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 № 1

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 подготовки данных для задач машинного обучения и применить полученные знания для обучения выборки на заданном датасете.

**Задание к лабораторной работе:**

Исследовать представленный набор данных, составить описание отдельных признаков, произвести их нормализацию и векторизацию. Полностью подготовить набор данных для дальнейшего их использования при обучении модели. Подготовить обучающие и тестовые данные для модели, обучить модель, проверить показатели качества на обучающей и тестовой выборках. Исследовать важность признаков, проверить, на сколько влияет удаление незначимых признаков на итоговый результат классификации.

**Ход работы:**

Импортируем данные при помощи pandas (рис. 1).



Рисунок 1. Импортирование данных

При помощи команды data.shape определим количество записей в нашей таблице и количество столбцов. Заметим, что PassengerId совпадает с точностью до порядка индексации с id Pandas, поэтому столбец PassengerId зададим в качестве столбца индексов (рис. 2)

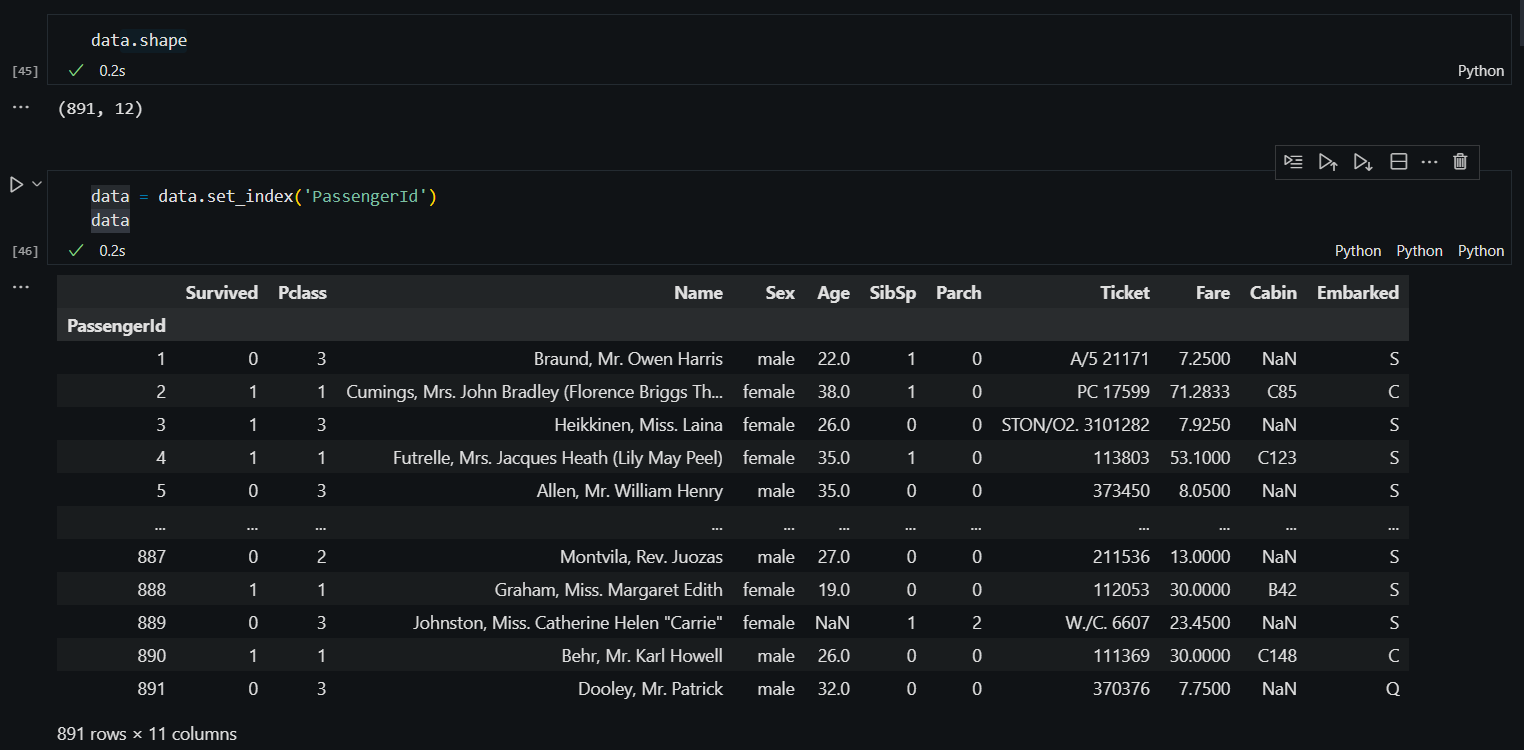


Рисунок 2. PassengerId в качестве столбца индексов

При помощи методов tail и head получим первые 5 последних и первых записей data соответственно (рис. 3)

