|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное автономное 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:***

|  |
| --- |
|  |
| ***Обработка набора данных*** |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ5-63Б |  |  |  | В.А. Беспалова |
|  | (группа) |  | (подпись, дата) |  | (И.О. Фамилия) |
| Руководитель НИР |  |  |  |  | Ю.Е. Гапанюк |
|  |  |  | (подпись, дата) |  | (И.О. Фамилия) |
|  |  |  |  |  |  |

*2025 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное автономное образовательное учреждение**

**высшего образования**

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

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

|  |  |  |
| --- | --- | --- |
| Утверждаю | | |
|  | | |
| Заведующий кафедрой | | ИУ5 |
|  | | (индекс) |
|  | В.И. Терехов | |
|  | (И.О. Фамилия) | |
| (подпись) |  | |
|  | (дата) | |

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| по теме | | Обработка набора данных | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | ИУ5-63Б | | | |  | | | | | | | |
|  | | | | Беспалова Виктория Андреевна | | | | | | | | | | | |
| Направленность НИР (учебная, исследовательская, практическая, производственная, др.) | | | | | | | | | | | | | | | |
| ИССЛЕДОВАТЕЛЬСКАЯ | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| Источник тематики (кафедра, предприятие, НИР) | | | | | | | | | | КАФЕДРА | | | | | |
| График выполнения НИР: | | | | | |  | | | | | | | | | |
| 25% к |  | | нед., 50% к | | |  | нед., 75% к | |  | | | нед., 75% к |  | нед |
| ***Техническое задание:*** | | | | | решение задачи машинного обучения на основе материалов | | | | | | | | | | |
| дисциплины. Выбор датасета, первичный анализ, выбор метрик для оценки качества моделей, | | | | | | | | | | | | | | | |
| построение базового решения, оценка качества, подбор гиперпараметров. | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***Оформление научно-исследовательской работы:*** | | | | | | | | | | |  | | | | |
| Расчетно-пояснительная записка на \_\_\_\_\_ листах формата А4. | | | | | | | | | | | | | | | |
| Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.) | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |

Дата выдачи задания « 07 » \_февраля\_\_\_ 2025 г.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Руководитель НИР** |  |  |  | Ю.Е. Гапанюк |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |
| **Студент** |  |  |  | В.А. Беспалова |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

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

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc197287981)

[1 ПОСТАНОВКА ЗАДАЧИ 5](#_Toc197287982)

[2 АНАЛИЗ ДАННЫХ 7](#_Toc197287983)

[3 ВЫБОР МОДЕЛЕЙ И МЕТРИК ДЛЯ ОЦЕНКИ КАЧЕСТВА 13](#_Toc197287984)

[4 ФОРМИРОВАНИЕ ОБУЧАЮЩЕЙ И ТЕСТОВОЙ ВЫБОРОК 15](#_Toc197287985)

[5 ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ 16](#_Toc197287986)

[6 ПОДБОР ГИПЕРПАРАМЕТРОВ 19](#_Toc197287987)

[7 СРАВНЕНИЕ РЕШЕНИЙ 22](#_Toc197287988)

[8 ВЕБ-ПРИЛОЖЕНИЕ 23](#_Toc197287989)

[ЗАКЛЮЧЕНИЕ 26](#_Toc197287990)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 27](#_Toc197287991)

# ВВЕДЕНИЕ

В условиях ускоренных темпов глобальных изменений, включая урбанизацию, изменение климата и сокращение природных ресурсов, вопрос оценки качества питьевой воды приобретает особую значимость. Определение пригодности воды для употребления представляет собой жизненно важную задачу, которая требует надежных методов анализа и современных технологических решений.

Целью данного исследования является построение и оценка моделей машинного обучения для классификации водных ресурсов на пригодные и непригодные для питья. Исследование основано на анализе выборки данных, включающей различные химические показатели качества воды. Важным аспектом работы является применение методов предварительной обработки данных, таких как очистка данных, обработка пропущенных значений и их масштабирование, что позволяет повысить точность и эффективность последующей классификации.

В рамках введения в аналитику мы проводим разведочный анализ данных для идентификации основных статистических характеристик выборки, выявления аномалий и оценки корреляционных связей между переменными. Дальнейший этап включает в себя тщательный отбор и настройку гиперпараметров алгоритмов машинного обучения.

Ожидается, что результаты данной работы окажут значительное влияние на практику мониторинга качества питьевой воды, предоставят ценную информацию для экологов, специалистов в области здравоохранения и разработчиков политики в области управления водными ресурсами. Особое внимание в исследовании уделяется аспектам устойчивости и точности предложенных моделей, что имеет критическое значение для их практического применения в условиях изменяющейся экологической ситуации.

# 1 ПОСТАНОВКА ЗАДАЧИ

Целью исследования является построение и оценка моделей машинного обучения для прогнозирования пригодности воды для питья на основании признаков, представленных в Таблице 1. Целевой переменной в решаемой задаче бинарной классификации является пригодность воды для питья.

Таблица 1 – Описание полей датасета

| Поле | Описание | Тип данных |
| --- | --- | --- |
| ph | pH воды, колеблется от 0 до 14, представляя меру активности водородных ионов в растворе. | float64 |
| Hardness | Жесткость воды, выраженная в миллиграммах на литр, способность воды осаждать мыло. | float64 |
| Solids | Общее количество растворенных твердых веществ в воде, выраженное в частях на миллион (ppm). | float64 |
| Chloramines | Концентрация хлораминов в воде, измеряемая в частях на миллион (ppm). | float64 |
| Sulfate | Концентрация растворенных сульфатов в воде, указанная в миллиграммах на литр. | float64 |
| Conductivity | Электрическая проводимость воды, измеряемая в микросименс на сантиметр (μS/cm). | float64 |
| Organic\_carbon | Содержание органического углерода в воде, выраженное в частях на миллион (ppm). | float64 |
| Trihalomethanes | Концентрация тригалометанов в воде, измеренная в микрограммах на литр. | float64 |
| Turbidity | Турбидность воды, измеряемая в нефелометрических единицах турбидности (NTU), показатель степени рассеивания света. | float64 |
| Potability | Показатель пригодности воды для питья: 1 — вода пригодна для потребления, 0 — вода непригодна для питья. | int64 |

# 2 АНАЛИЗ ДАННЫХ

**Основные характеристики датасета**

Основные характеристики датасета, такие как: количество строк и столбцов, типы данных полей, уникальные значения целевого признака – приведены на Рисунке 1. Целевой признак является бинарным и содержит только значения 0 и 1. Значит, решается задача бинарной классификации.

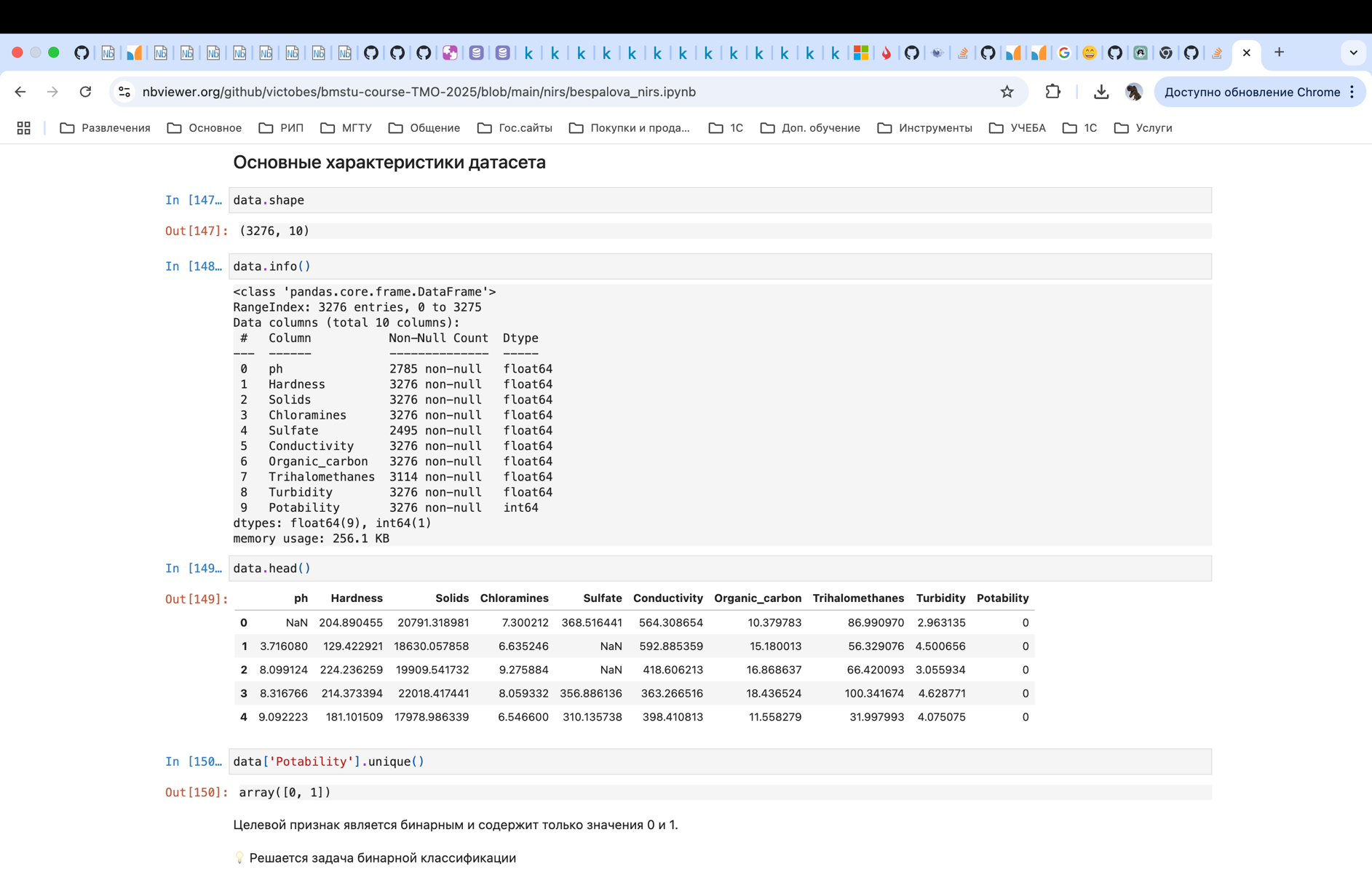


Рисунок 1 – Основные характеристики датасета

В анализируемом наборе данных есть пропуски в колонках «ph», «Sulfate» и «Trihalomethanes» (см. Рисунок 2). Можно отметить, что в строках, относящихся к классу «0» в 2 раза больше пропусков, чем в строках, принадлежащих классу «1» (см. Рисунок 3).

