Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский Томский политехнический Университет»



Центр цифровых

образовательных технологий

01.03.02 «Прикладная математика и информатика»

Табличное представление данных

Лабораторная работа № 2

по дисциплине:

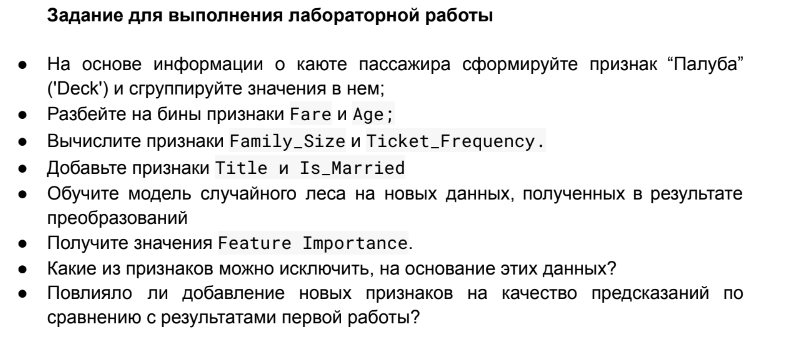
**Python для анализа данных**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Исполнитель:** |  | | | | |
| студент группы | 0В02 |  | Редько Д.А. |  |  |
|  |  |  |  |  |  |
| **Руководитель:** |  | | | | |
| преподаватель |  |  | Гоморов А.Л. |  |  |
|  |  |  |  |  |  |

Томск – 2023

**Цель работы:**

Ознакомиться с основными методами feature engineering для задач машинного обучения и применить полученные знания для обучения выборки на заданном датасете.



**Ход работы:**

Импортируем данные при помощи pandas (рис. 1).



Рисунок 1. Импортирование данных

При помощи replace заменим все T на A. При помощи словаря deck\_map сформируем Deck по первому символу данных Cabin (рис. 2)



Рисунок . Группировка значений в признаке “Deck”

Разделим Fare на бины, построим столбчатую диаграмму (рис. 3)

