**Московский авиационный институт**

**(Национальный исследовательский университет)**

Институт: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

**Дисциплина: «Искусственный интеллект»**

Студент: Нелюбин В.С.

Группа: М8О-406Б-19

Преподаватель: Сошников Д.В.

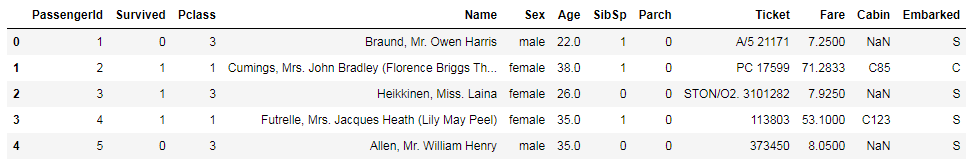
Дата: 07.12.2022

Оценка:

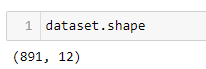
**Москва 2022**

**Загрузка датасета и первичный анализ**

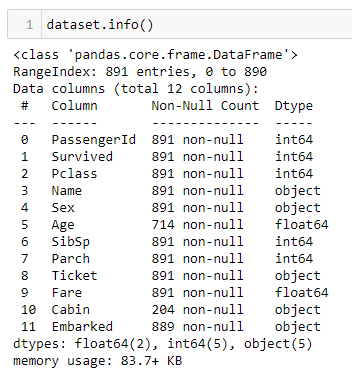
Выведем первые 5 строк датасета



Отобразим размер



Выведем общую информацию



**Описание столбцов датасата**

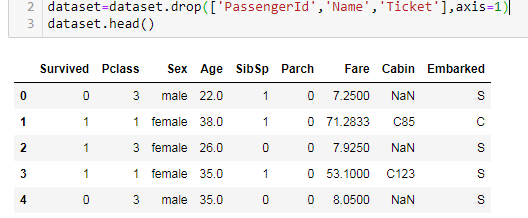
* Passengerid - это id пассажира.
* Survived - столбец, говорит о том выжил ли пассажир, принимает значение 0 или 1 . 1 если выжил , 0 если нет. Является целивой перменной, которую будем предсказывать. Только 342 человека выжило
* Pclass - фича, говорящая о том ,к какому классу в корабле относится пассажир. Может принимать значения 1,2,3. В основном люди 3 класса
* Name- имя пассажира.
* Sex - пол пассажира. Мужчин больше чем женщин
* Age - возраст пассажира. 5 самых распрастараненных возрастов 24, 22, 18, 18 , 30. В этом столбце есть пропуски.
* SibSp — содержит информацию о количестве родственников 2-го порядка (муж, жена, братья, сетры).Пропусков нет. Большинство пассажиров не имеют родственников. 209 имеют только одного родственника
* Parch-содержит информацию о количестве родственников на борту 1-го порядка (мать, отец, дети). У большинства людей нет на борту родственников первого порядка
* Ticket- код билета пассажира
* Fare - цена билета. Пропусков в данных нету. 5 самые частые цен на билет 8.0500, 3.0000, 7.8958, 7.7500, 26.0000.
* Cabin- вид каюты пассажира. Пропусков данных очень много, порядка 77%.
* Embarked- порт посадки. Есть немного пропусков. Возможные значения S,C,Q, где C — Cherbourg, Q — Queenstown, S — Southampton. Самый частый порт посадки S — Southampton

**Выводы из первичного анализа данных**

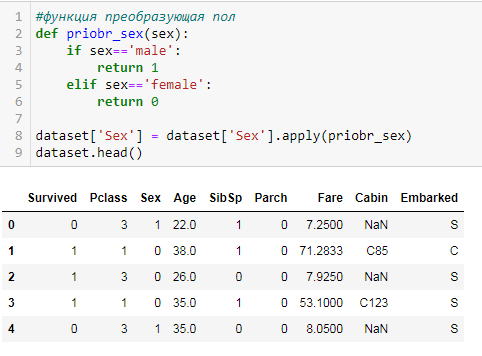
* Passengerid, Name, Ticket удаляем
* К Sex заменяем на 0 и 1
* Age в пропусках заполним через среднее.
* Cabin может быть преобразована в логическое значение
* Embarked - порт посадки пропуски заполню через самое часто встречаемое значение.

**Преобразования признаков**

Удаляем PassengerId, Name, Ticke



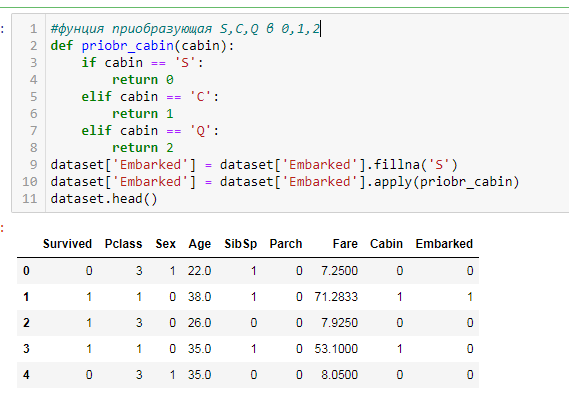
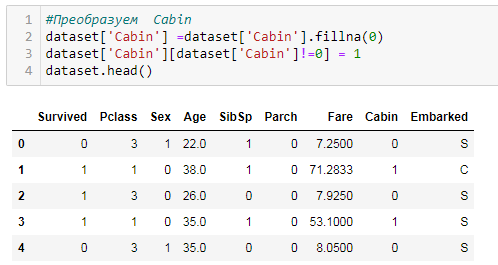
Преобразуем пол



