|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Системы обработки информации и управления»

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_­­\_\_Использование методов машинного обучения для \_\_\_\_\_\_\_\_\_решения задачи классификации\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент группы ИУ5-31М **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Крюков Г.М.

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Гапанюк Ю.Е.

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*Москва, 2022 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |
| --- | --- |
| по теме | Использование методов машинного обучения для решения задачи классификации |
|  | |

|  |  |
| --- | --- |
| Студент группы | ИУ5-31М |

|  |
| --- |
| Крюков Геннадий Максимович |

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

|  |
| --- |
| исследовательская |

|  |  |
| --- | --- |
| Источник тематики (кафедра, предприятие, НИР) | кафедра |

График выполнения НИР: 25% к 4 нед., 50% к 8 нед., 75% к 12 нед., 100% к 17 нед.

|  |  |
| --- | --- |
| ***Техническое задание*** |  |
|  | |
|  | |
|  | |

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \*\* листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

|  |
| --- |
|  |
|  |
|  |

Дата выдачи задания « \_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ г.

|  |  |  |  |
| --- | --- | --- | --- |
| **Руководитель НИР** |  |  | Ю.Е. Гапанюк |
|  | (Подпись, дата) |  | (И.О.Фамилия) |
| **Студент** |  |  | Г.М. Крюков |
|  | (Подпись, дата) |  | (И.О.Фамилия) |

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Содержание**

[Введение 4](#_Toc123060785)

[Основная часть 5](#_Toc123060786)

[Список литературы 33](#_Toc123060787)

# **Введение**

Пульсары - это редкий тип нейтронных звёзд, которые производят радиоизлучение, обнаруживаемое здесь, на Земле. Они представляют значительный научный интерес для исследования пространства-времени, межзвездной среды и состояний материи.

Когда пульсары вращаются, их луч их излучения проносится по небу, и когда он пересекает линию видимости, образуется обнаруживаемая картина широкополосного радиоизлучения. При динамичном вращении пульсаров эта картина периодически повторяется. Таким образом, поиск пульсаров предполагает поиск периодических радиосигналов с помощью больших радиотелескопов.

Каждый пульсар производит уникальную картину излучения, которая слегка меняется с каждым вращением. Таким образом, потенциальное обнаружение сигнала, рассматриваемого как "кандидат", усредняется по многим вращениям пульсара, определяемым длиной наблюдения. В отсутствие дополнительной информации каждый кандидат потенциально мог бы описать реальный пульсар. Однако на практике почти все обнаружения вызваны радиочастотными помехами (RFI) и шумом, что затрудняет поиск законных сигналов.

В данной работе ставится задача определения принадлежности звезды к классу пульсаров по различным параметрам с помощью методов машинного обучения.

# **Основная** **часть**

Поиск и выбор набора данных для построения моделей машинного обучения.

Описание выбранного датасета

Выбранный набор данных описывает выборку звёзд-кандидатов в пульсары, собранную во время исследования «The High Time Resolution Universe Survey».

Информация об атрибутах:

Каждый кандидат описывается 8 непрерывными переменными и одной переменной класса. Первые четыре являются простыми статистическими данными, полученными из интегрированного импульсного профиля (сложенного профиля). Это массив непрерывных переменных, описывающих разрешенную по долготе версию сигнала, которая была усреднена как по времени, так и по частоте . Остальные четыре переменные аналогично получены из кривой DM-SNR.

* Среднее значение интегрального профиля.
* Стандартное отклонение интегрированного профиля.
* Избыточный эксцесс интегрального профиля.
* Асимметрия интегрированного профиля.
* Среднее значение кривой DM-SNR.
* Стандартное отклонение кривой DM-SNR.
* Избыточный эксцесс кривой DM-SNR.
* Асимметрия кривой DM-SNR.
* Класс

Всего примеров 17 898.

* 1639 положительных примеров.
* 16 259 негативных примеров.

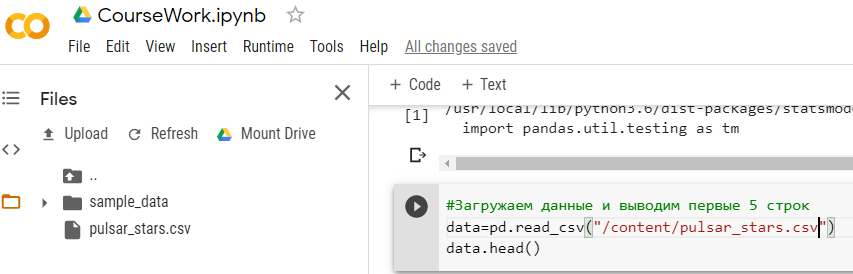
В рассматриваемом примере будем решать задачу классификации. Для классификации в качестве целевого признака будем использовать "target\_class" (Класс). Поскольку признак содержит только значения 0 и 1, то это задача бинарной классификации.

Импорт библиотек



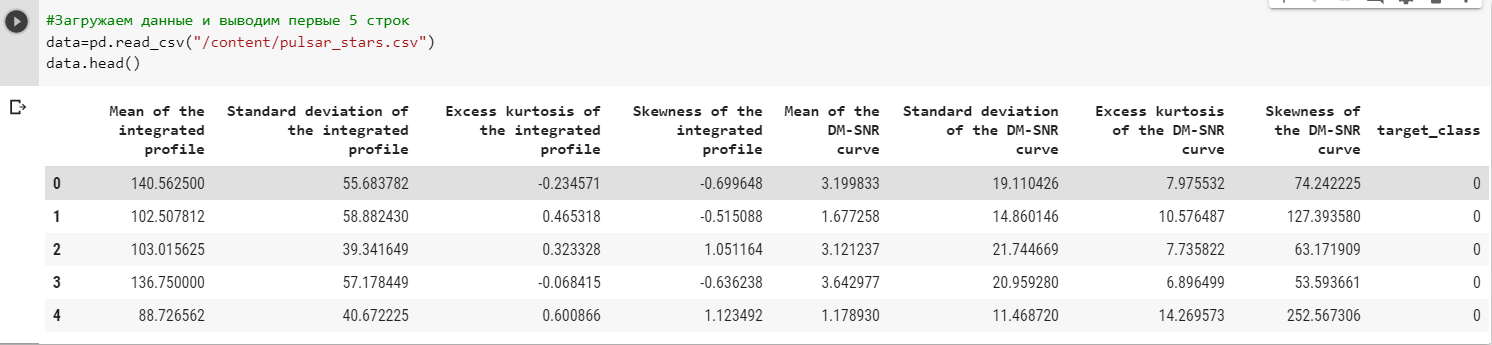
Загрузка данных.

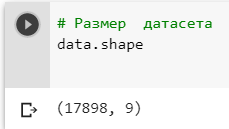
Загрузим файл датасета в помощью библиотеки Pandas.

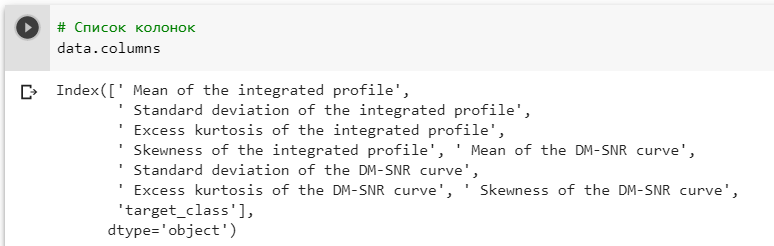


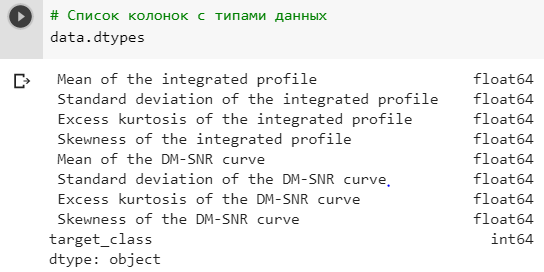
Проведем разведочный анализ. Построим графики, необходимые для понимания структуры данных. Анализируем и заполненям пропуски в данных.