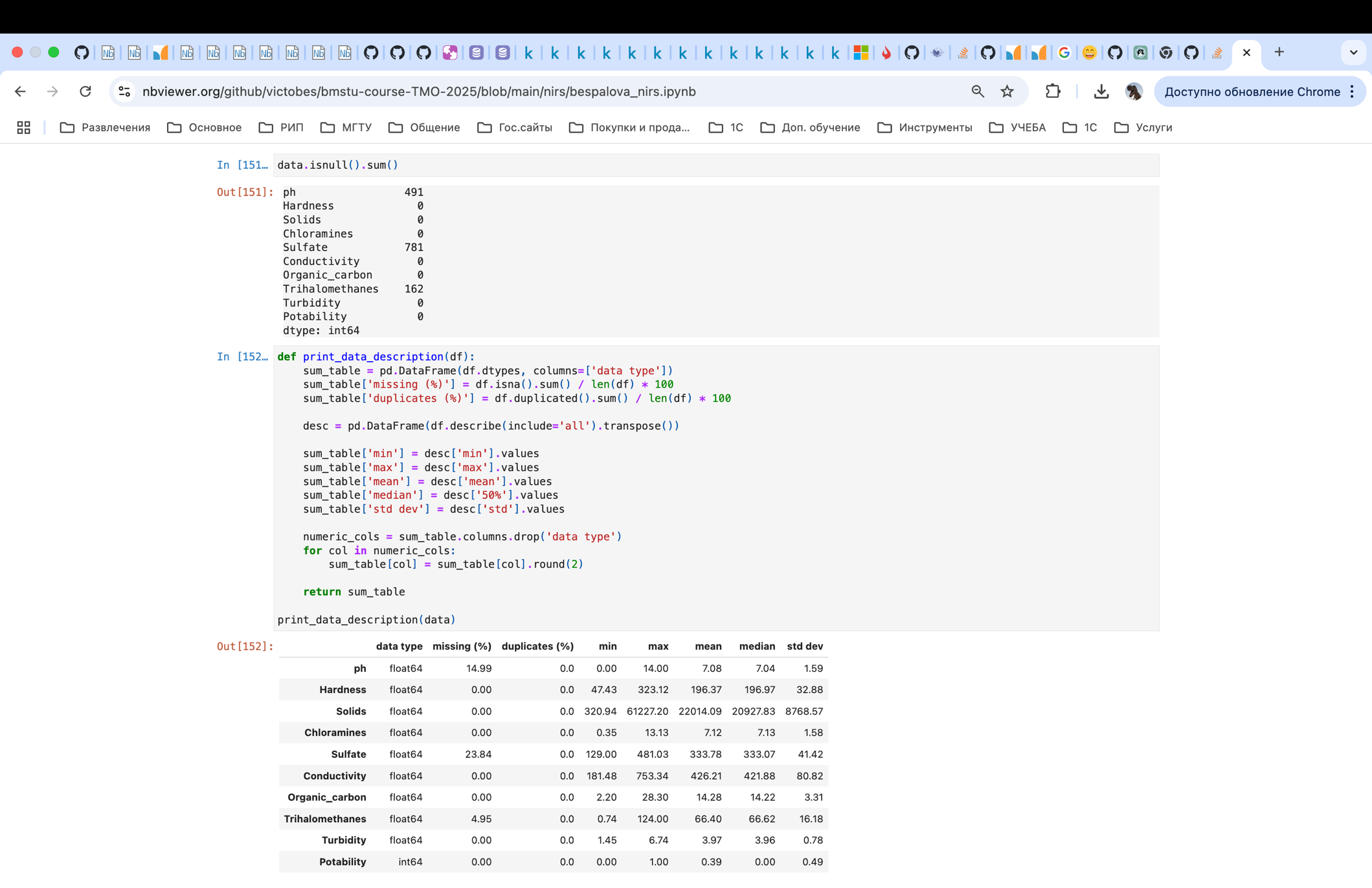


Рисунок 2 – Количество пропусков данных по столбцам

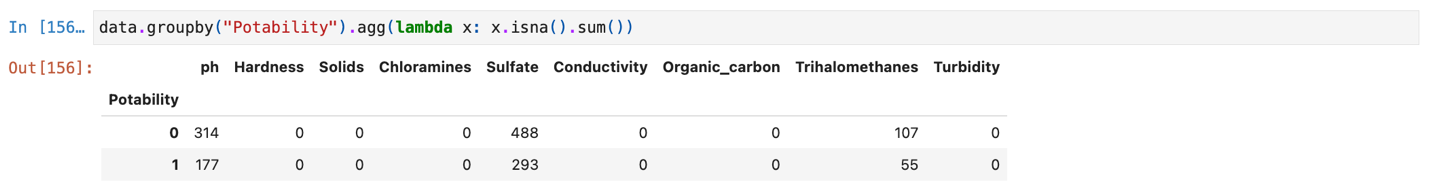


Рисунок 3 – Количество пропусков данных по столбцам (детализация)

На Рисунке 4 представлены основные характеристики столбцов датасета.

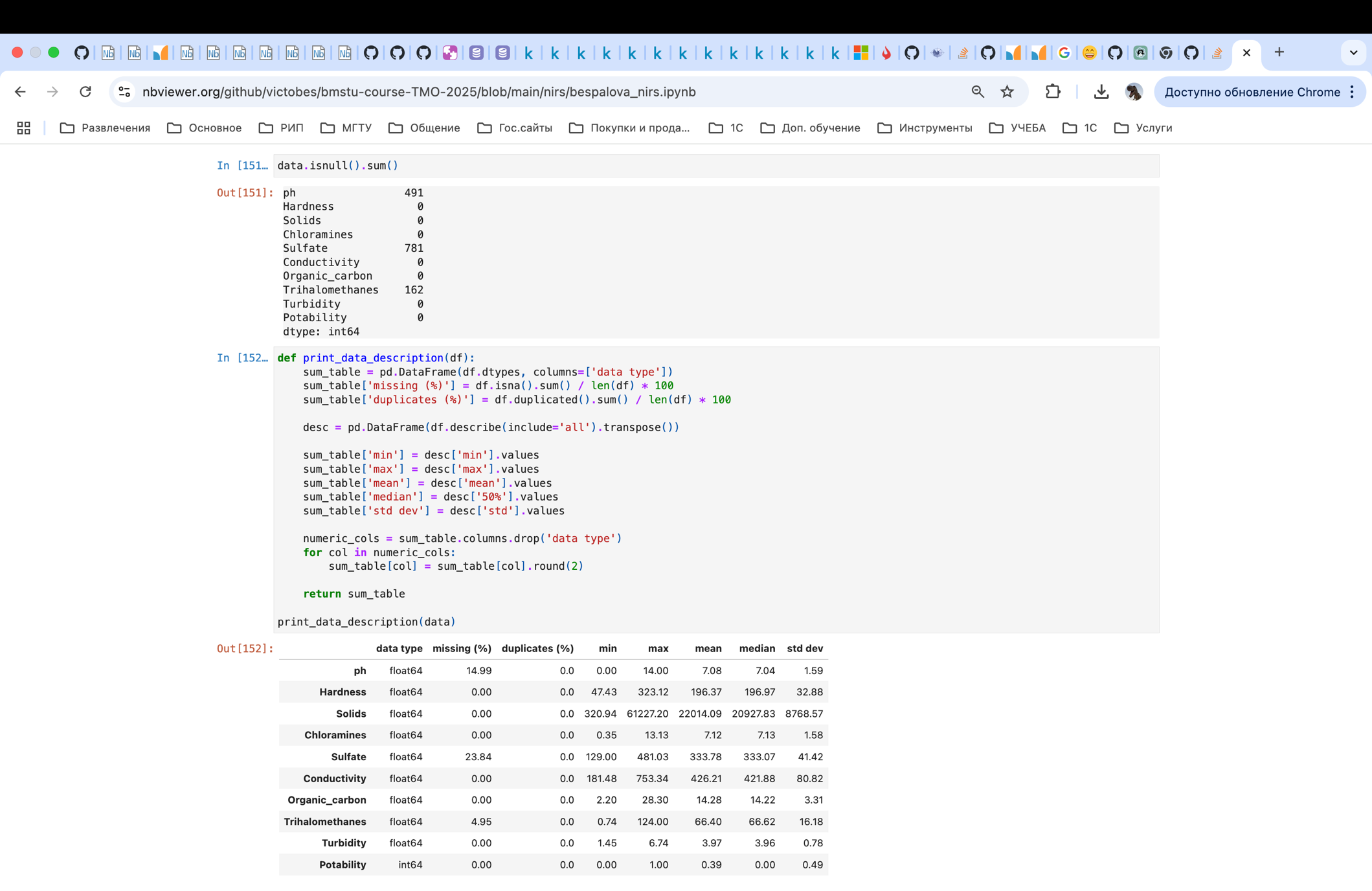


Рисунок 4 – Основные характеристики столбцов датасета

Можем отметить, что пропуски данных составляют 15%, 5% и около 25% в колонках «ph», «Trihalomethanes» и «Sulfate» соответственно. Кроме этого, не наблюдается нереальных с физической точки зрения значений величин.