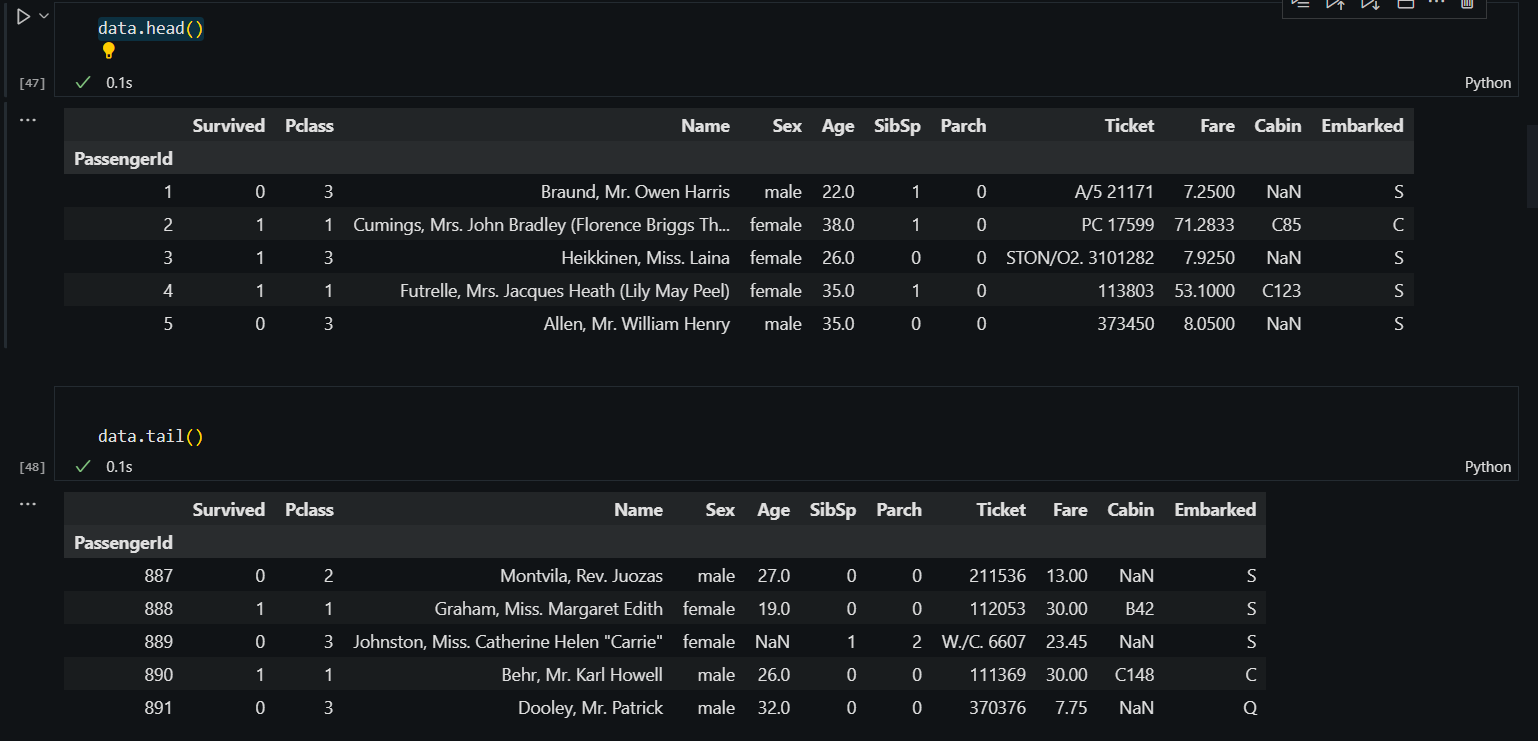


Рисунок 3. Пять первых и последних элементов

При помощи метода describe получим описание данных по столбцам, а именно количество записей, среднее значение, стандартное отклонение, минимум, максимум и квантили 25%, 50%, 75%. Заметим, что столбец “Age” содержит NaN значения. Заменим все NaN на среднее по столбцу (рис. 4)



Рисунок 4. Describe и заполнение NaN ячеек

Разделим данные на категориальные и числовые, для категориальных функцией describe получим описание (рис. 5)