Основные характеристики датасета

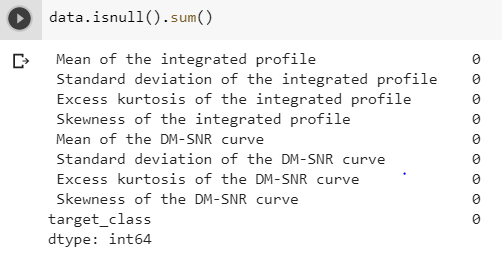






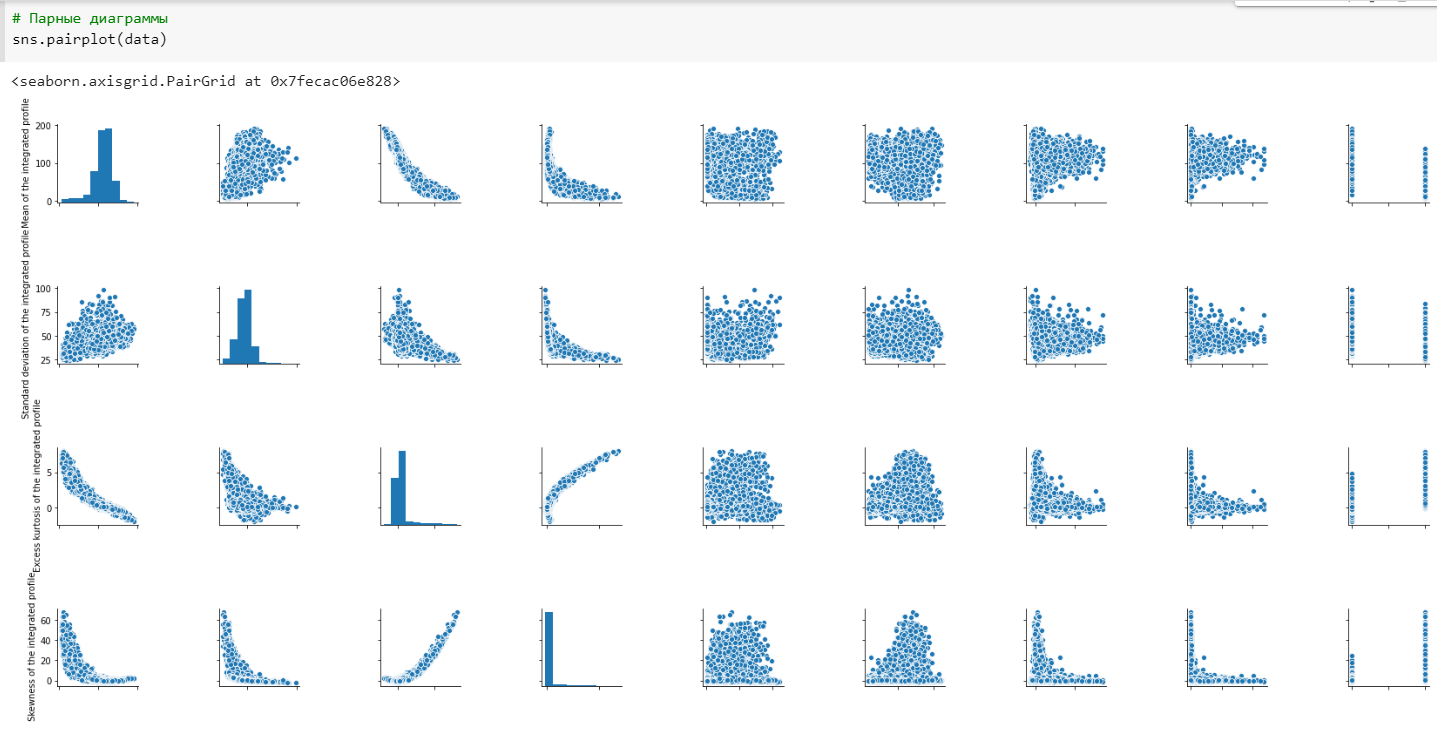


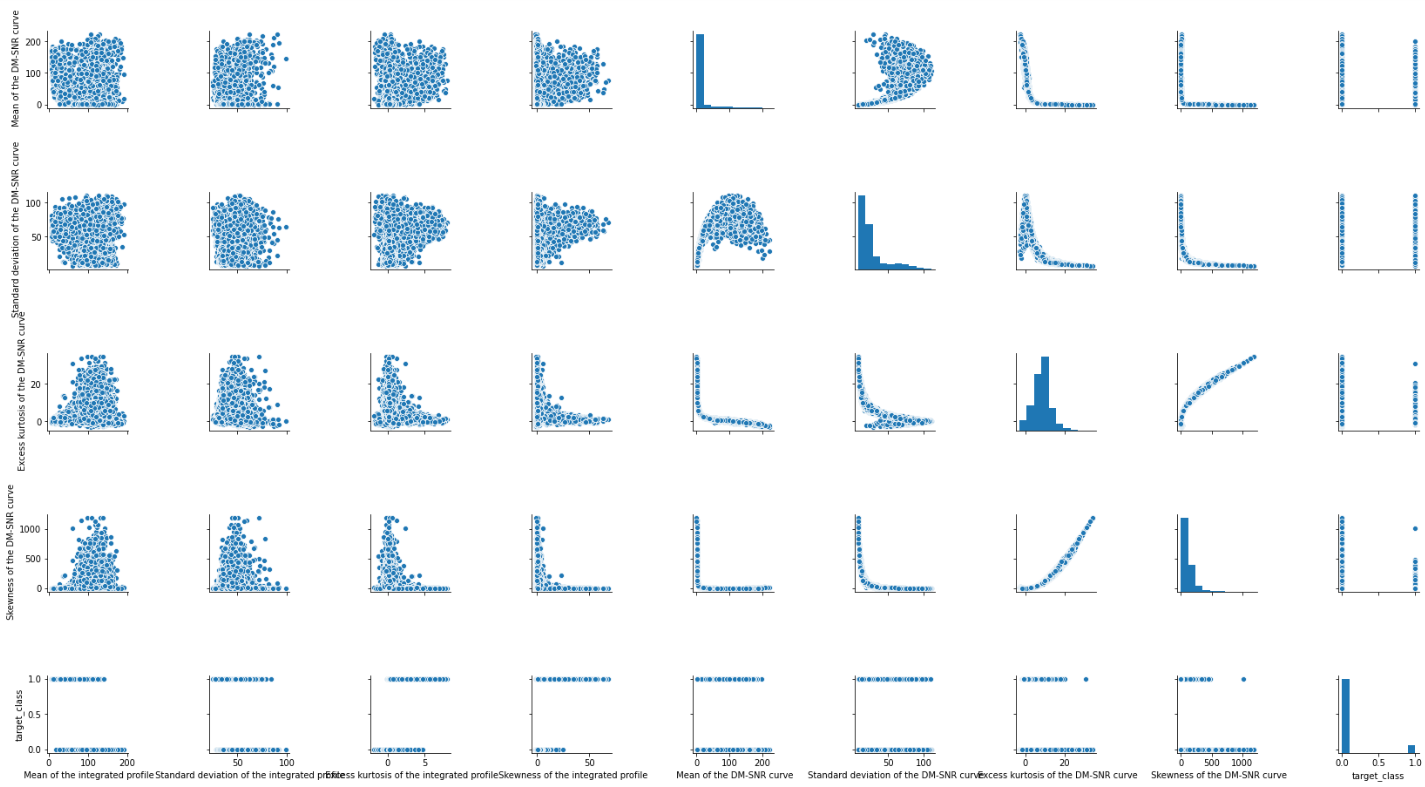
Посмотрим заполненность датасета. Возможно есть пропуски.



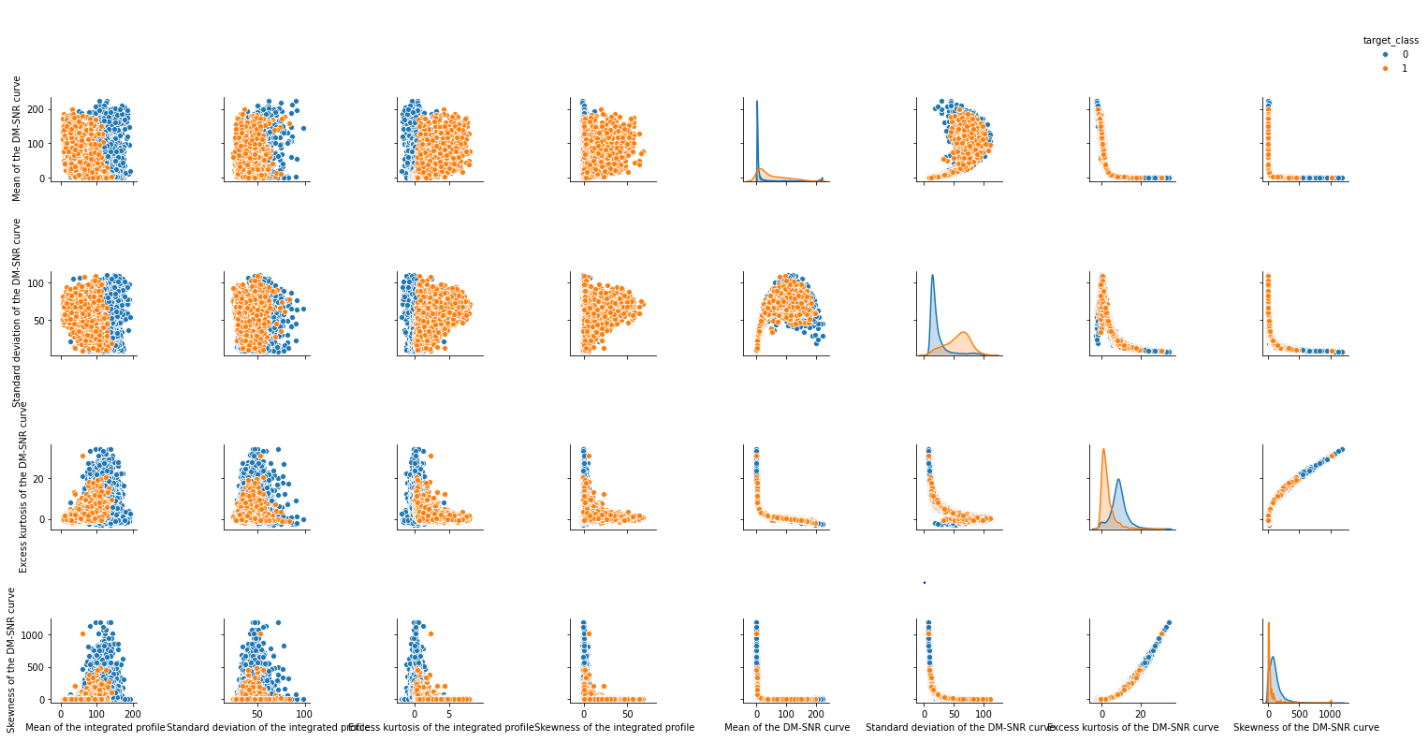
Пропуски данных отсутствуют

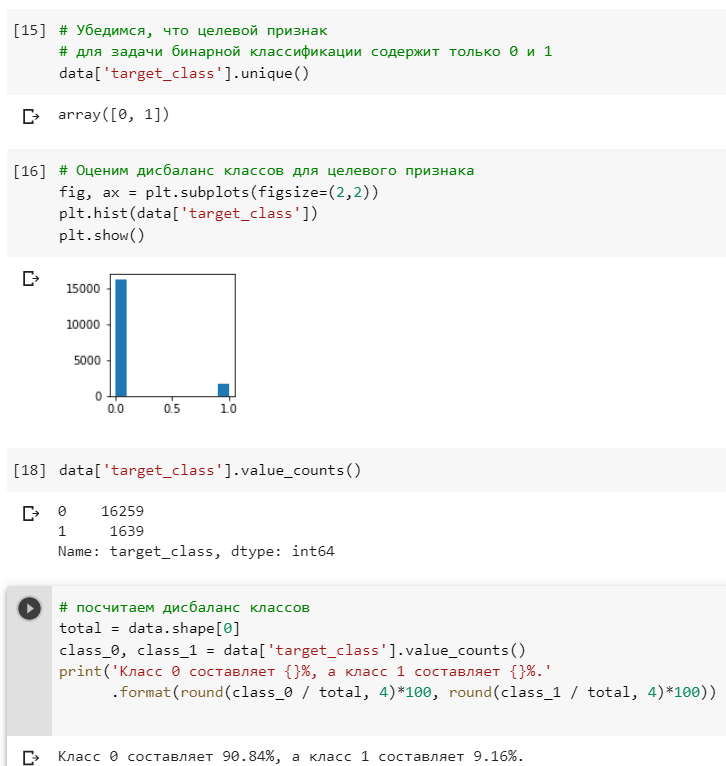
Построим некоторые графики для понимания структуры данных.