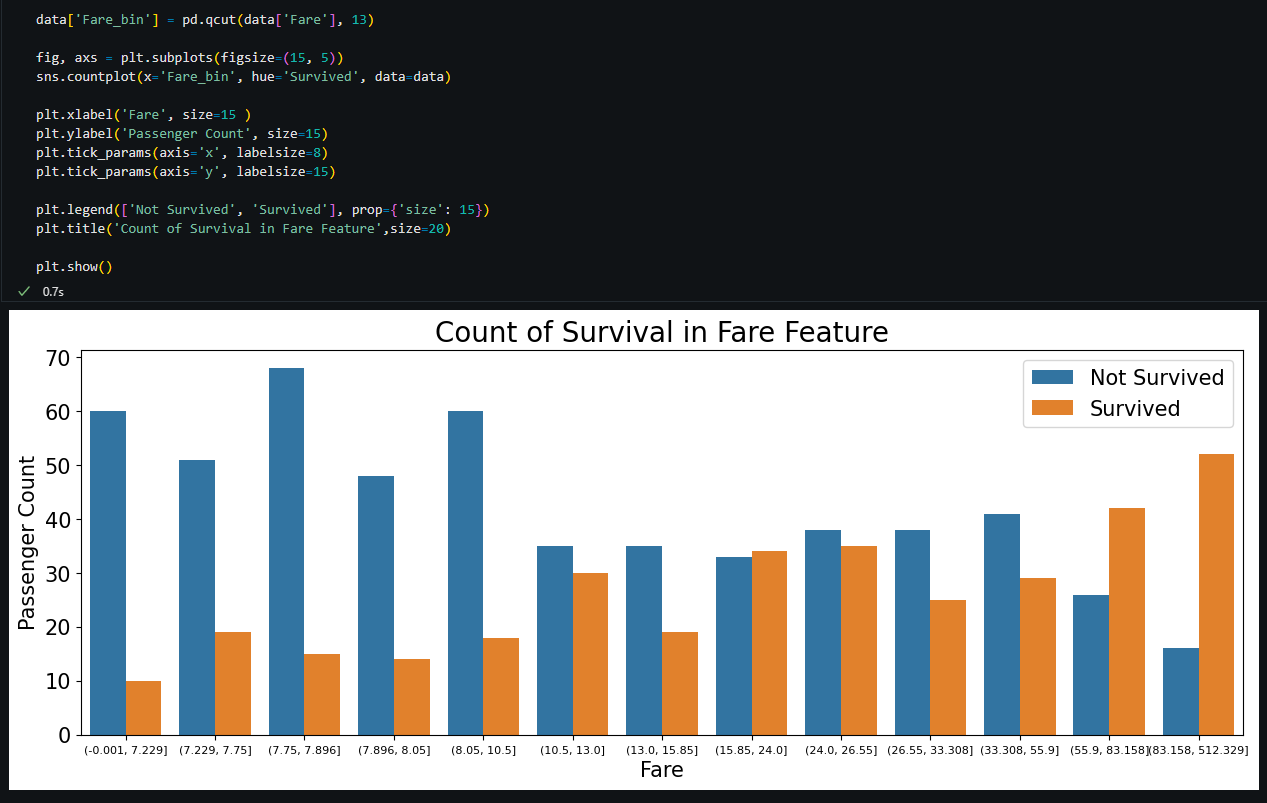


Рисунок . Разделение Fare на бины.

Аналогичным образом поступим со столбцом Age (рис. 4)

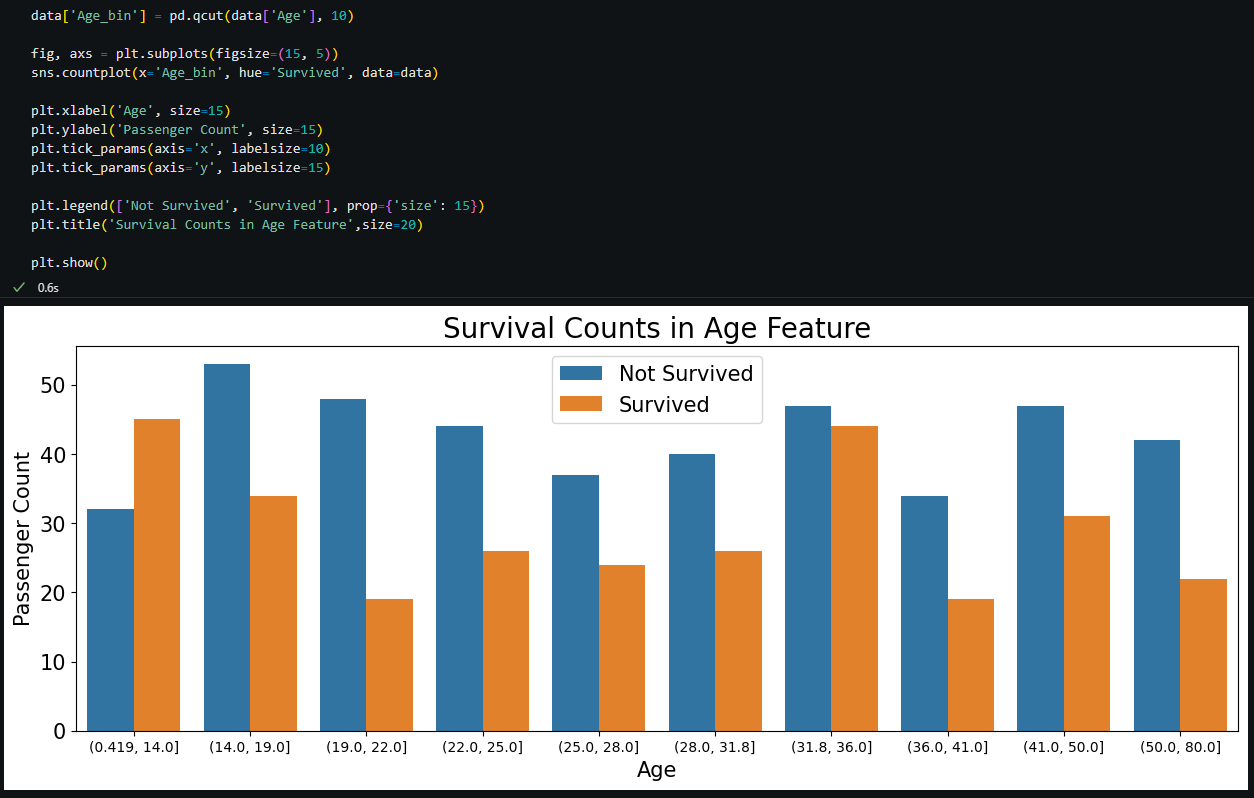


Рисунок . Разделение Age на бины.