**Баланс классов**

Из Рисунка 5 видно, что присутствует незначительный дисбаланс классов. В наборе данных 39% принадлежат к классу «1» (1278 значений), т.е. пригодной для питья воды и 61% – к классу «0» (1998 значений), т.е. не пригодной для питья воды.

Дисбаланс действительно ожидаем в данных о качестве воды в реальном мире, поскольку вода в необработанных образцах чаще оказывается не пригодной для питья, чем пригодной.

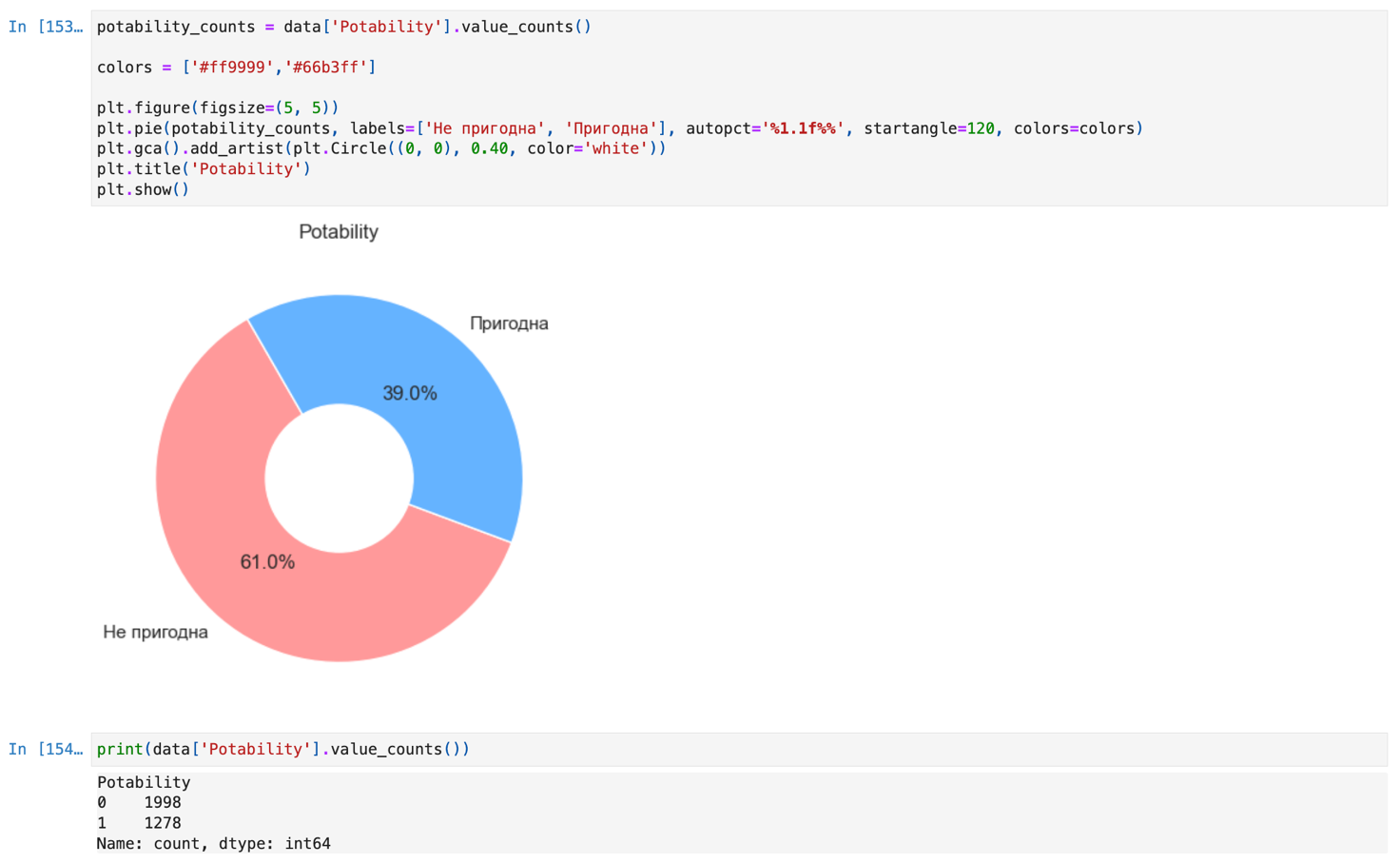


Рисунок 5 – Баланс классов

**Визуальный анализ датасета**

Распределение значений признаков проиллюстрировано на Рисунке 6. Отметим, что большинство признаков имеют распределение близкое к нормальному. Показатели такие как «Solids» и «Conductivity» демонстрируют положительную асимметрию.

Все признаки содержат выбросы, однако сохранение выбросов оправдано, т.к. они представляют собой возможные крайние значения измерений качества воды.

Признаки измеряются в значительно разных масштабах. Значит, необходимо провести масштабирование данных.

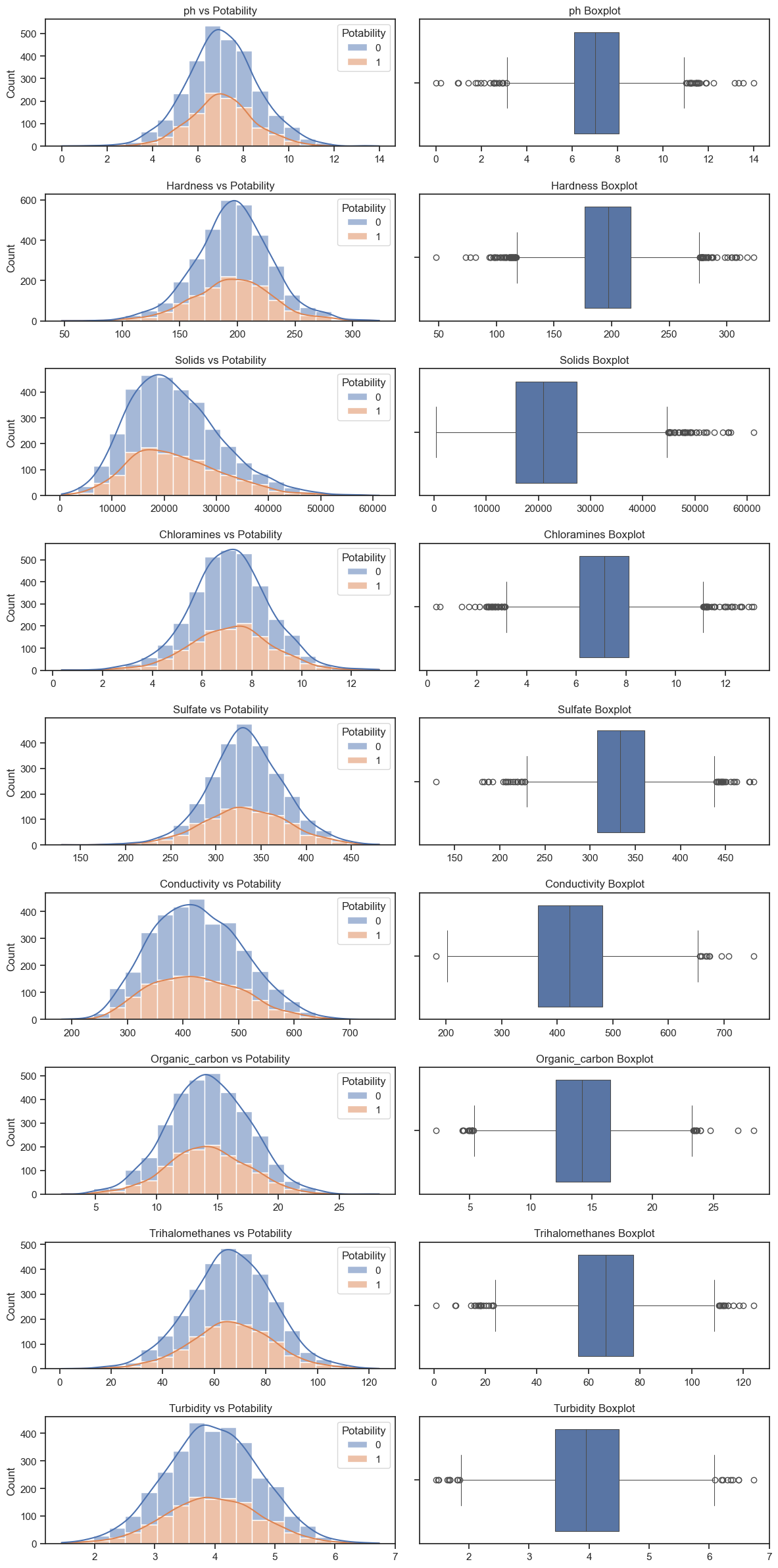


Рисунок 6 – Распределение значений признаков

**Заполнение пропусков**

Перед заполнением пропусков сравним средние и медианные значения для колонок, содержащих пропуски. Заметим, что медианные и средние значения для колонок, содержащих пропуски, близки (Рисунок 7).