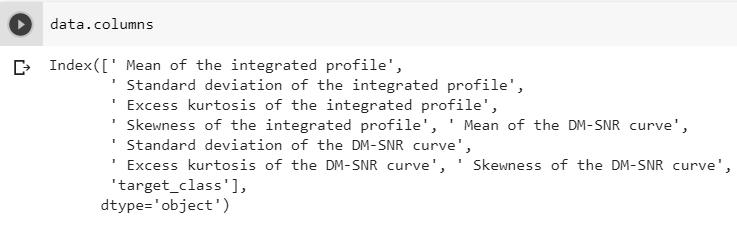


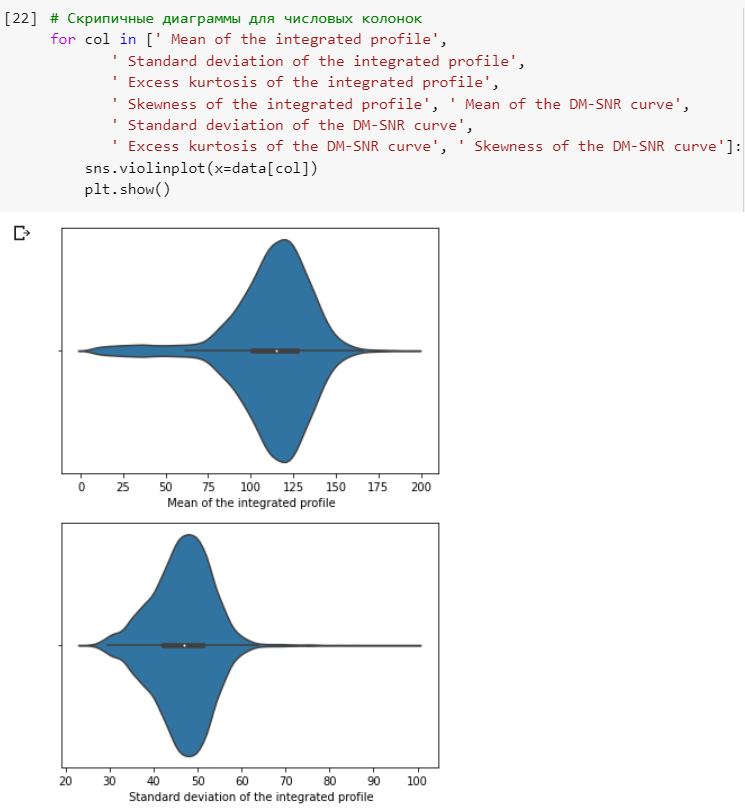


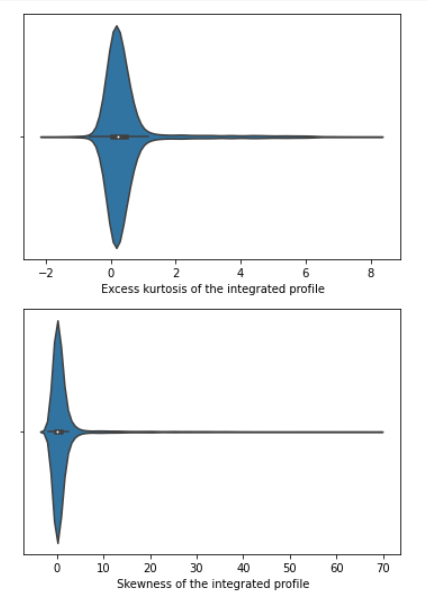
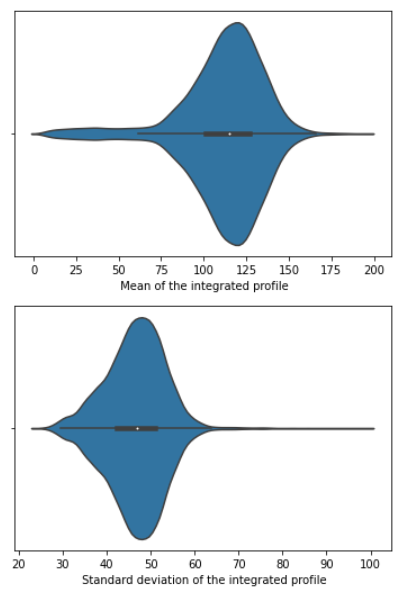


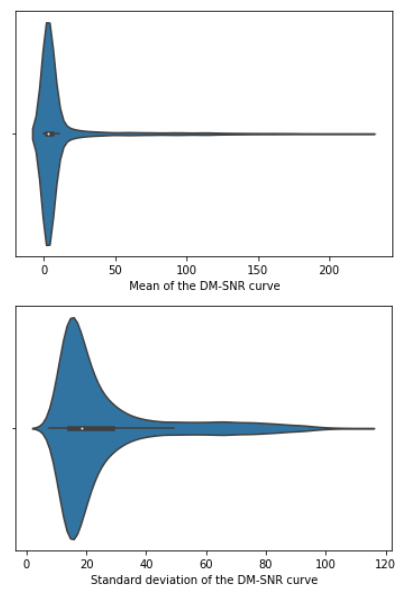
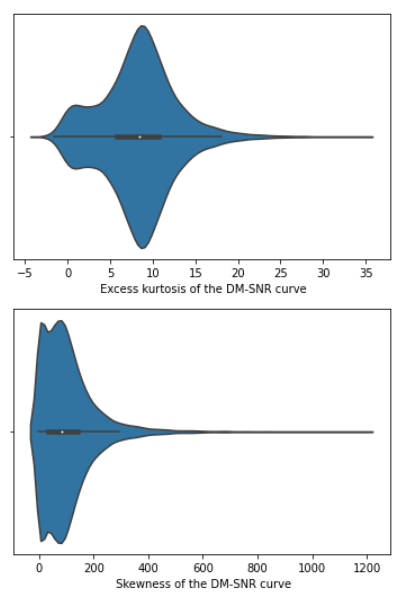


Можем наблюдать явно выраженный дисбаланс классов. Но он является приемлемым.

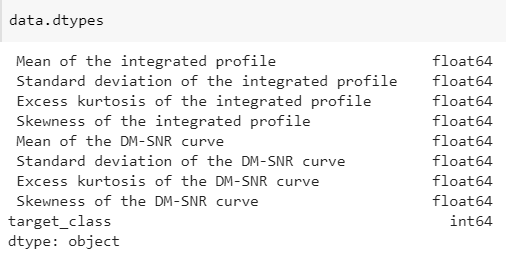






Выбираем признаки, подходящие для построения моделей.

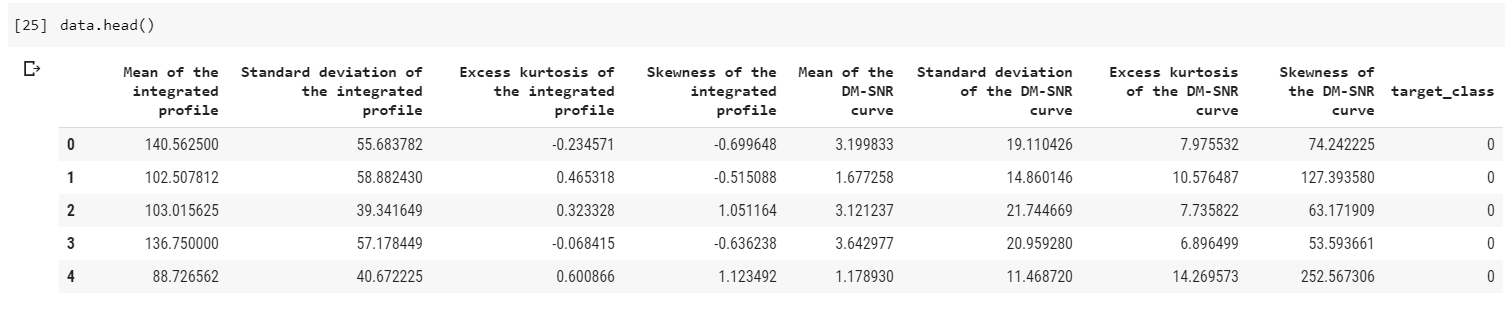


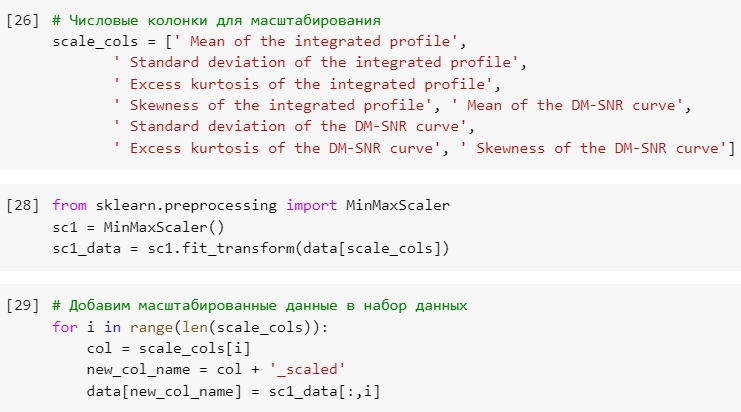
Для построения моделей будем использовать все признаки.

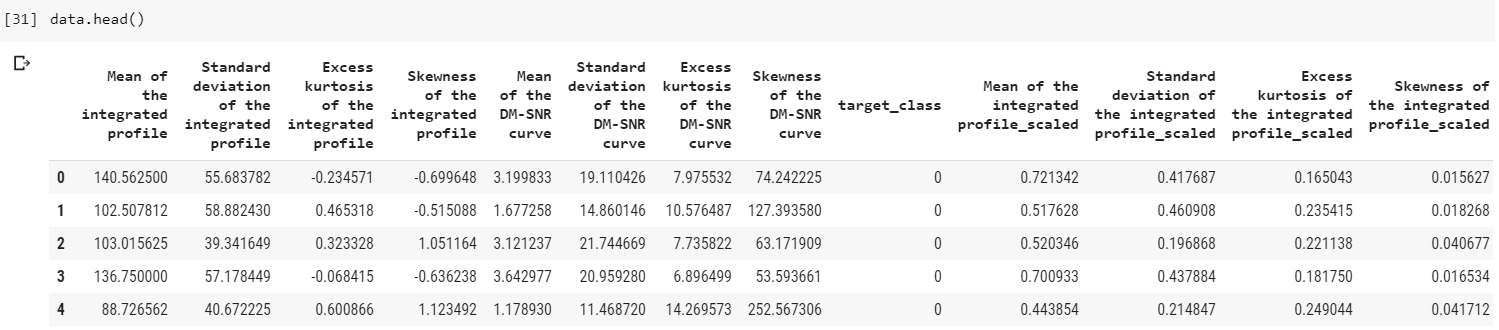
Категориальные признаки отсутствуют, их кодирования не требуется. Исключением является признак target\_class, но в представленном датасете он уже закодирован на основе подхода LabelEncoding.