Посчитаем размер семьи каждого пассажира (рис. 5)

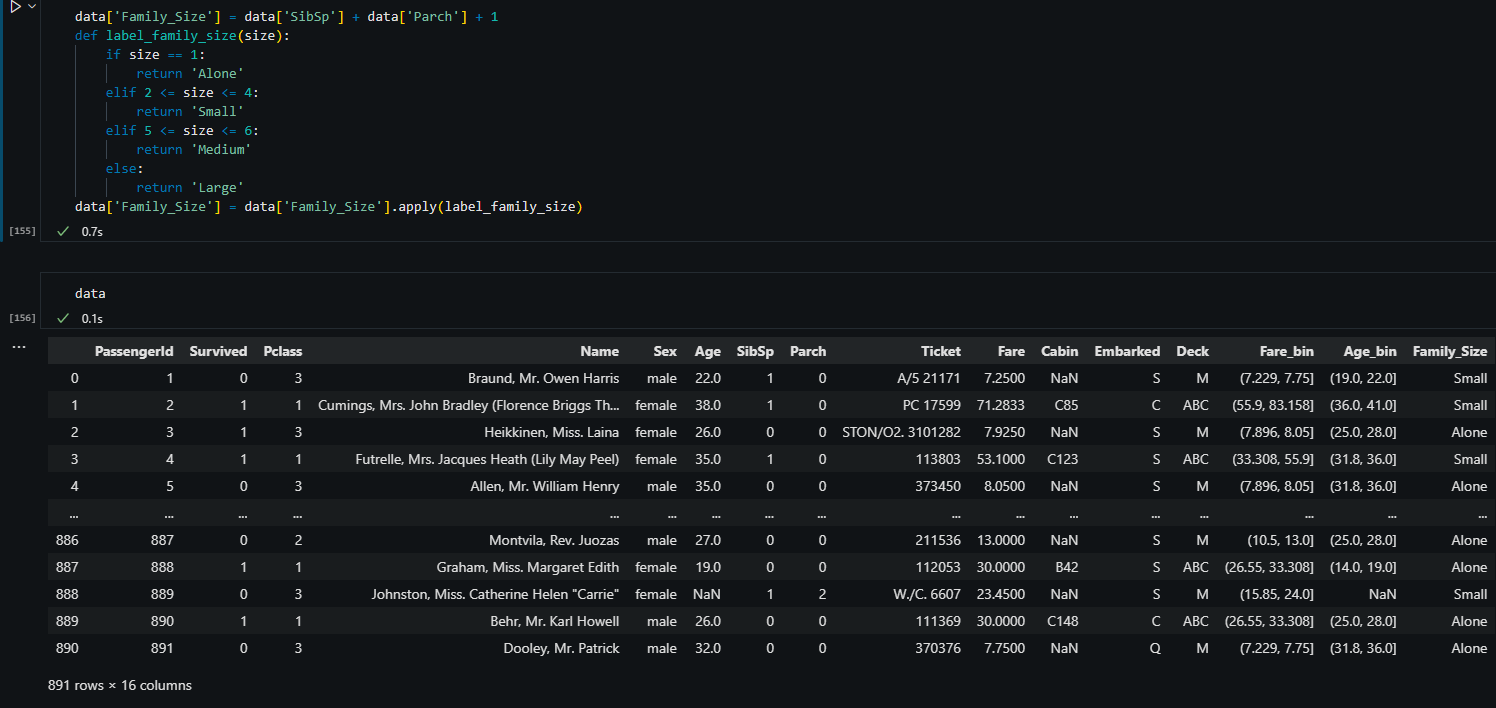


Рисунок . Размер семьи пассажиров

Заметим, что преобладающее количество пассажиров были без семьи или с маленькой семьей (рис. 6, 7).