Заполняем возраст через среднее

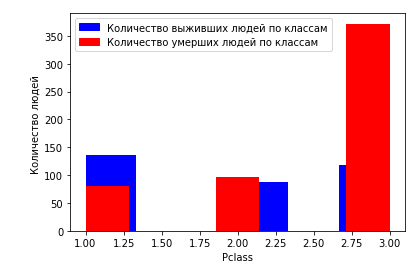


Преобразуем Cabin

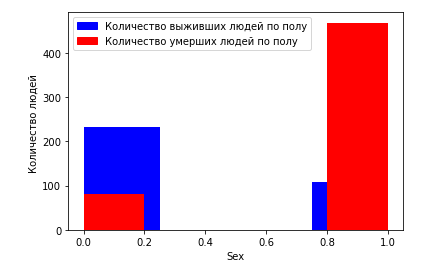


**Визуализации**

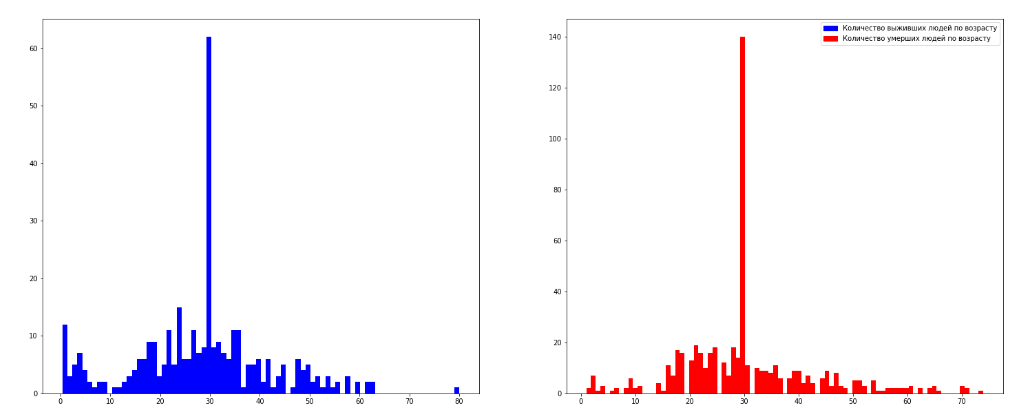
Визуализация выживших и умерших по соц. положению



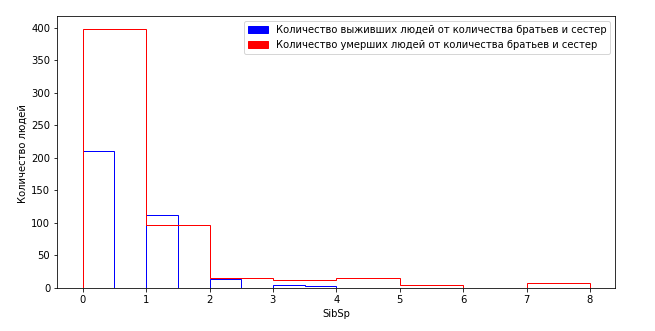
Распределение выживших и у умерших по полу



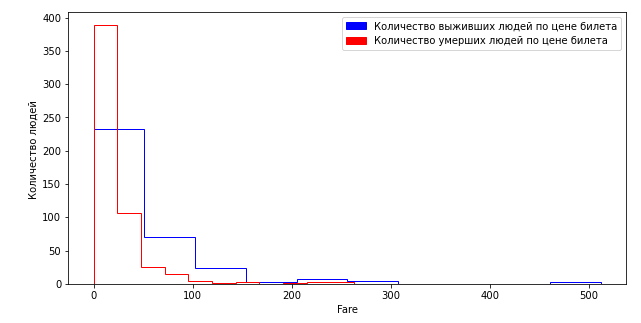
Распределение выживших и у умерших по возрастам



Распределение выживших и умерших от количества братьев и сестер



Распределение выживших и умерших по цене билета Fare



**Выводы по визуализации**

* По первому графику можно сделать вывод, что люди из 3-ого класса умирают чаше всех остальных. Выживают же чаще всех люди из первого класса. Люди из второго класса могут с примерно равной вероятностью выжить или умереть
* По второму графику можно сделать вывод,что мужчины значительно чаще умирали на титанике чем женщины.При том женщины выживали примерно в 2 раза чаще чем умирали
* По третьему графику можно сделать вывод, что количество выживших и умерших людей примерно 30 лет больше всех остальных.
* По четвертому графику можно сделать вывод, что если у пассижира нету братьв или сестер, то он с большей вероятностью погибнет
* По пятому графику можно сделать вывод,что больше всех умирают люди с билетами до примерно 100 долларов. Также можно сказать, что люди чья стоимость билетов находится около 500 долларов не умирают.