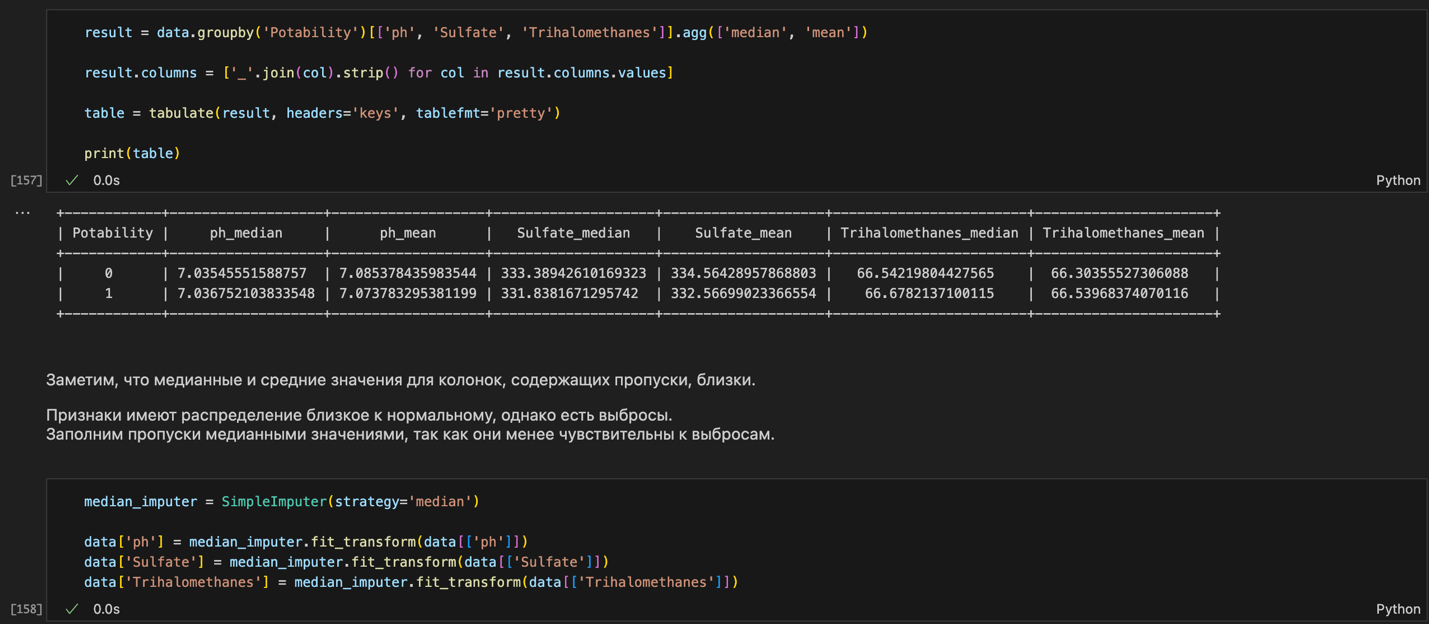


Рисунок 7 – Медианные и средние значения для столбцов с пропусками

Признаки имеют распределение близкое к нормальному, однако есть выбросы. Заполним пропуски медианными значениями, так как они менее чувствительны к выбросам (см. Рисунок 8). Затем убедимся в том, что пропусков действительно не осталось.

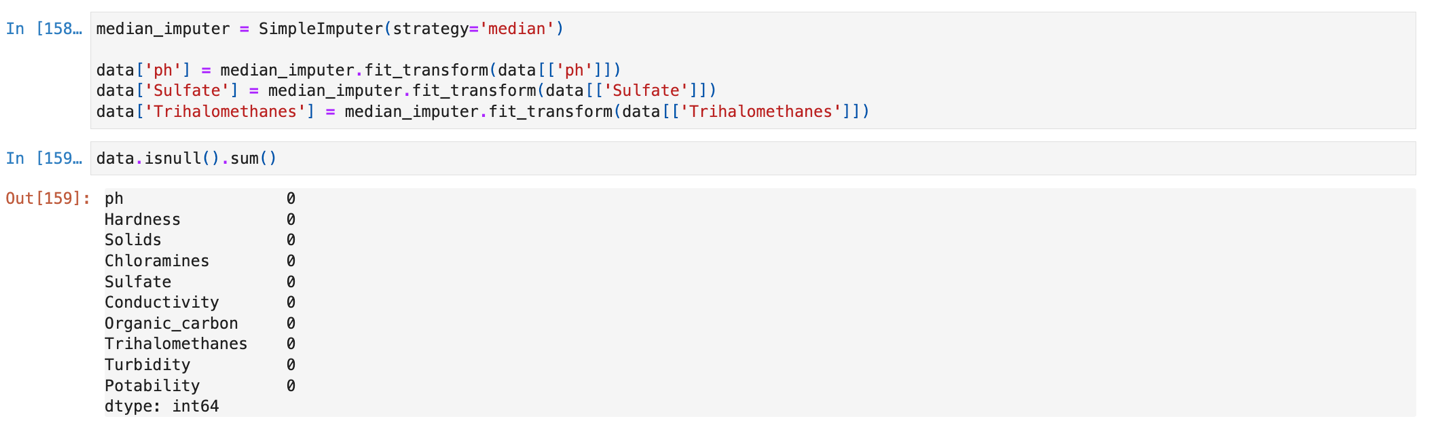


Рисунок 8 – Заполнение пропусков

**Информация о корреляции признаков**

На Рисунке 9 представлена корреляционная матрица. Признаки не имеют корреляции между собой, а, следовательно, ни один из них не является избыточным. Целевой признак не имеет корреляции с остальными признаками.

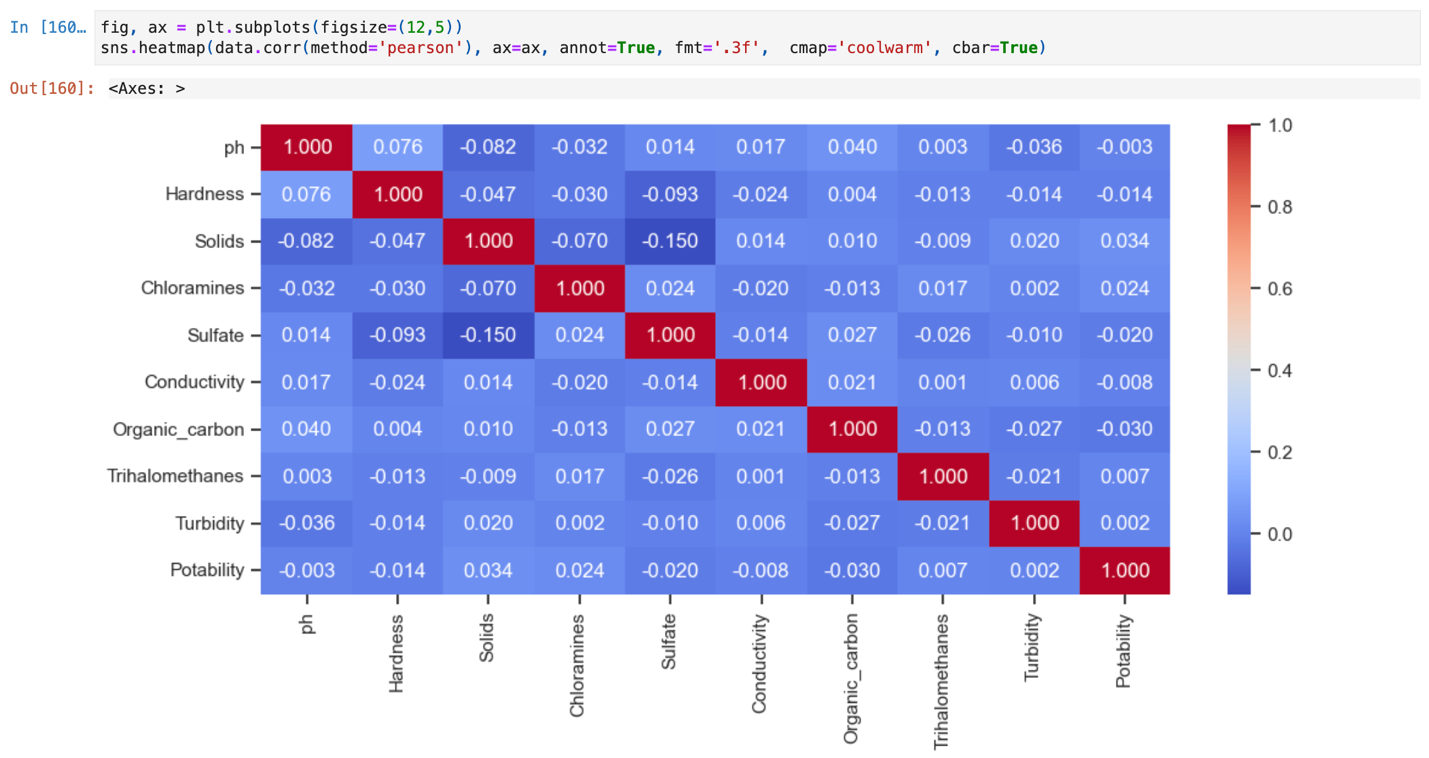


Рисунок 9 – Корреляционная матрица

# 3 ВЫБОР МОДЕЛЕЙ И МЕТРИК ДЛЯ ОЦЕНКИ КАЧЕСТВА

**Выбор метрик для оценки качества моделей**

1. ***Accuracy***

Метрика вычисляет процент (долю в диапазоне от 0 до 1) правильно определенных классов.