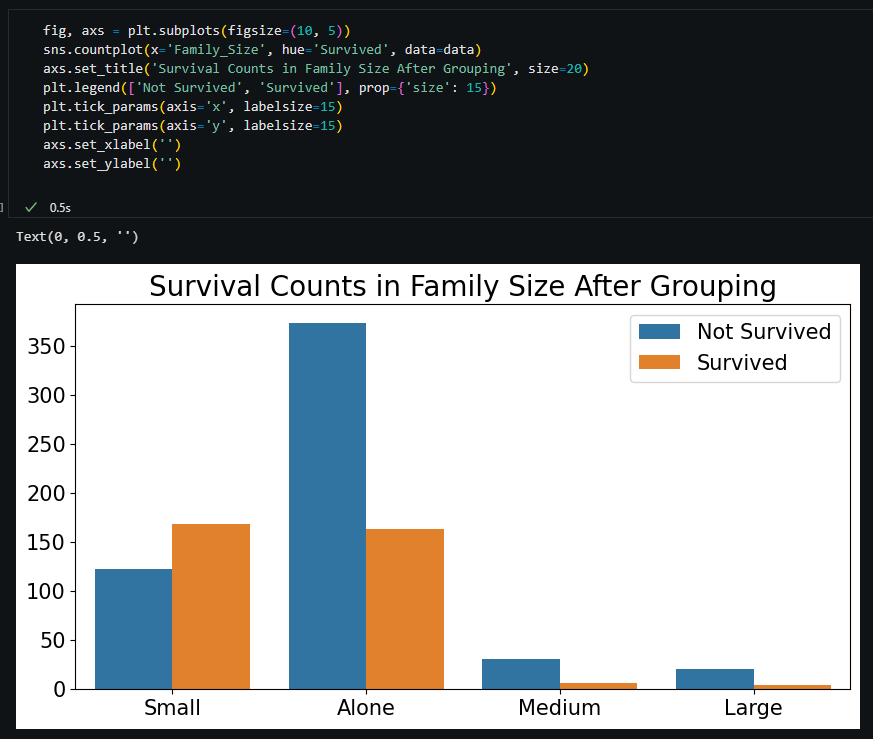


Рисунок . Размер семьи в соответствии с выжившими

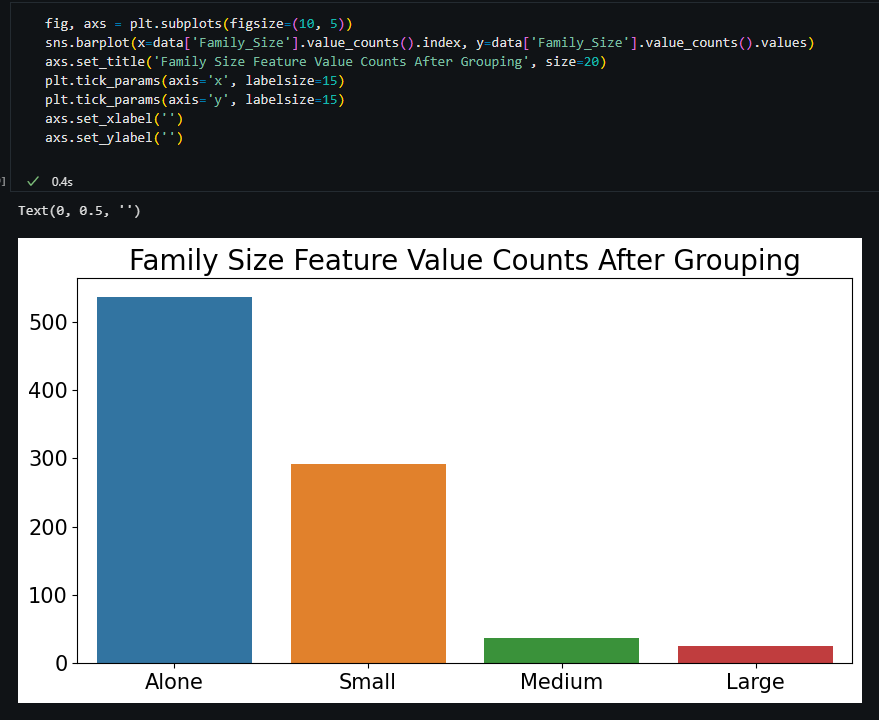


Рисунок . Размер семьи всех пассажиров

Чтобы уменьшить количество уникальных данных, вычислим Ticket\_Frequency (рис. 8)