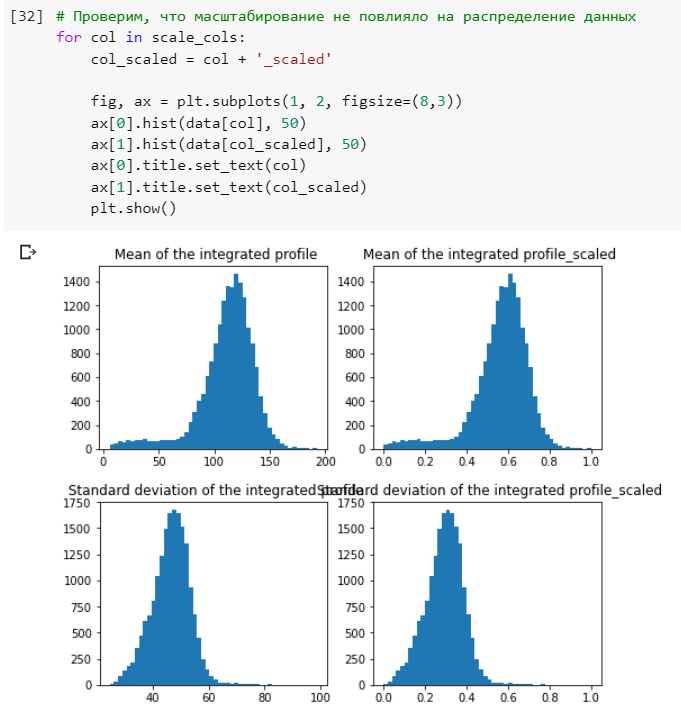
Вспомогательные признаки для улучшения качества моделей строить не будем.

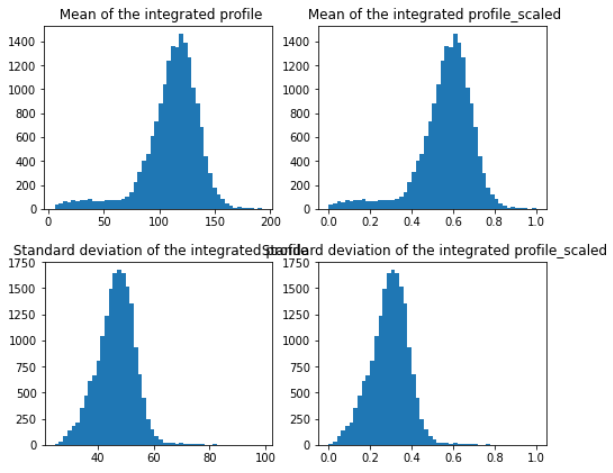
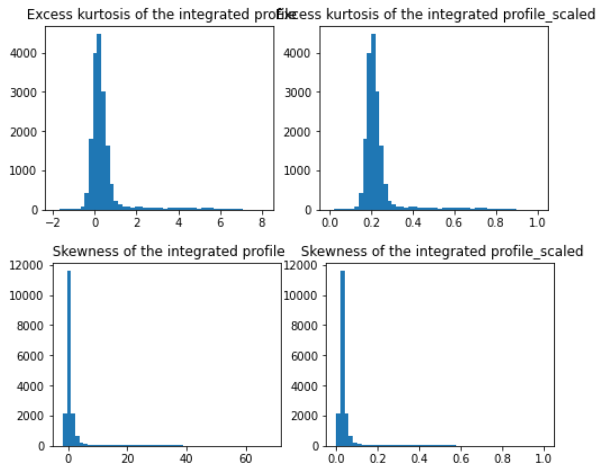
Выполним масштабирование данных.

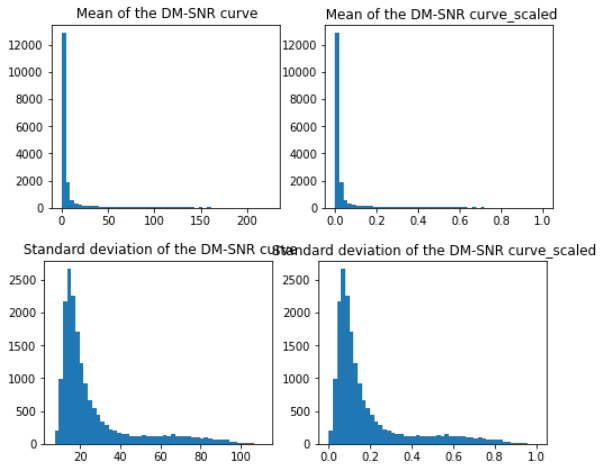
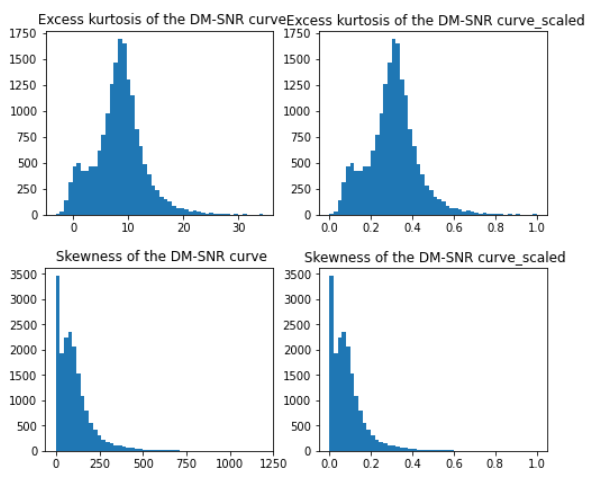




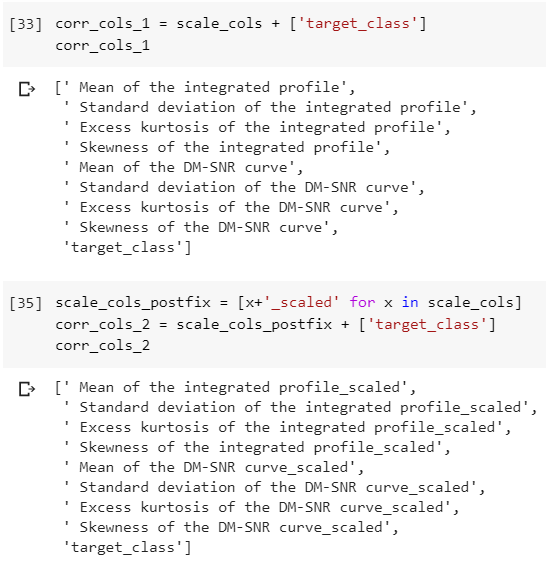




Проведем корреляционный анализ данных.







На основе корреляционной матрицы можно сделать следующие выводы:

1. Корреляционные матрицы для исходных и масштабированных данных совпадают.
2. Целевой признак классификации "target\_class" наиболее сильно коррелирует с признаками:

«Excess kurtosis of the integrated profile» (Избыточный эксцесс интегрального профиля) - 0.79

«Skewness of the integrated profile» (Асимметрия интегрированного профиля) - - 0.71

«Standard deviation of the DM-SNR curve» (Стандартное отклонение кривой DM-SNR) - 0.49.

Эти признаки следует оставить в модели классификации.

1. Признаки «Excess kurtosis of the DM-SNR curve» (Избыточный эксцесс кривой DM-SNR) и «Skewness of the DM-SNR curve» (Асимметрия кривой DM-SNR) имеют корреляцию, близкую по модулю к 1, поэтому оба признака не следуют включать в модели. Следовало бы исключить один из них, но учитывая слишком малую корреляцию их обоих с целевым признаком, можем сделать вывод о необходимости исключения каждого их этих признаков из модели.
2. Признаки «Mean of the integrated profile» (Среднее значение интегрального профиля) и «Standard deviation of the integrated profile» (Стандартное отклонение интегрированного профиля) слишком слабо коррелируют с целевым признаком, поэтому их следует исключить из модели, так как они могут ухудшить её качество.
3. Достаточно большие по модулю значения коэффициентов корреляции свидетельствуют о значимой корреляции между исходными признаками и целевым признаком. На основании корреляционной матрицы можно сделать вывод о том, что данные позволяют построить модель машинного обучения.

**Выбираем метрики для последующей оценки качества моделей.**

#### В качестве метрик для решения задачи классификации будем использовать:

#### Метрика precision:

#### 

#### Доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор верно или неверно определил, как положительные.

#### Используется функция *precision\_score*.

#### Метрика recall (полнота):

#### 

#### Доля верно предсказанных классификатором положительных объектов, из всех действительно положительных объектов.

#### Используется функция *recall\_score*.

#### Метрика F1-мера

#### Для того, чтобы объединить precision и recall в единую метрику используется Fβ-мера, которая вычисляется как среднее гармоническое от precision и recall:

#### 

#### На практике чаще всего используют вариант F1-меры (которую часто называют F-мерой) при β=1:

#### 

#### Для вычисления используется функция *f1\_score*.

#### Метрика ROC AUC

#### Используется для оценки качества бинарной классификации. Основана на вычислении следующих характеристик:

#### 

#### True Positive Rate, откладывается по оси ординат. Совпадает с recall.

#### False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно.

#### Чем сильнее отклоняется кривая от верхнего левого угла графика, тем хуже качество классификации.

#### В качестве количественной метрики используется площадь под кривой - ROC AUC (Area Under the Receiver Operating Characteristic Curve). Чем ниже проходит кривая тем меньше ее площадь и тем хуже качество классификатора.

#### Для получения ROC AUC используется функция *roc\_auc\_score*.

#### Сохранение и визуализация метрик

Разработаем класс, который позволит сохранять метрики качества построенных моделей и реализует визуализацию метрик качества.