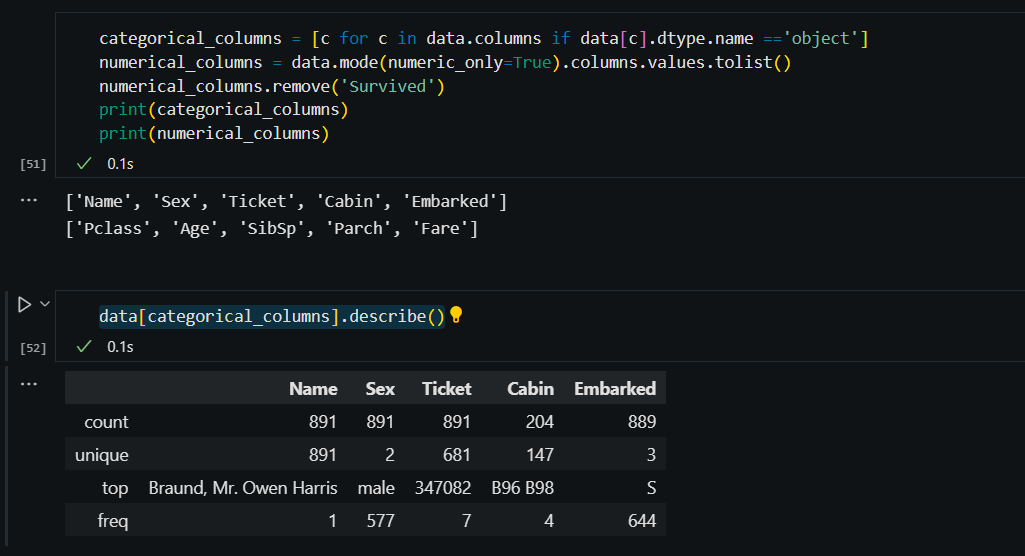


Рисунок 5. Описание категориальных данных

Заметим, что столбец Cabin содержит всего 204 не NaN ячейки, а Embarked 889 при размере выборки 891. Таким образом целесообразно удалить столбец Cabin, в столбце Embarked заменим NaN на наиболее часто встречающееся значение (рис. 6).

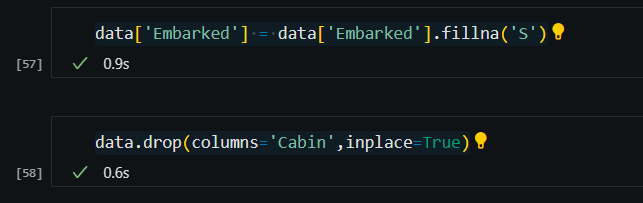


Рисунок 6. Удаление Cabin и заполнение NaN у Embarked

Выделим бинарные и небинарные категориальные данные (рис. 7).