1. ***ROC-кривая и ROC AUC***

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.

TPR содержит в знаменателе количество истинных 1.

FPR содержит в знаменателе количество истинных 0.

Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1), то есть через верхний левый угол графика. 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.

1. ***Матрица ошибок***

Количество верно и ошибочно классифицированных данных, представленное в виде матрицы.

В случае бинарной классификации матрица ошибок выглядит как представлено в Таблице 2.

Таблица 2 – Матрица ошибок в случае бинарной классификации

| Предсказанное/истинное значение | y=1 | y=0 |
| --- | --- | --- |
|  | True Positive (TP) | False Positive (FP) |
|  | False Negative (FN) | True Negative (TN) |

1. ***F1-мера***

Объединяет precision и recall в единую метрику.

Precision – доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор верно или неверно определил как положительные.

Recall – доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор верно или неверно определил как положительные.

**Выбор моделей для решения задачи бинарной классификации**

С учетом следующих условий:

* Требование задания НИРС: не менее 5 моделей, не менее 2 ансамблевых моделей;
* Решение задачи бинарной классификации;
* Датасет содержит только числовые признаки;

выберем модели:

* Логистическая регрессия;
* Метод опорных векторов;
* Дерево решений;
* Случайный лес;
* Градиентный бустинг.

# 4 ФОРМИРОВАНИЕ ОБУЧАЮЩЕЙ И ТЕСТОВОЙ ВЫБОРОК

Разделим выборку на обучающую и тестовую с помощью train\_test\_split (см. Рисунок 10).



Рисунок 10 – Разделение выборки

Как было установлено ранее, признаки имеют очень разные масштабы измерения, поэтому проведем масштабирование (см. Рисунок 11). Масштабирование предполагает изменение диапазона измерения величины.

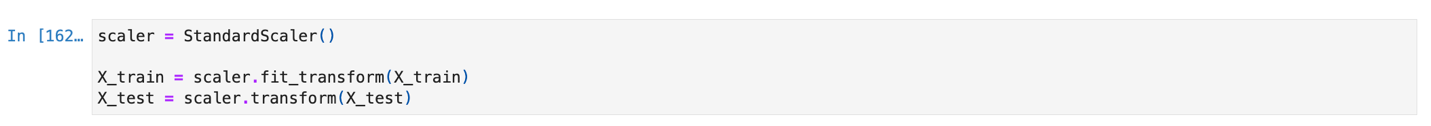


Рисунок 11 – Масштабирование данных

# 5 ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ

Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки.

**Обучение моделей**

Обучим 5 выбранных моделей без подбора гиперпараметров (см. Рисунок 12). Используем модели из библиотек scikit-learn и xgboost.

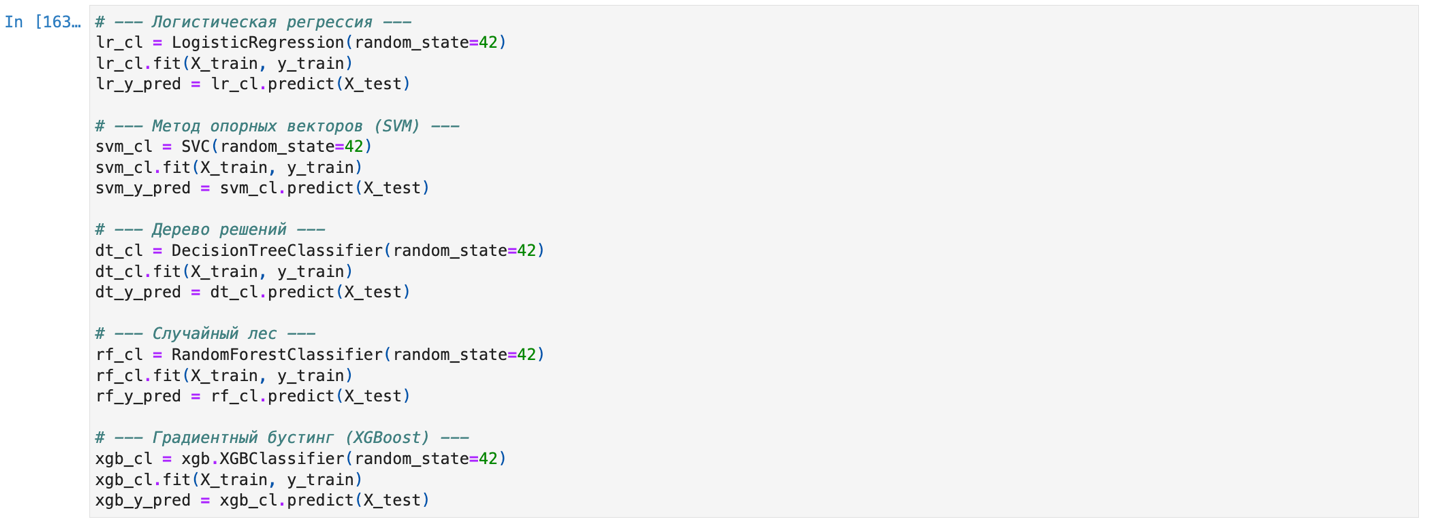


Рисунок 12 – Обучение моделей без подбора гиперпараметров

**Оценка качества моделей**

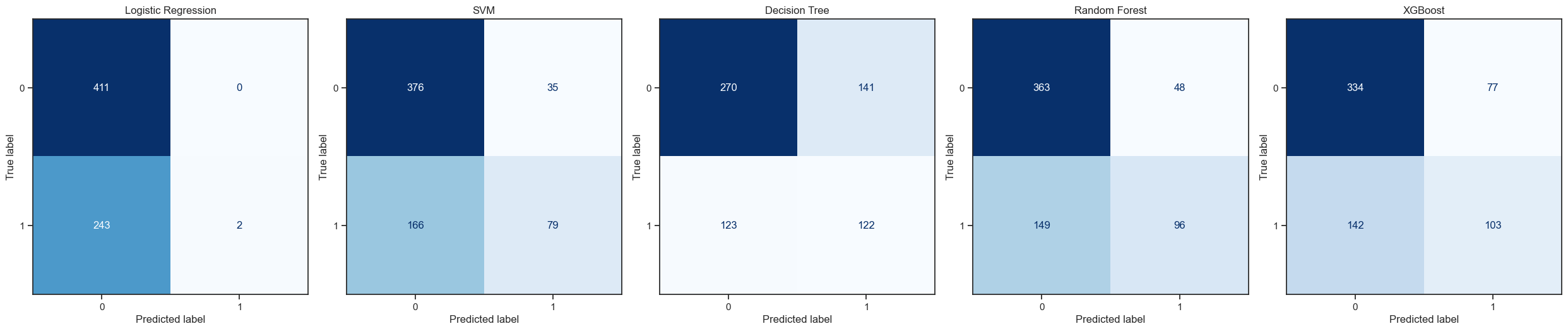
Оценим качество работы моделей. На Рисунке 13 приведены 3 метрики: Accuracy, Balanced Accuracy и F1-мера. Добавление метрики Balanced Accuracy, вычисляемой с помощью balanced\_accuracy\_score, обусловлено важностью определения точности предсказания для каждого класса.



Рисунок 13 – Оценка качества работы моделей baseline-решения

Лучшие результаты показала модель Случайного леса: Balanced Accuracy – 0,64 и F1-мера – 0,49. Худшие результаты у модели Логистической регрессии: Balanced Accuracy – 0,5 и F1-мера – 0,02.

На Рисунке 14 представлены матрицы ошибок для каждой модели базового решения.



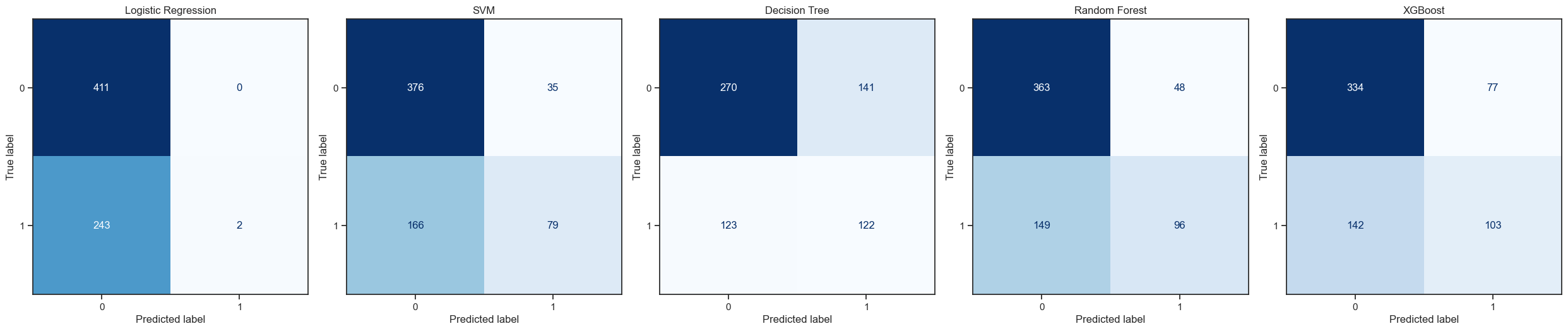


Рисунок 14 – Матрица ошибок для baseline-решения

ROC-кривая и ROC AUC приведены на Рисунке 15.