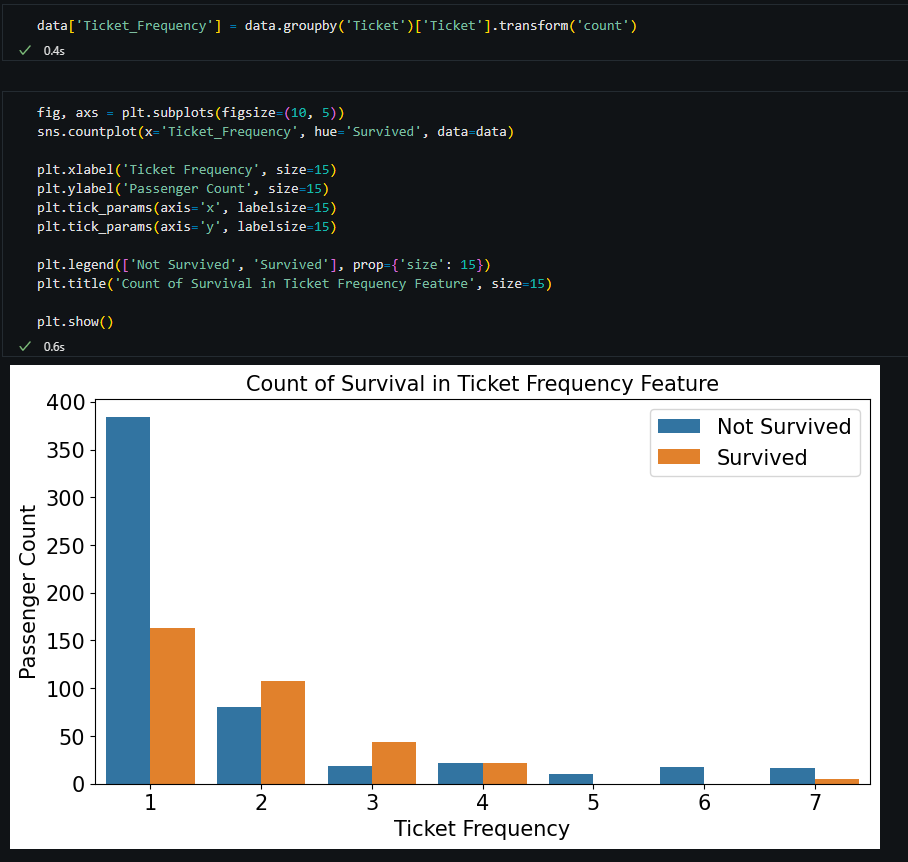


Рисунок . Частота билетов

По данным Name получим статус пассажира, определим женат(замужем) или нет, а также сгруппируем статусы, чтобы не было слишком много уникальных значений (рис. 9, 10).