**Выбираем наиболее подходящие модели для решения задачи классификации или регрессии.**

Для задачи классификации будем использовать следующие модели:

* Логистическая регрессия

Метод, используемый для решения задачи бинарной классификации.

Метод выдает вероятность принадлежности объекта к нулевому/единичному классам.

Используется класс *LogisticRegression*.

* Машина опорных векторов

Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Для решения задачи классификации используем класс:

*SVC* - основной классификатор на основе SVM. Поддерживает различные ядра.

* Решающее дерево

Для текущего выбранного признака (колонки) из N признаков построить все варианты ветвления по значениям (для категориальных признаков) или по диапазонам значений (для числовых признаков).

Если подвыборке соответствует единственное значение целевого признака, то в дерево добавляется терминальный лист, который соответствует предсказанному значению.

Если в подвыборке больше одного значения целевого признака, то предыдущие пункты выполняются рекурсивно для подвыборки.

Для решения задачи классификации используется класс *DecisionTreeClassifier*.

* Случайный лес (ансамблевая)

Случайный лес можно рассматривать как алгоритмом бэггинга над решающими деревьями.

Но при этом каждое решающее дерево строится на случайно выбранном подмножестве признаков. Эта особенность называется "feature bagging" и основана на методе случайных подпространств.

Случайный лес для задача классификации реализуется в scikit-learn с помощью класса *RandomForestClassifier*.

Задание параметра n\_jobs=-1 распараллеливает алгоритм на максимально возможное количество процессоров.

* Градиентный бустинг (ансамблевая)

В отличие от методов бэггинга и случайного леса, которые ориентированы прежде всего на минимизизацию дисперсии (Variance), методы бустинга ориентированы прежде всего на минимизацию смещения (Bias) и, отчасти, на минимизизацию дисперсии.

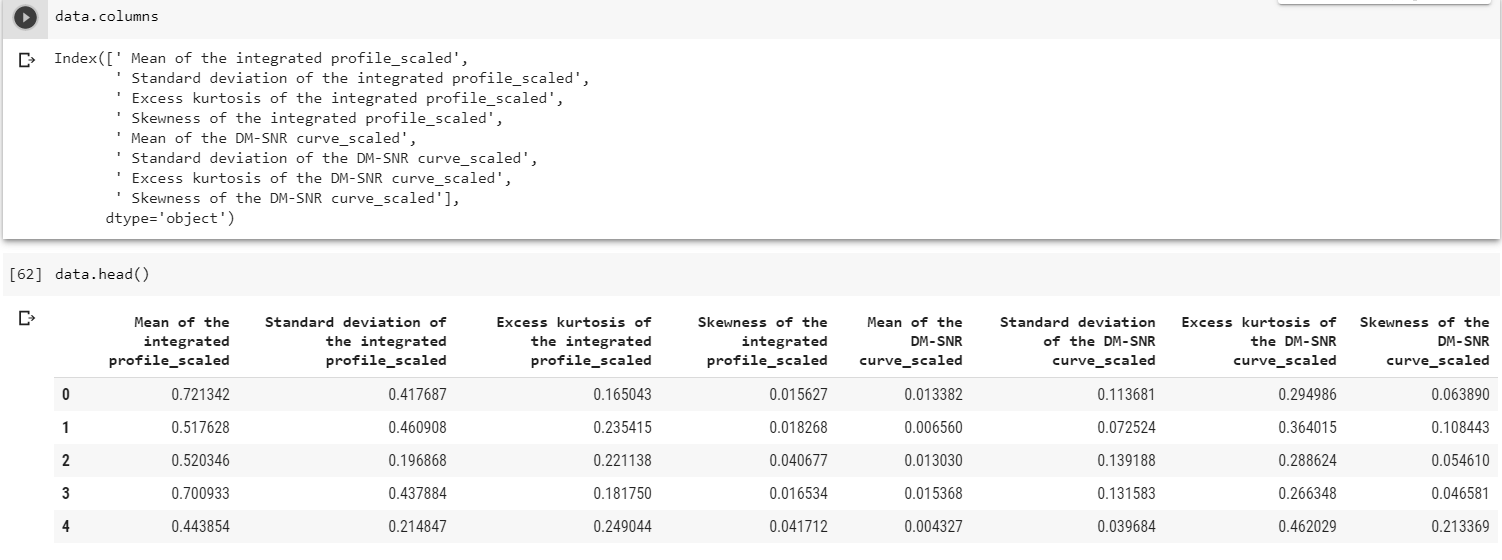
Исторически первым полноценным алгоритмом бустинга считается алгоритм AdaBoost.

AdaBoost реализуется в scikit-learn с помощью класса *AdaBoostClassifier* для задач классификации.

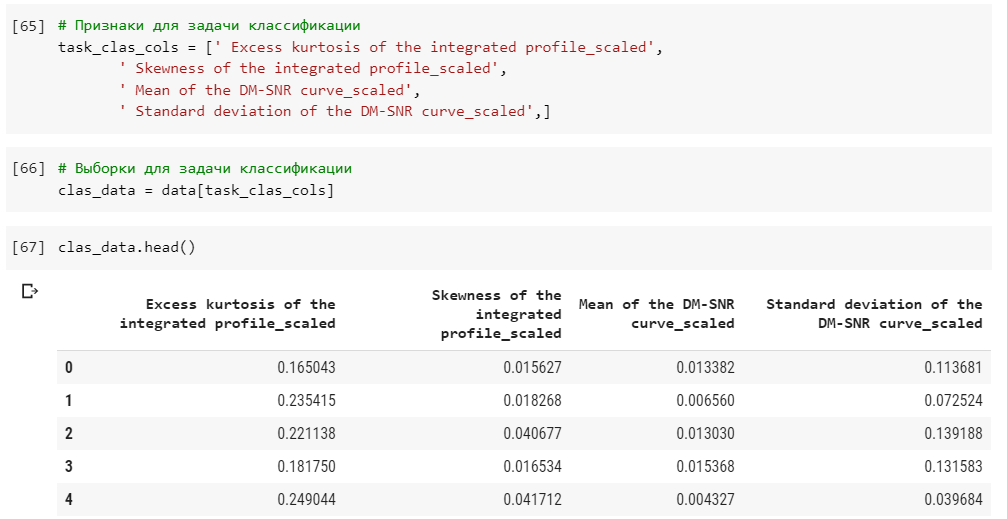
**Формируем обучающую и тестовую выборку на основе исходного набора данных.**

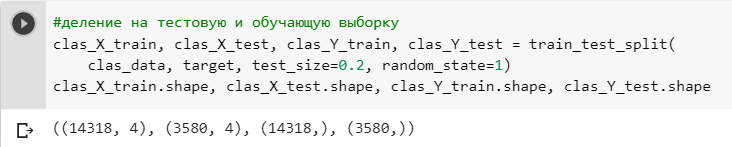
Имеем масштабированные данные:



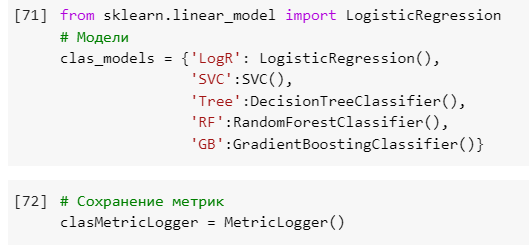


На основе масштабированных данных выделим обучающую и тестовую выборки:

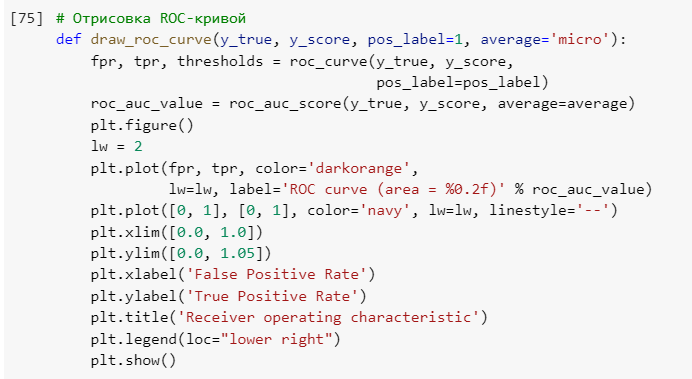


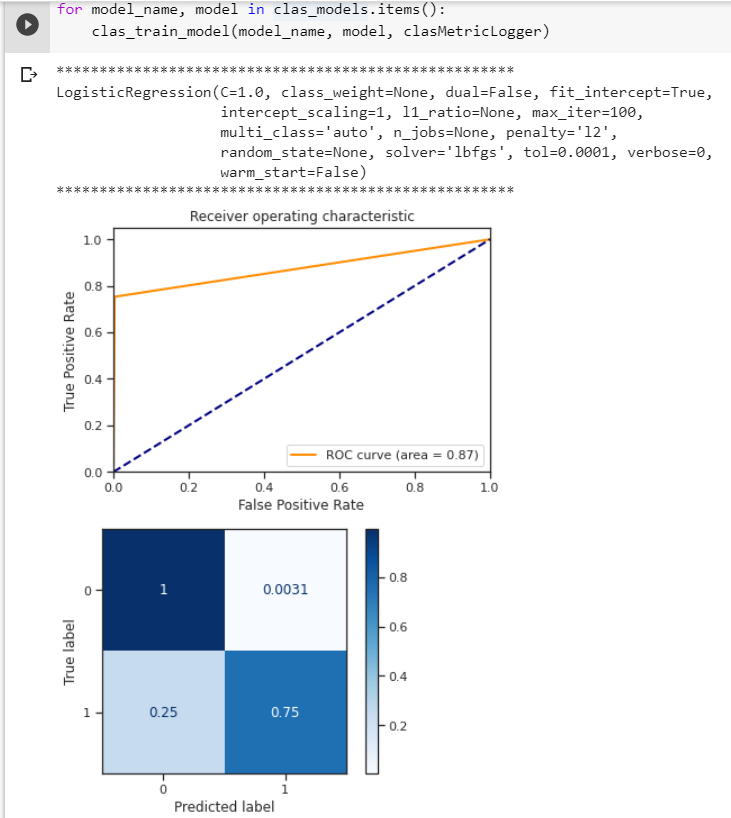


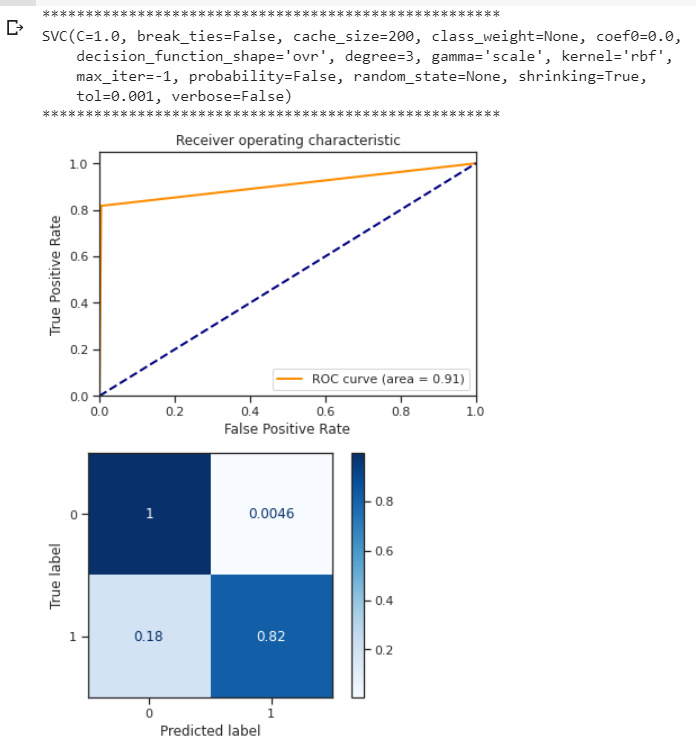
**Построим базовое решение (baseline) для выбранных моделей без подбора гиперпараметров**

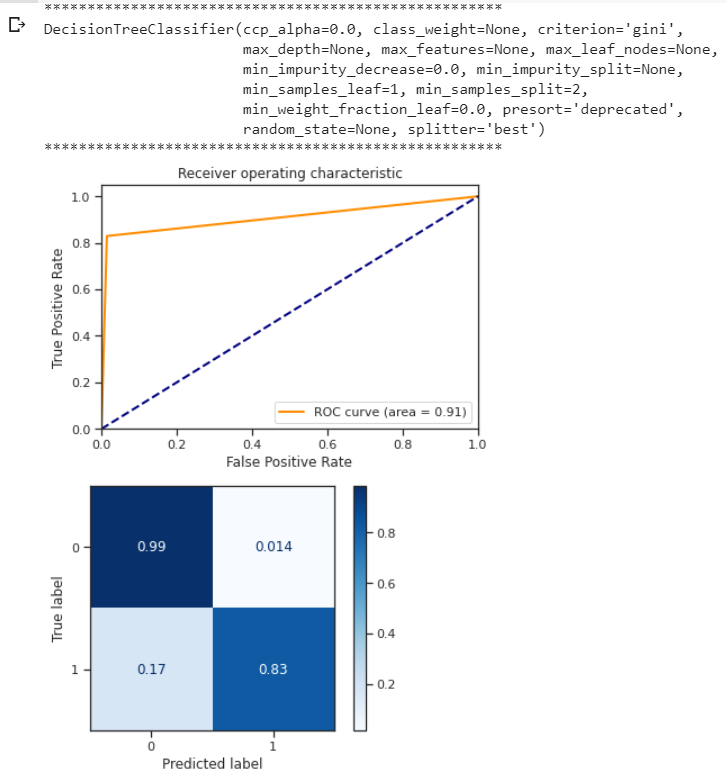


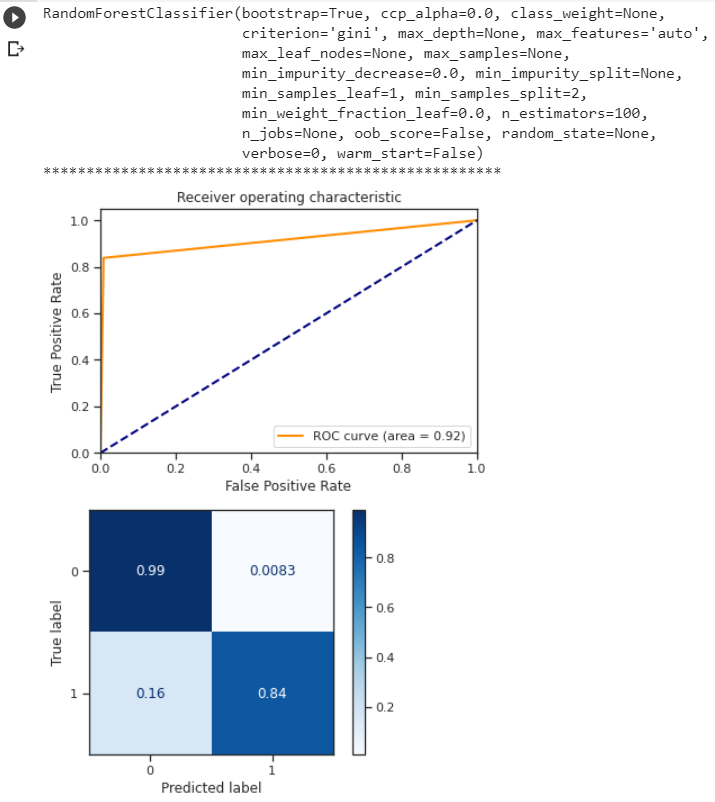


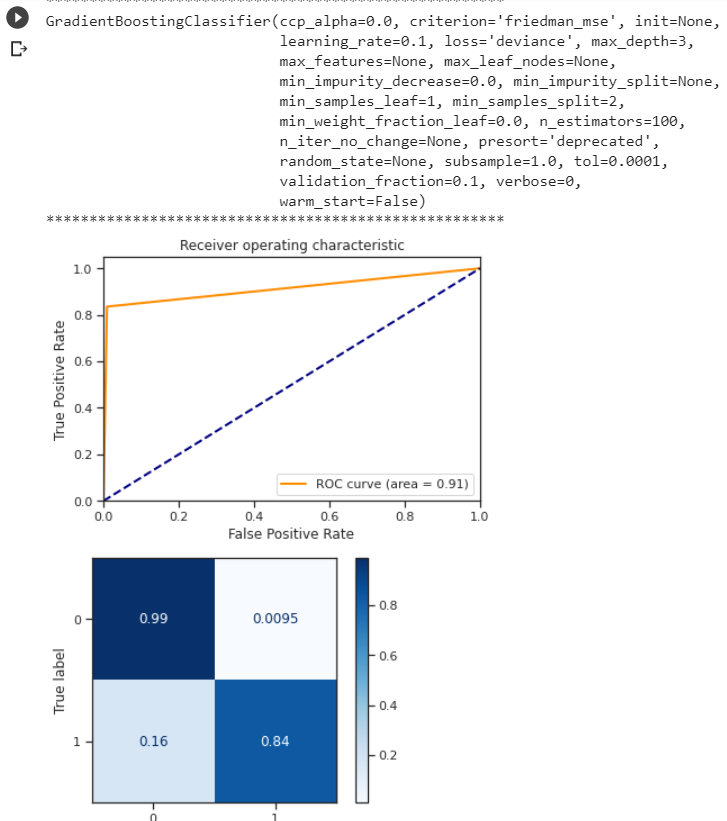






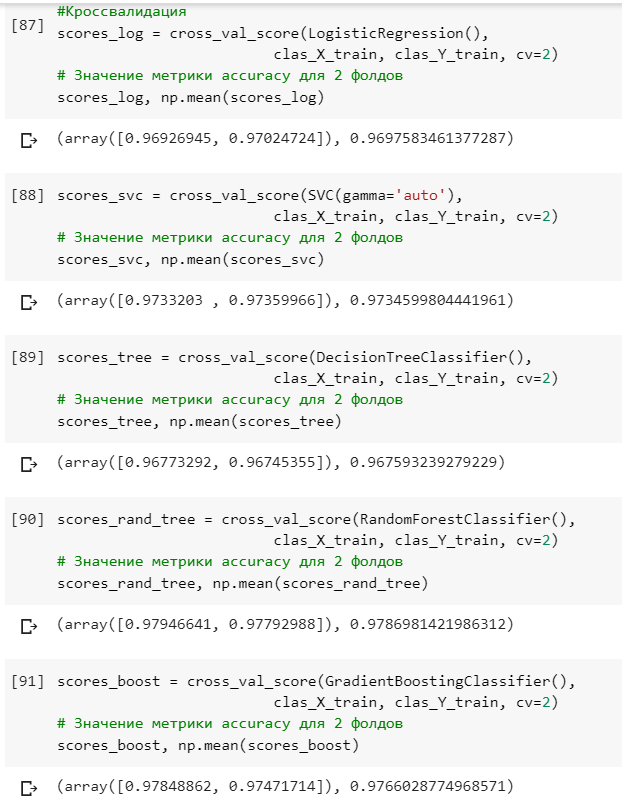


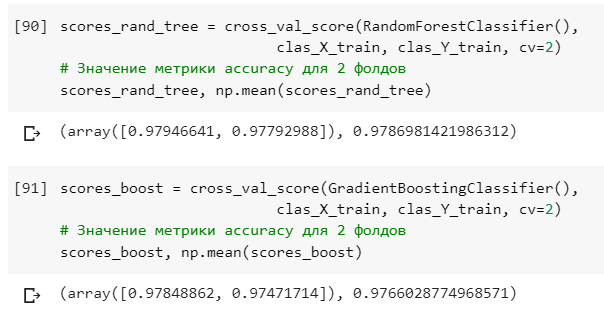




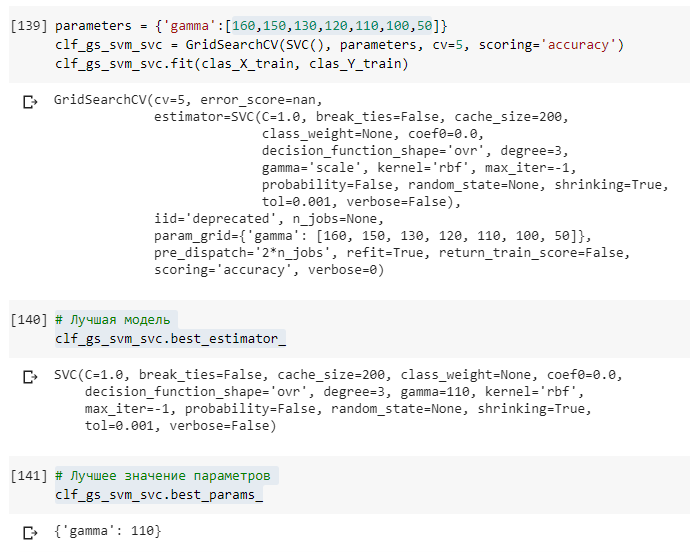
**Подбираем гиперпараметры для выбранных моделей**