Рисунок 15 – ROC-кривая и ROC AUC для baseline-решения

Мы видим, что модель Случайного леса показала лучшее качество классификации. Модели Метода опорных векторов и Градиентного бустинга имеют схожие оценки.

# 6 ПОДБОР ГИПЕРПАРАМЕТРОВ

Производится подбор гиперпараметров для выбранных моделей.

**Обучение моделей**

Обучим выбранные модели, подбирая гиперпараметры с помощью GridSearchCV и кросс-валидации по методу K-Fold (см. Рисунок 16).

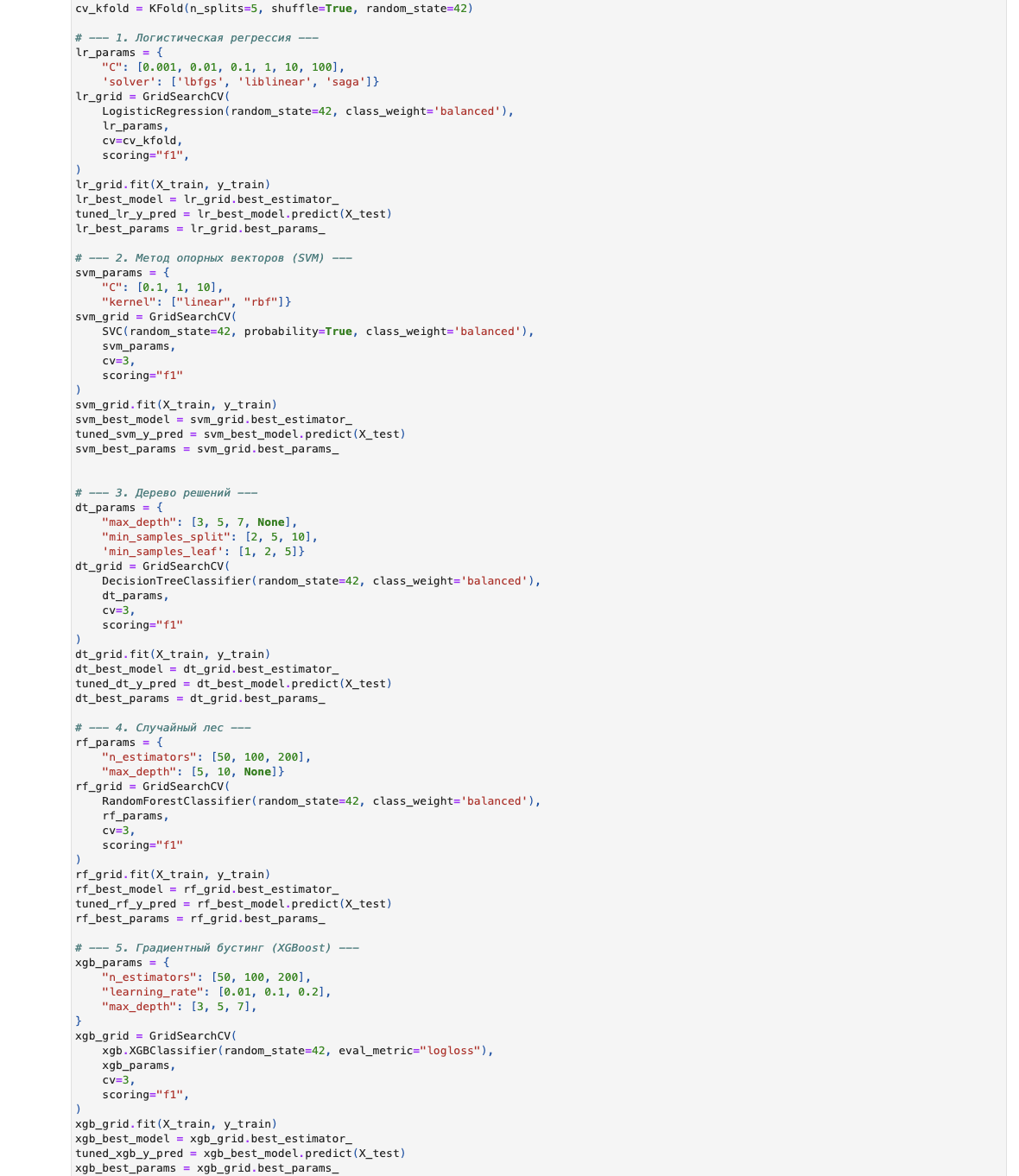


Рисунок 16 – Обучение моделей с подбором гиперпараметров

**Оценка качества моделей**

Оценим качество работы моделей с подбором гиперпараметров. На Рисунке 17 приведены 3 метрики: Accuracy, Balanced Accuracy, F1-мера – и подобранные гиперпараметры.

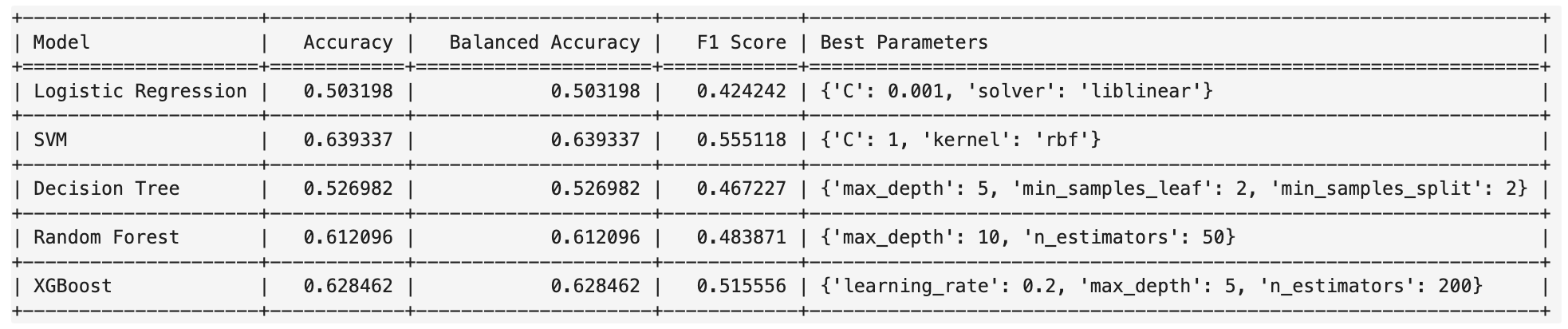
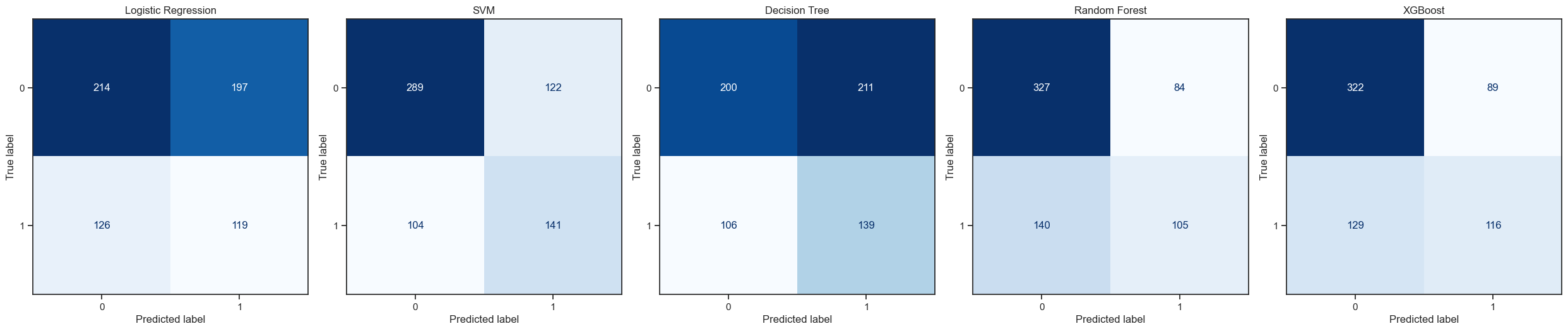


Рисунок 17 – Оценка качества работы моделей оптимального решения

На Рисунке 18 представлены матрицы ошибок для каждой модели решения с подбором гиперпараметров.