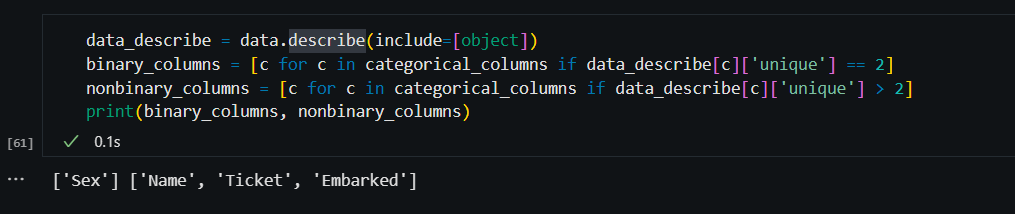


Рисунок 7. Бинарные и небинарные категориальные данные

Проведем перевод из категориальных данных в численные для столбца Embarked (рис. 8)

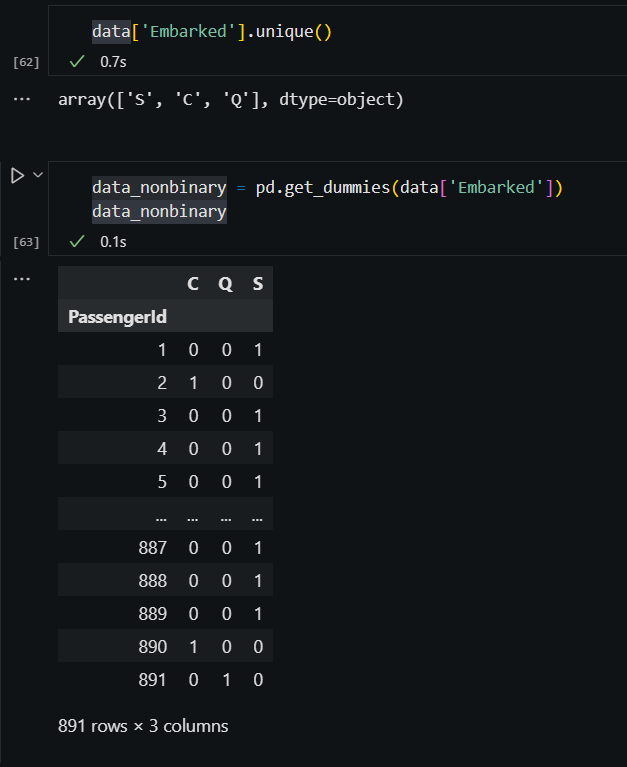


Рисунок 8. Преобразование Embarked

Столбцы Ticket и Name не влияют на выживаемость, поэтому их можно удалить. Столбец с полом представим в виде бинарного набора. (рис. 9).

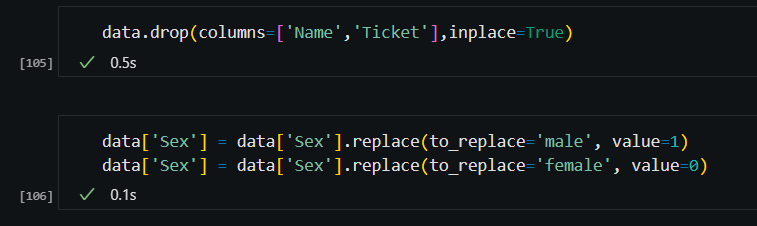


Рисунок 9. Удаление неинформативных столбцов. Бинаризация Sex.

Числовые данные нужно нормализовать. Проведем нормализацию, сделав матожидание по столбцу 0 и дисперсию 1 (рис. 10).