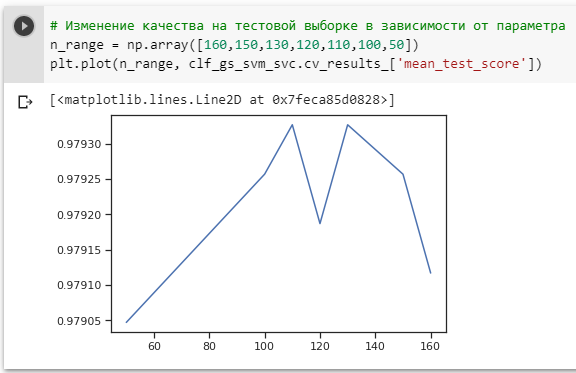
Кросс-валидация

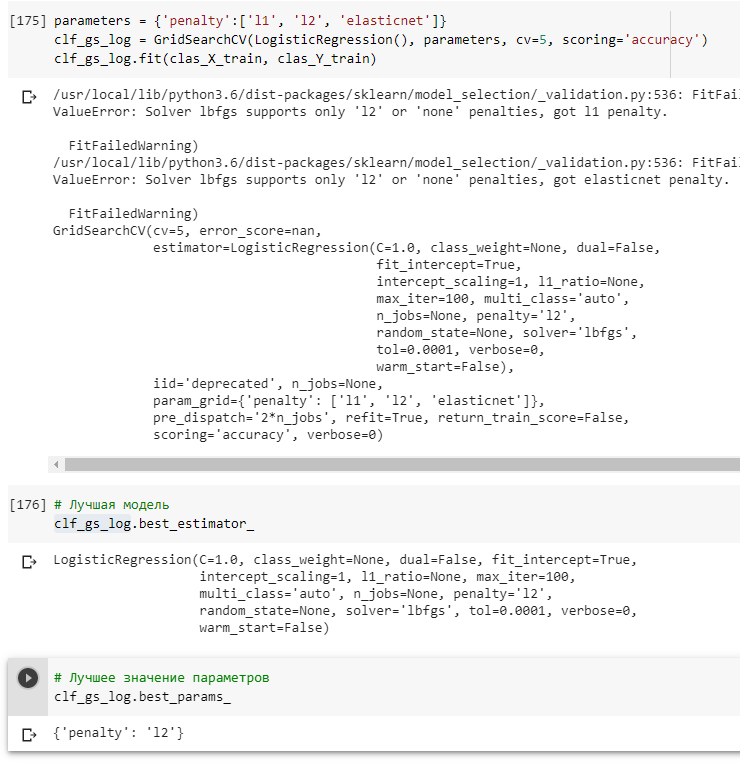


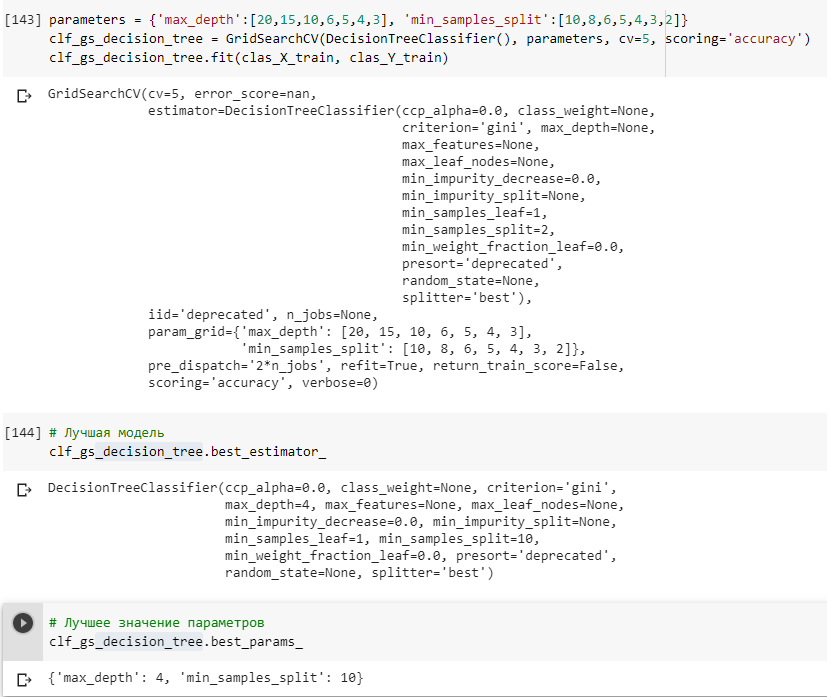


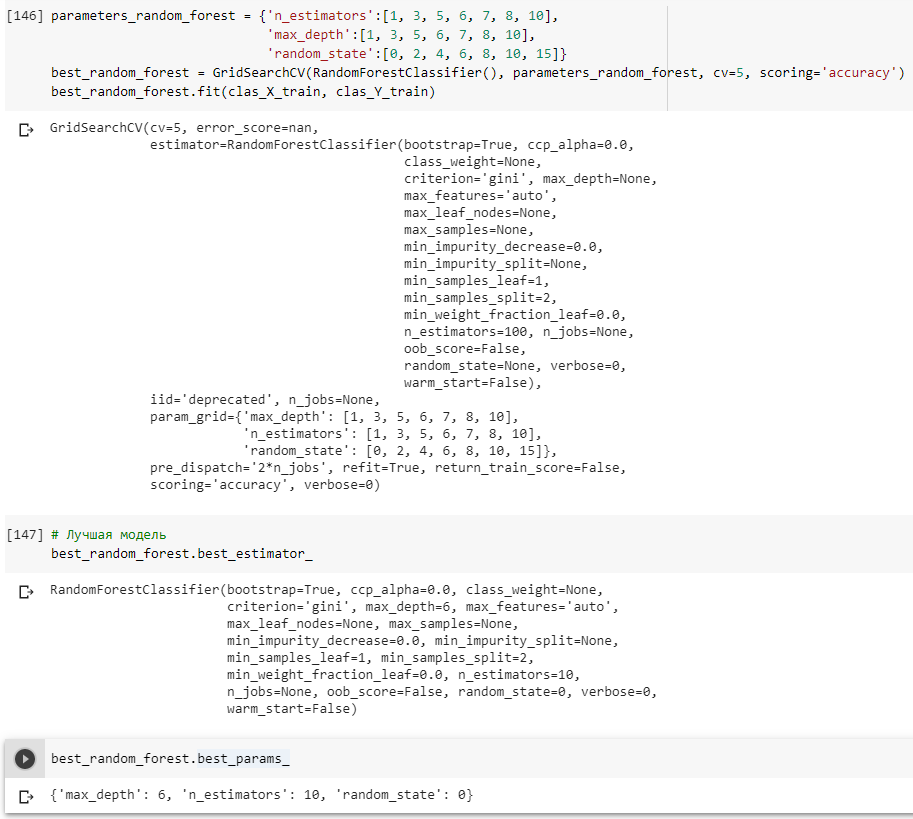
Подбор гиперпараметров:

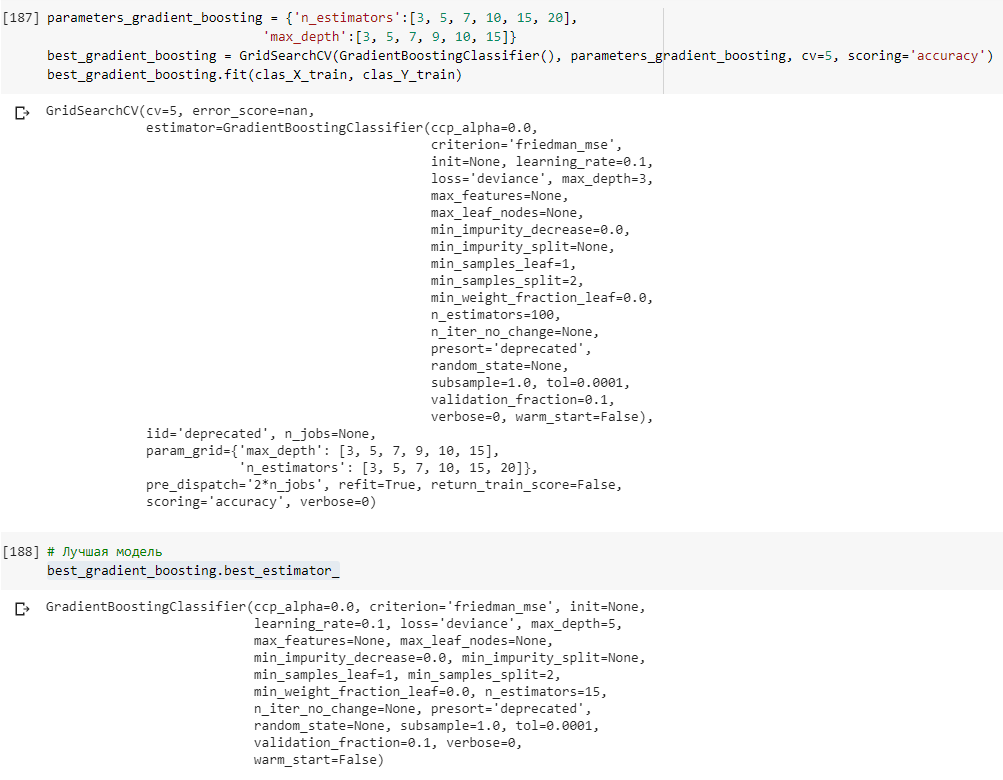


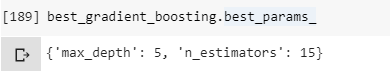




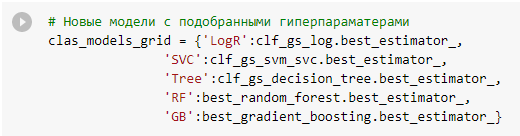


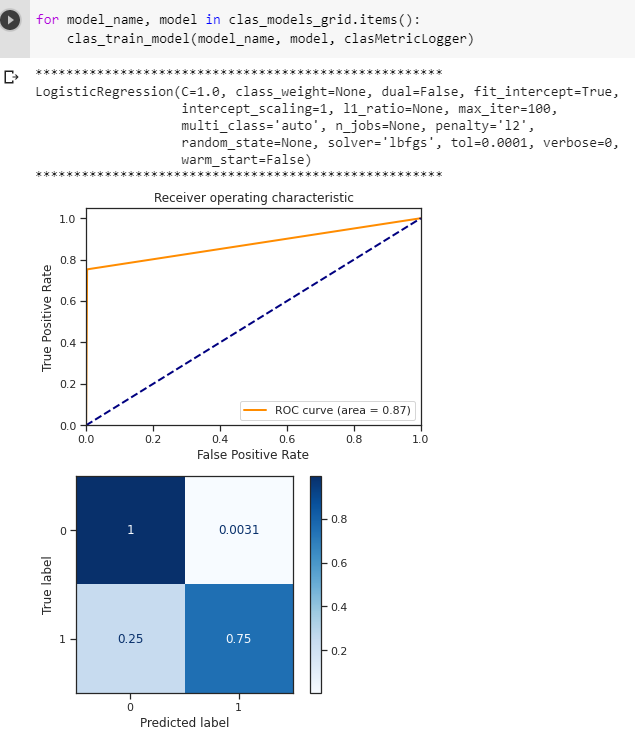


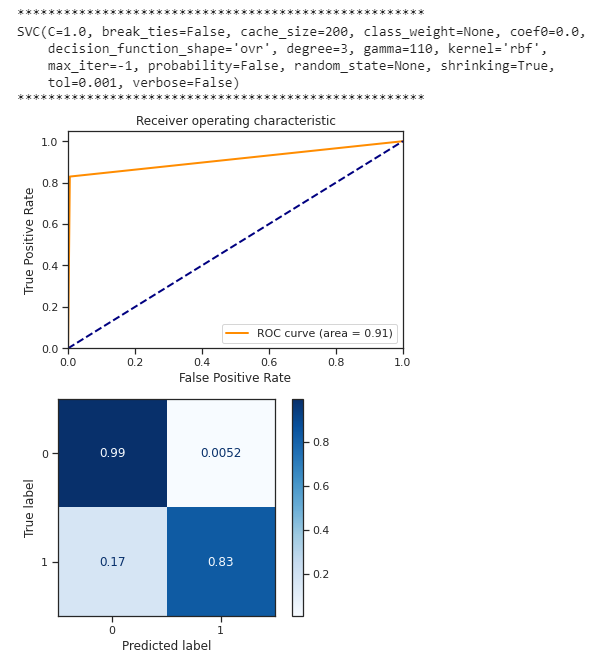


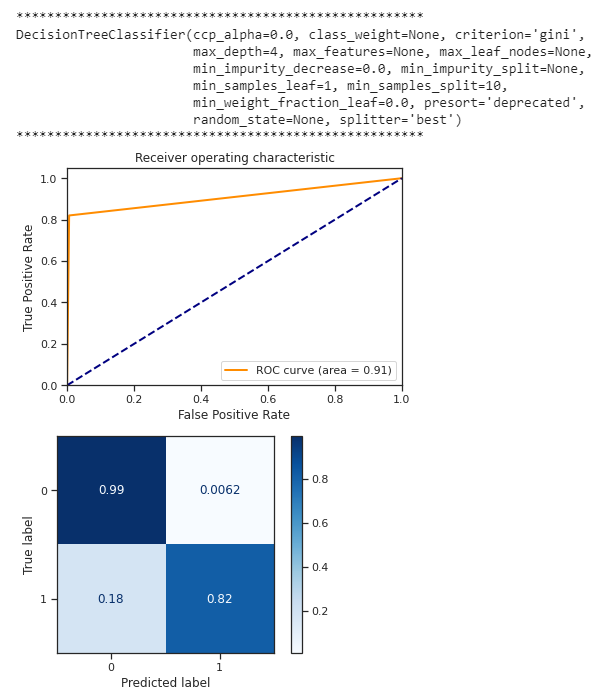


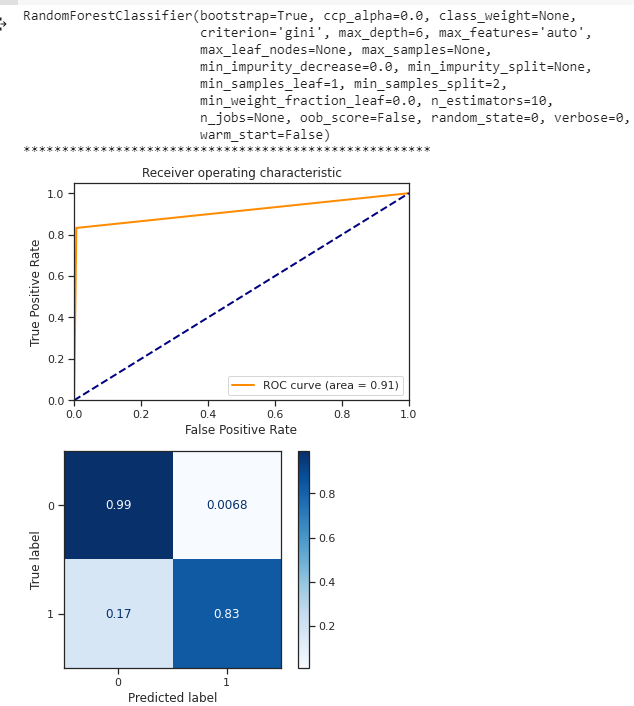
**Повторение для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.**

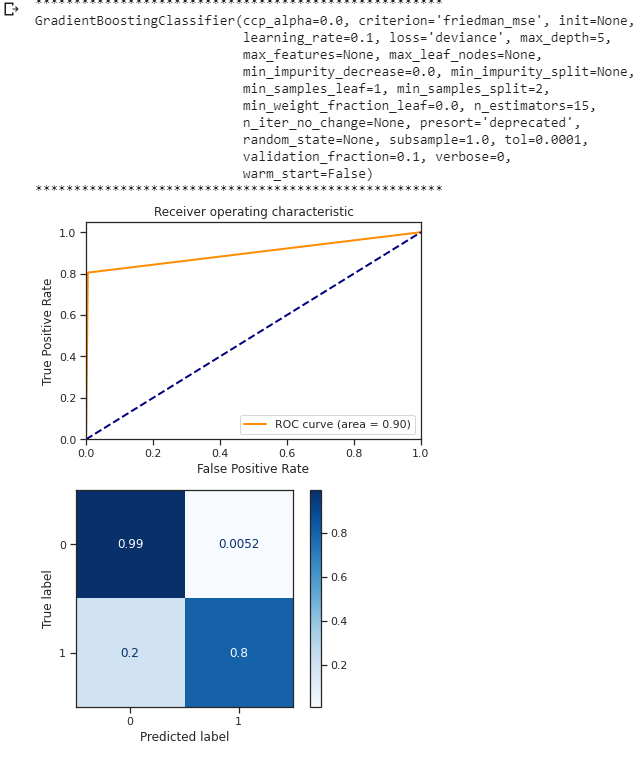








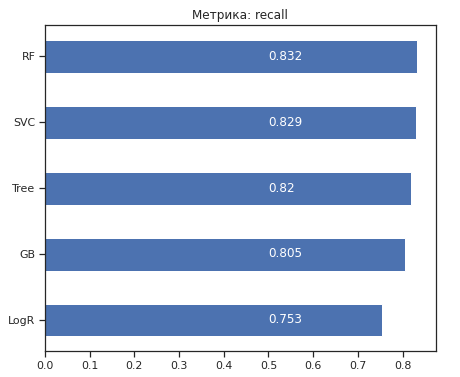




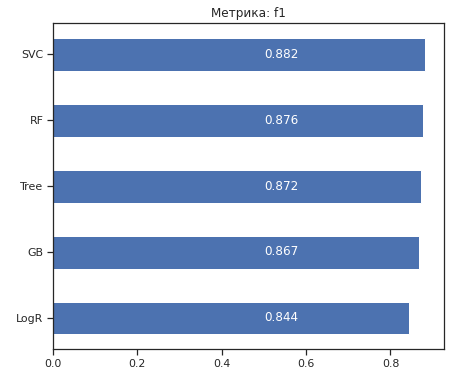
**Формируем выводы о качестве построенных моделей на основе выбранных метрик**



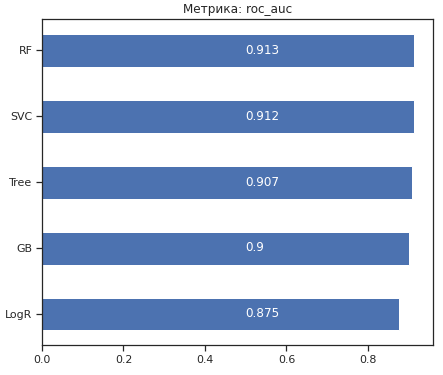
Лучшая модель по метрике precision: Логическая регрессия



Лучшая модель по метрике recall: Случайный лес



Лучшая модель по метрике f1: Метод опорных векторов



Лучшая модель по метрике ROC AUC: Случайный лес

Результат: на основании двух метрик из четырех используемых, лучшей оказалась модель случайного леса. Как известно Random forest борется с переобучением модели, следовательно можно сделать вывод о том, что датасет является довольно разрозненным, поэтому другие методы могли привести к возникновению проблемы переобучения, а "Random forest" успешно обошёл эту проблему.

# **Список литературы**

1. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. Лекции по теории машинного обучения. Ю.Е. Гапанюк [Электронный ресурс]. – Электрон. дан. - URL: <https://github.com/ugapanyuk/ml_course_2020/wiki/COURSE_TMO> (дата обращения: 25.12.2022)
2. Predicting a Pulsar Star [Электронный ресурс]. – Электрон. дан. - URL: <https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star> (дата обращения: 25.12.2022)
3. Машинное обучение (часть 1). А.М.Миронов [Электронный ресурс]. – Электрон. дан. - URL:

<http://www.intsys.msu.ru/staff/mironov/machine_learning_vol1.pdf> (дата обращения: 25.12.2022)

1. Scikit learn[Электронный ресурс]. – Электрон. дан. - URL: <https://scikit-learn.org/stable/index.html> (дата обращения: 25.12.2022)