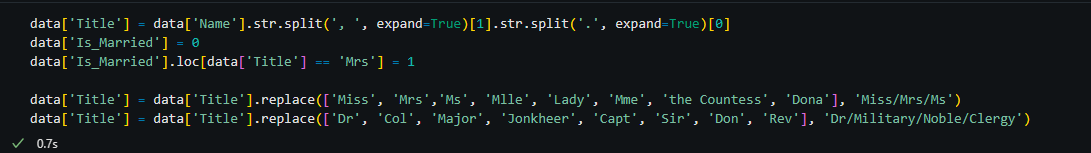


Рисунок . Формирование Title и Is\_Married

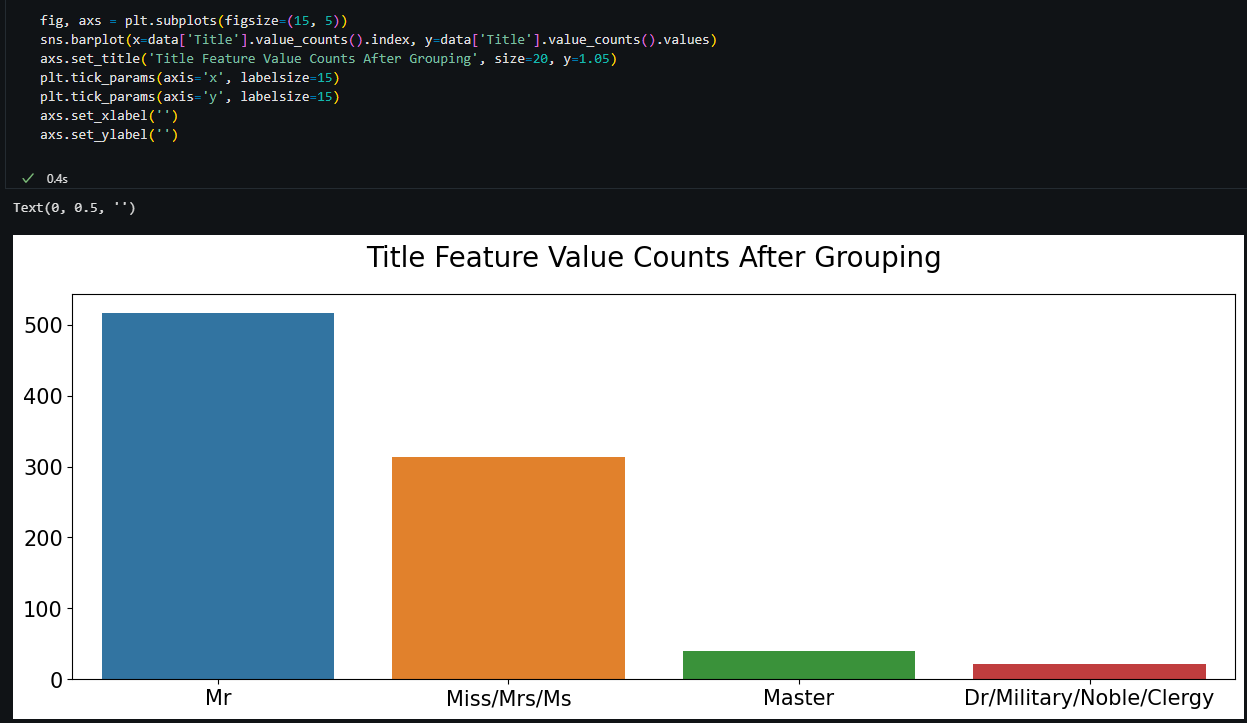


Рисунок . Подсчет количества пассажиров с каждым из статусов

Подготовим данные к обучению, преобразуя все нечисловые данные к числовым и удаляя ненужные (рис. 11).