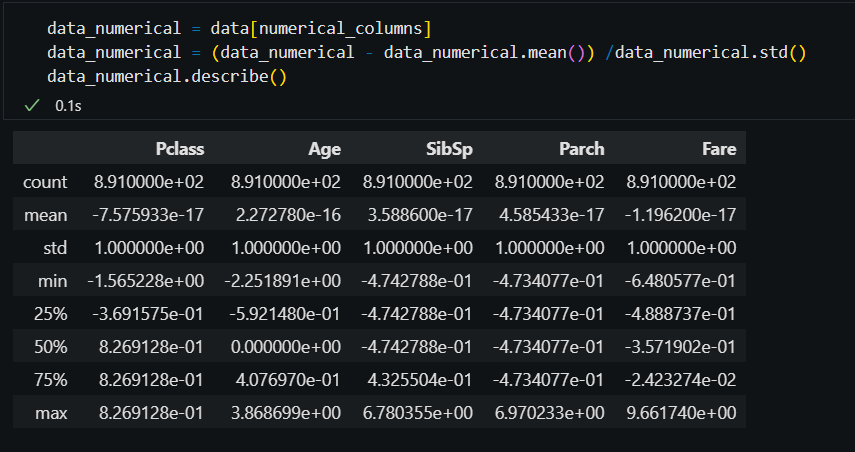


Рисунок 10. Нормализация данных

Соберем обучающую выборку по трем таблицам: числовые данные, бинарные данные и небинарные данные (рис. 11).

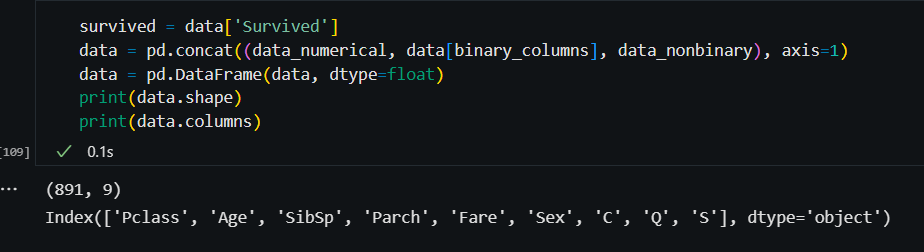


Рисунок 11. Конкатенация таблиц

Разделим полученную после конкатенации таблицу на обучающую и тестовую выборку (рис 12).

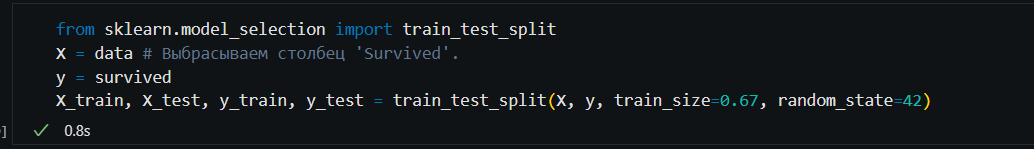


Рисунок 12. Подготовка обучающей и тестовой выборки

Методом случайного леса обучим наш классификатор на обучающей выборке правильно распознавать выжил пассажир или нет (рис. 13)

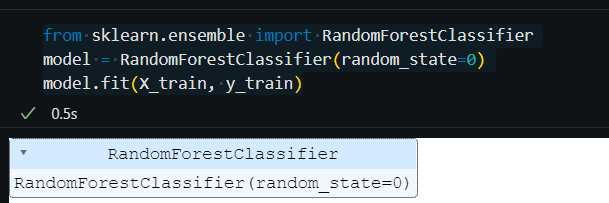


Рисунок 13. Классификатор случайный лес

Предскажем при помощи обученной модели значения для тестовой выборки, получим точность предсказаний на обучающем и тестовом наборе данных (рис. 14).

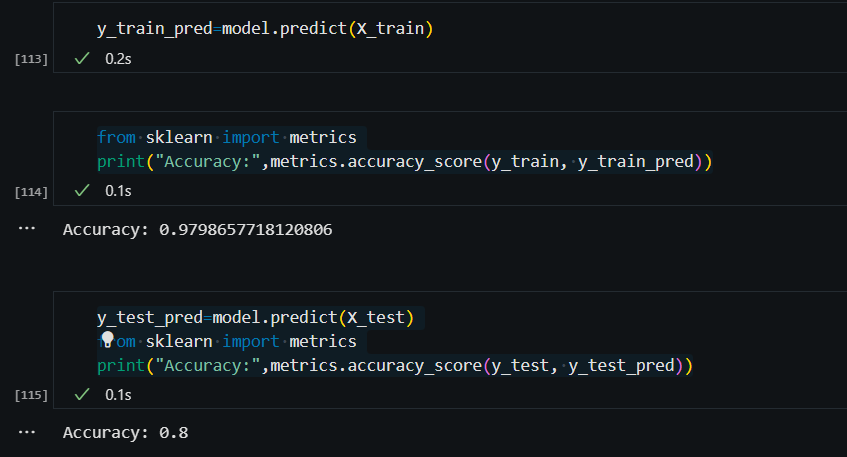


Рисунок 14. Определение точности предсказаний на обучающей и тестовой выборках

Определим, какие признаки оказались наиболее значимы при обучении модели (рис. 15).

