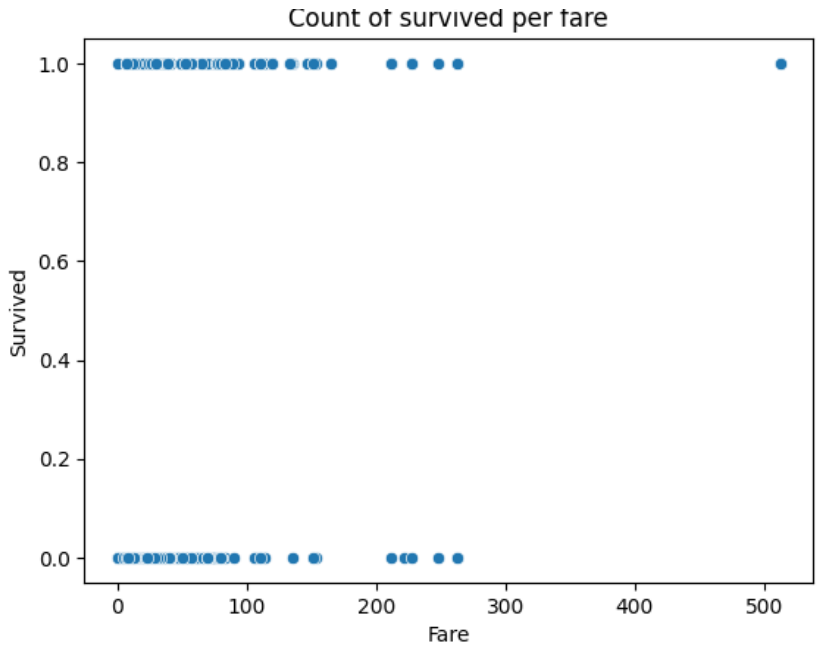
Count of survived per fare
1.0
0.8
0.6
0.4
0.2
0.0
Survived
0
100
200
300
400
500
Fare

Рис. 5:

```
df.corr()
```

| | Survived | Pclass | Sex | Age | SibSp | Parch | Fare | Embarked |
|---|---|---|---|---|---|---|---|---|
| Survived | 1.000000 | -0.338481 | -0.543351 | -0.069809 | -0.035322 | 0.081629 | 0.257307 | -0.167675 |
| Pclass | -0.338481 | 1.000000 | 0.131900 | -0.331339 | 0.083081 | 0.018443 | -0.549500 | 0.162098 |
| Sex | -0.543351 | 0.131900 | 1.000000 | 0.084153 | -0.114631 | -0.245489 | -0.182333 | 0.108262 |
| Age | -0.069809 | -0.331339 | 0.084153 | 1.000000 | -0.232625 | -0.179191 | 0.091566 | -0.026749 |
| SibSp | -0.035322 | 0.083081 | -0.114631 | -0.232625 | 1.000000 | 0.414838 | 0.159651 | 0.068230 |
| Parch | 0.081629 | 0.018443 | -0.245489 | -0.179191 | 0.414838 | 1.000000 | 0.216225 | 0.039798 |
| Fare | 0.257307 | -0.549500 | -0.182333 | 0.091566 | 0.159651 | 0.216225 | 1.000000 | -0.224719 |
| Embarked | -0.167675 | 0.162098 | 0.108262 | -0.026749 | 0.068230 | 0.039798 | -0.224719 | 1.000000 |

Рис. 6:

# 3 Выводы

В данном лабораторной работе я произвёл анализ датасета и подготовил его к машинному обучению, путём удаления ненужных признаков и перевода в числа нечисловых признаков. Работа с данными - это важный этап, потому что эффективность машинного обучения будет в том числе зависеть от того, насколько хорошо мы подготовили данные.