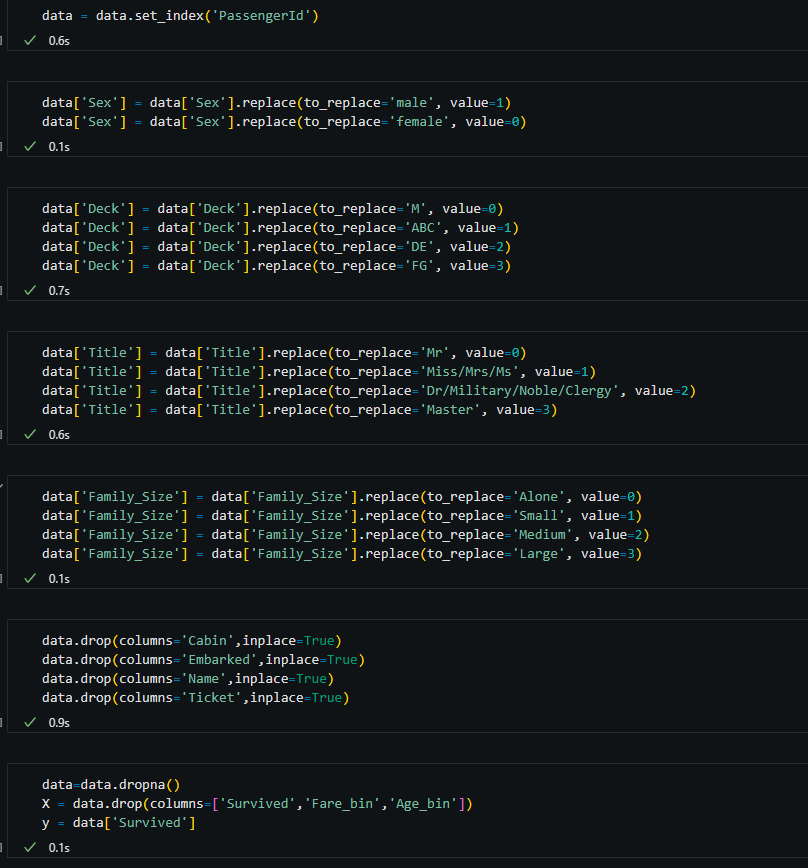


Рисунок . Формирование набора для обучения

Разделим полученную после конкатенации таблицу на обучающую и тестовую выборку. Методом случайного леса обучим наш классификатор на обучающей выборке правильно распознавать выжил пассажир или нет. Предскажем при помощи обученной модели значения для тестовой выборки, получим точность предсказаний на обучающем и тестовом наборе данных. Определим, какие признаки оказались наиболее значимы при обучении модели (рис. 12).

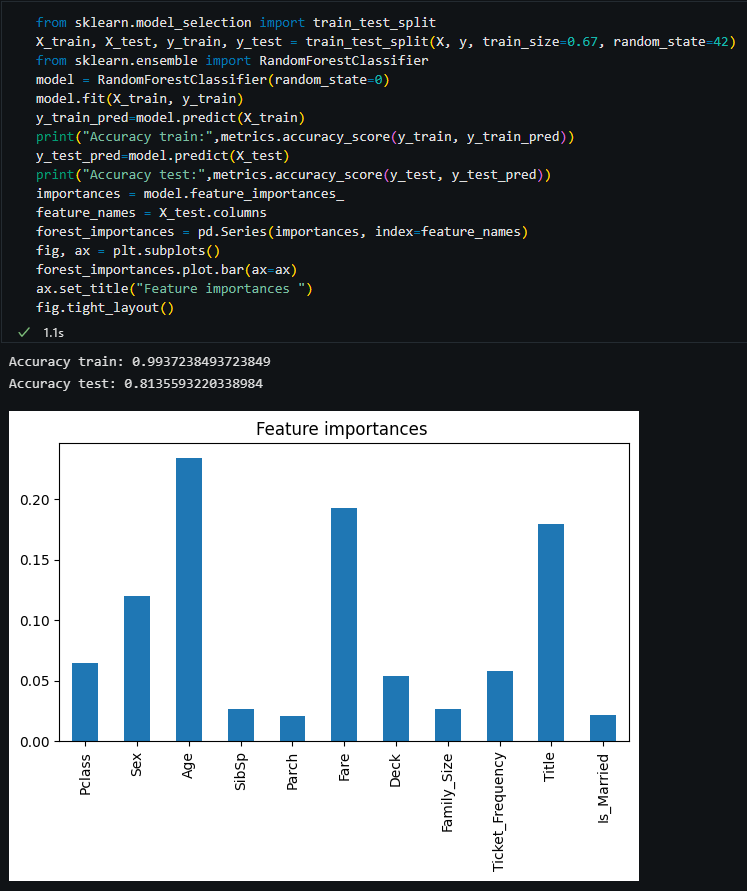


Рисунок 12. Значимость признаков при обучении

**Вывод:**

В ходе данной лабораторной работы исследован набор данных “Titanic”. На основе данных получены новые атрибуты размера семьи, титула и семейного положения. Представлено описание отдельных признаков. Набор подготовлен данных для дальнейшего их использования при обучении модели. Данные разделены на обучающую и тестовую выборки. Результаты обучения на обучающей выборке: 0.9937238493723849 и 0.8135593220338984 на тестовой. Предыдущие результаты: 0.979866, на тестовой – 0.8. Таким образом, видим положительную динамику на примерно две доли процента.