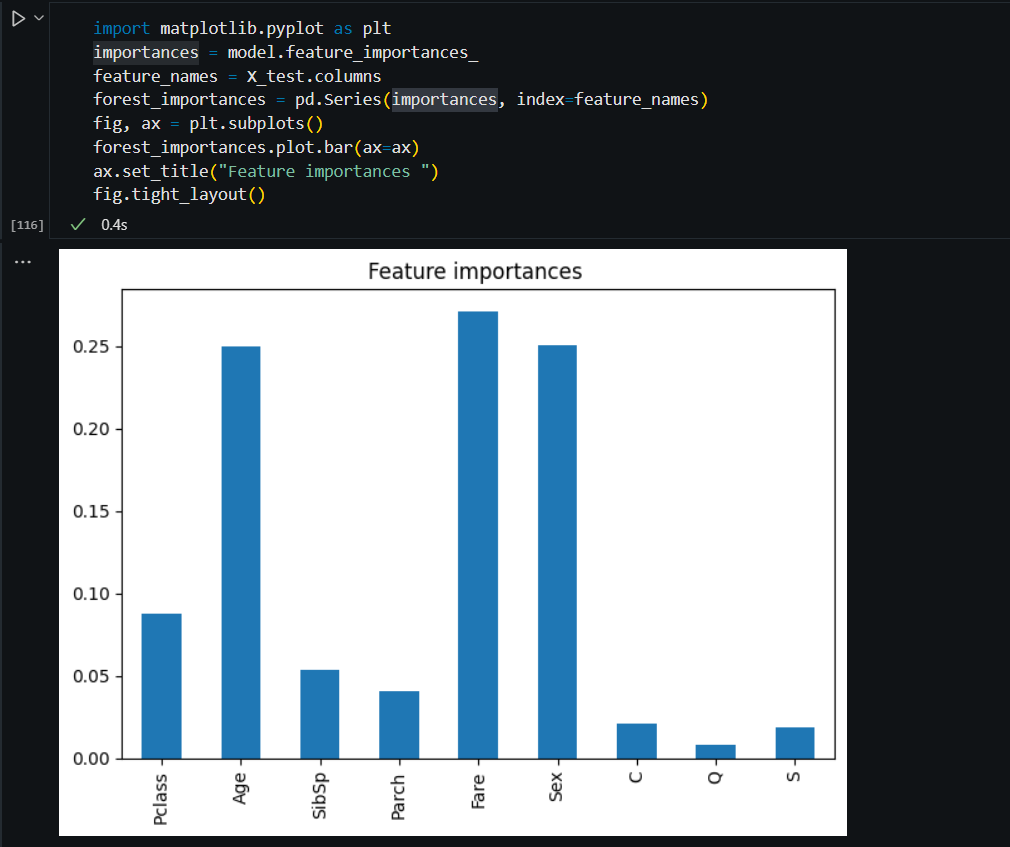


Рисунок 15. Значимость признаков при обучении

Видим, что столбцы C Q S практически не влияют на обучение. Удалим эти столбцы и аналогично обучим модель, чтобы сравнить результаты обучения с предыдущими (рис. 16, 17).