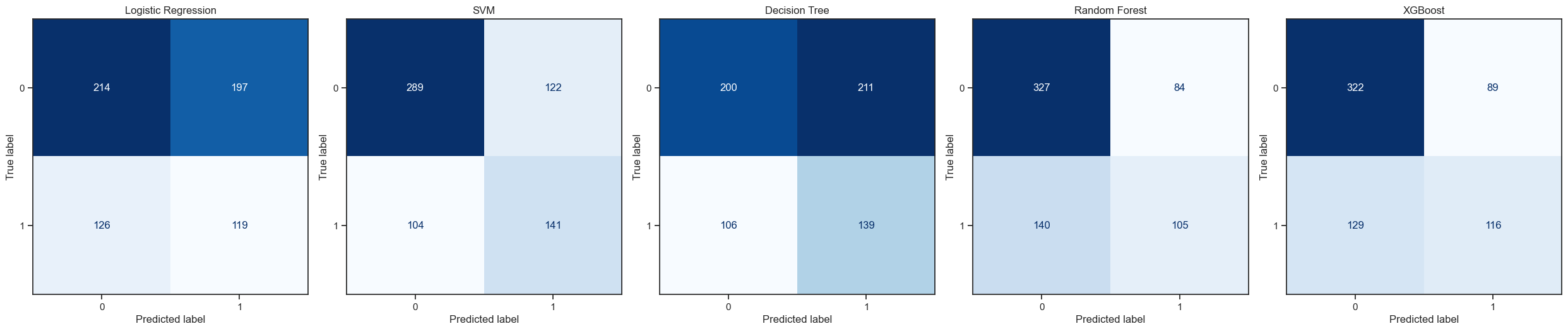


Рисунок 18 – Матрица ошибок для оптимального решения

ROC-кривая и ROC AUC изображены на Рисунке 19.

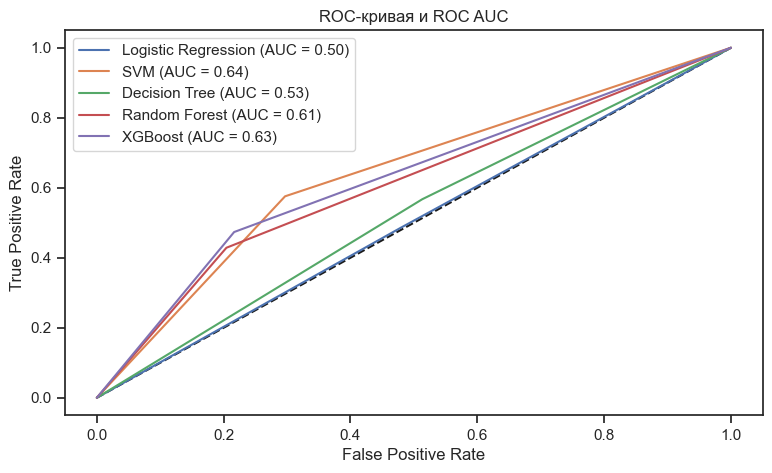


Рисунок 19 – ROC-кривая и ROC AUC для оптимального решения

Мы видим, что лучшие результаты показывает классификатор Метода опорных векторов. Подбор гиперпараметров не позволил значительно улучшить качество работы моделей.

# 7 СРАВНЕНИЕ РЕШЕНИЙ

На Рисунке 20 представлены столбчатые диаграммы позволяющие сравнить Balanced Accuracy, AUC ROC и F1-меру моделей базового и оптимального решения.

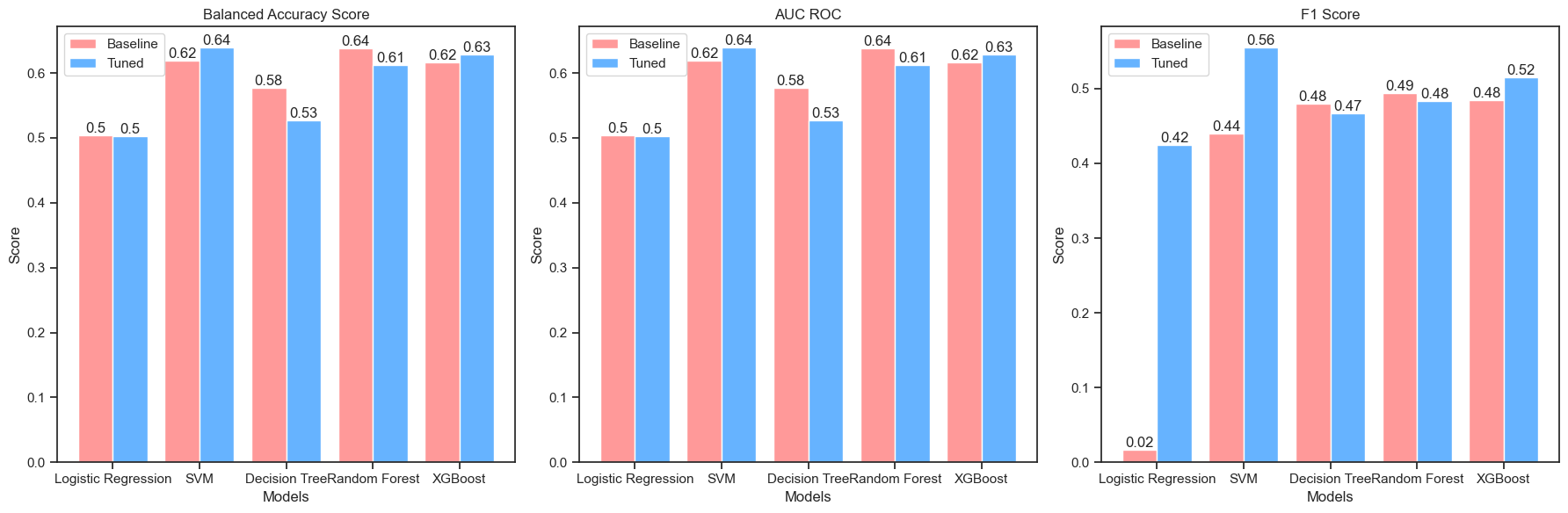


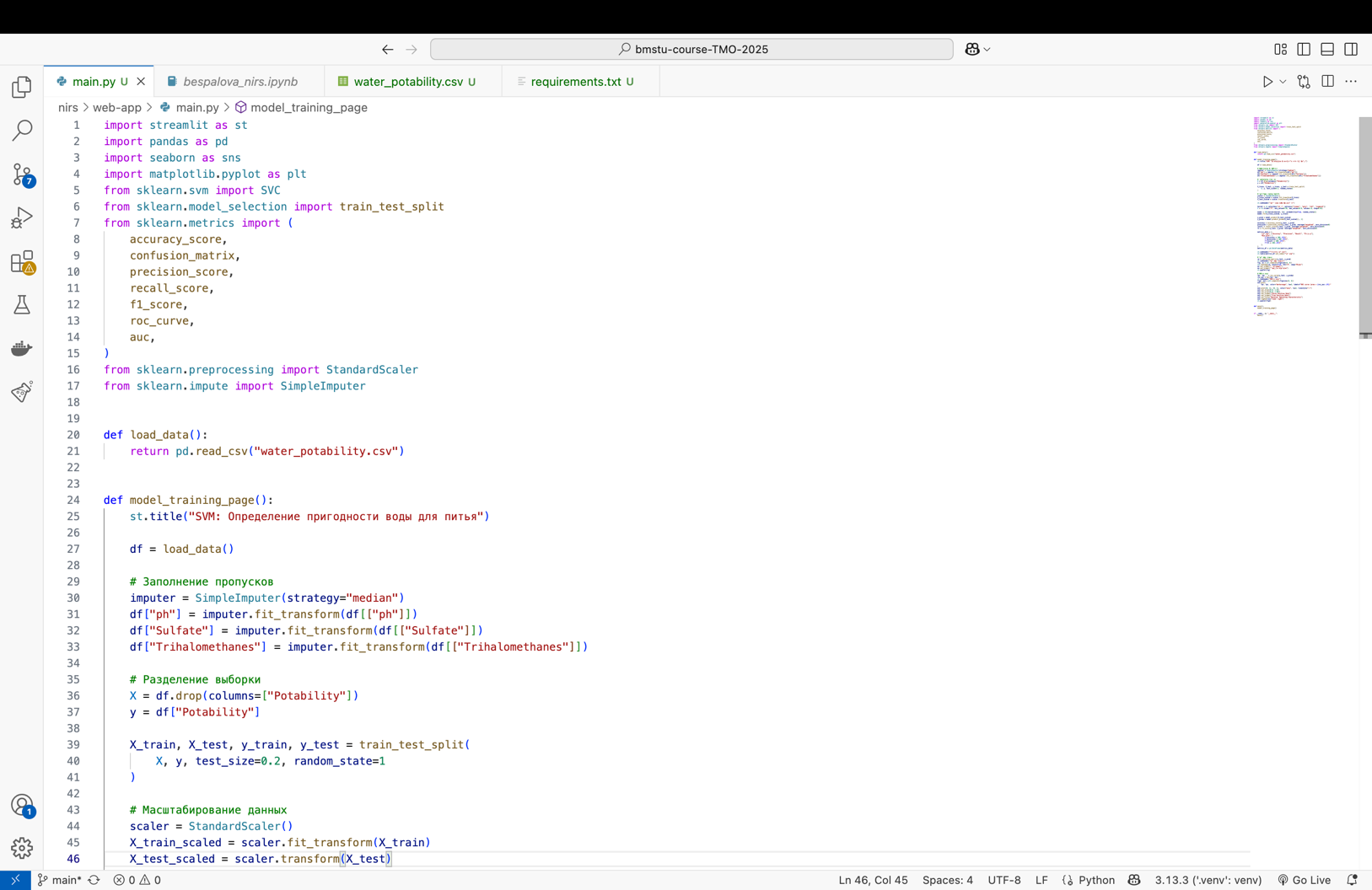
Рисунок 20 – Сравнение решений

Подбор гиперпараметров проводился с учетом того, что F1-мера приоритетнее при оценке качества работы классификаторов, так как она не зависит от соотношения классов. Стоит отметить, что у моделей Деревья решений и Случайного леса результаты ухудшились после подбора гиперпараметров.