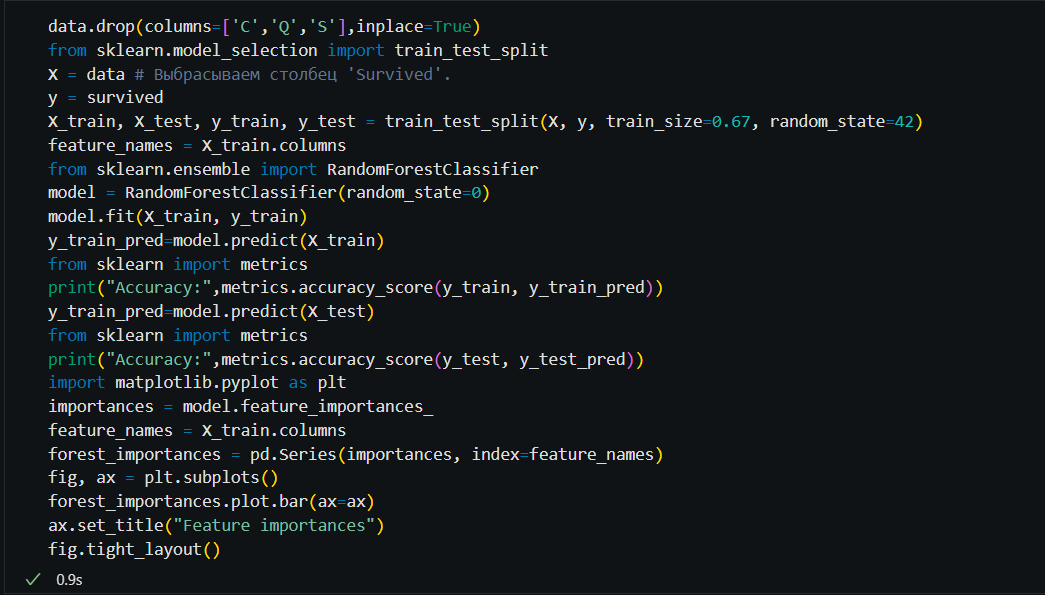


Рисунок 16. Код реализации обучения при удаленных столбцах

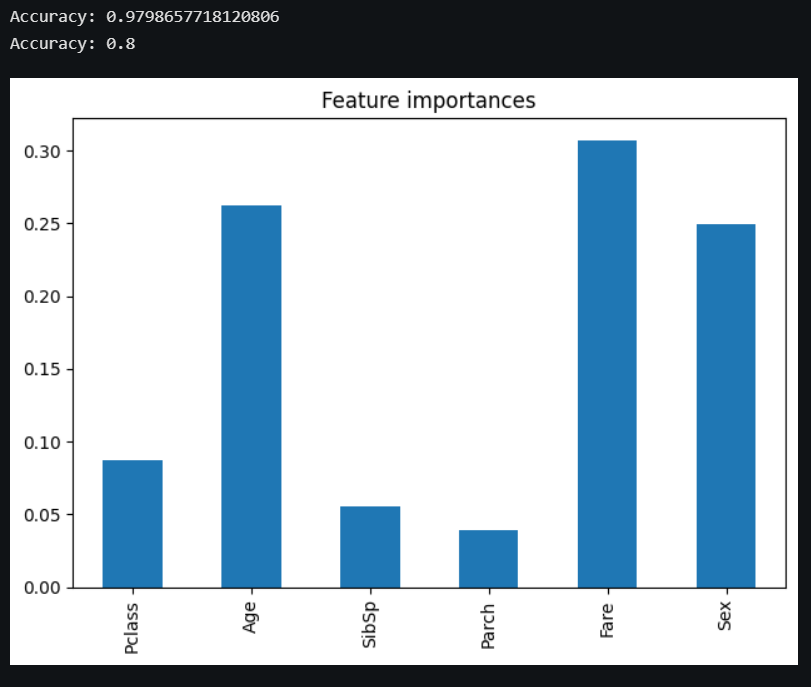


Рисунок 17. Результаты обучения при удалении столбцов.

Как и ожидалось, удаление незначимых столбцов не привело в качественному изменению результатов.

**Вывод:**

В ходе данной лабораторной работы исследован набор данных, составлено описание отдельных признаков, произведена их нормализация и векторизация. Набор подготовлен данных для дальнейшего их использования при обучении модели. Данные разделены на обучающую и тестовую выборки. Результаты обучения до удаления признаков на обучающей выборке: 0.979866, на тестовой – 0.8. Результаты обучения после удаления признаков на обучающей выборке: 0.979666, на тестовой – 0.8. Удаление признаков значимо не повлияло на варианты.