Мы видим, что модель Метода опорных векторов с подобранными гиперпараметрами и модель Случайного леса без подбора гиперпараметров показали наивысшие результаты: Balanced Accuracy – 0.64, AUC ROC – 0.64 и F1-мера – 0.56 и 0.49 соответственно.

# 8 ВЕБ-ПРИЛОЖЕНИЕ

Создадим веб-приложение для демонстрации модели Метода опорных векторов (SVM) с помощью фреймворка streamlit. На Рисунке 21 приведен код реализации.





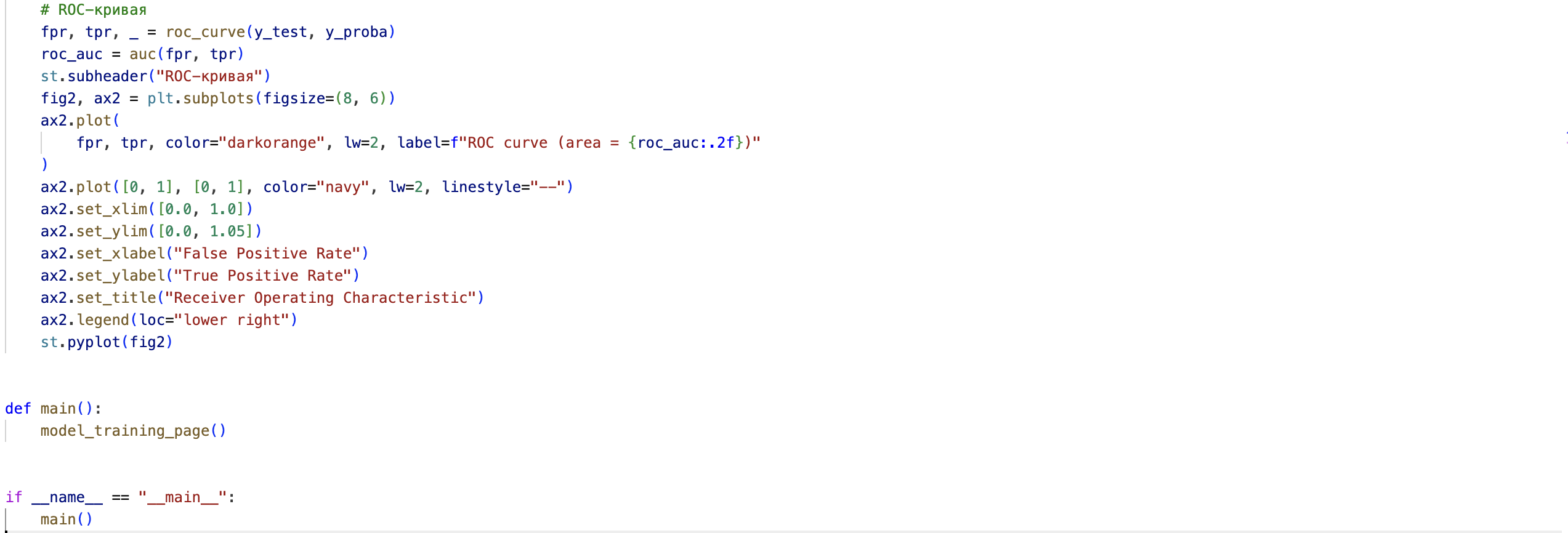
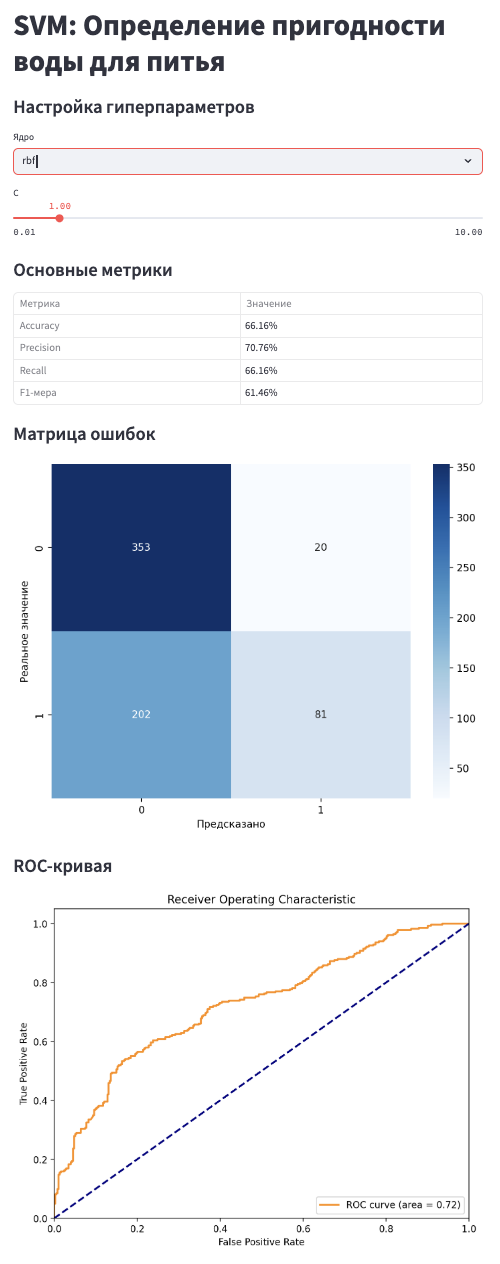


Рисунок 21 – Реализация веб-приложения

У пользователя есть возможность изменения двух гиперпараметров модели: ядра и C (см. Рисунок 22). При изменении гиперпараметров обновляются разделы: «Основные метрики», «Матрица ошибок» и «ROC-кривая».



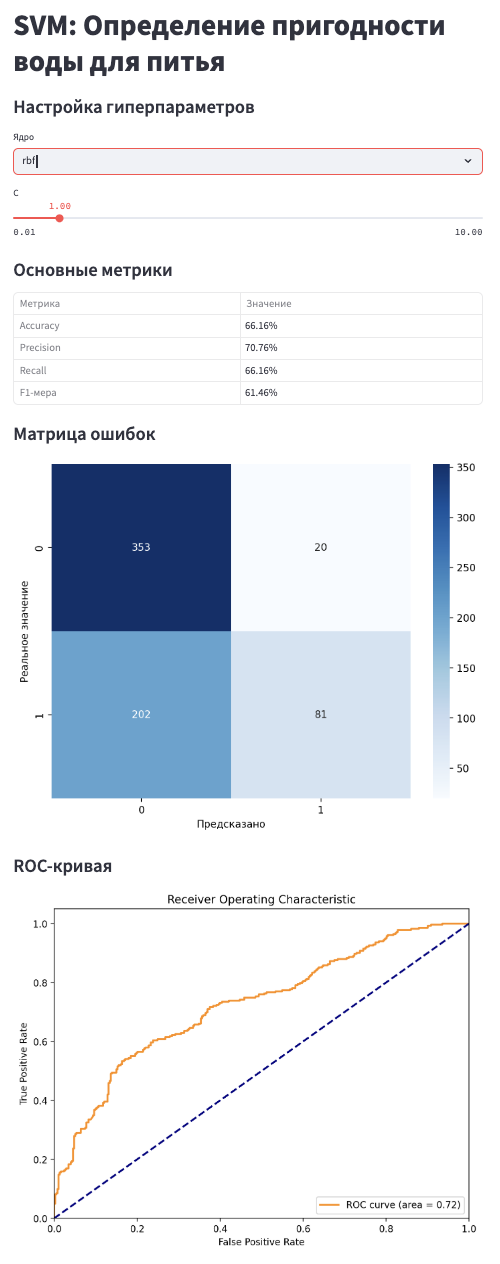


Рисунок 22 – Интерфейс веб-приложения

# ЗАКЛЮЧЕНИЕ

В рамках научно-исследовательской работы была решена задача бинарной классификации – определение пригодности воды для питья. В ходе исследования были получены два решения: базовое и оптимальное за счет подбора гиперпараметров.

Наилучшего качества классификации по выбранным метрикам достигла модель Метода опорных векторов после подбора гиперпараметров.

Для демонстрации влияния гиперпараметров на результаты работы модели было разработано веб-приложение.

Построенные модели показывают недостаточно хорошие результаты классификации для применения в реальной жизни, так как ошибка при определении пригодности воды для питья может стать источником страшных последствий. Тем не менее, результаты исследования могут служить основой для дальнейших научных и прикладных работ в области изучения качества воды с помощью машинного изучения.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Дьяконов А. Блог Александра Дьяконова [Электронный ресурс]. URL: https://alexanderdyakonov.wordpress.com/ (дата обращения: 03.05.2025).

2. Streamlit. Документация Streamlit [Электронный ресурс]. URL: https://streamlit.io/ (дата обращения: 03.05.2025).

3. Открытый курс машинного обучения. Тема 10: Градиентный бустинг [Электронный ресурс]. URL: https://habr.com/ru/companies/ods/articles/327250/ (дата обращения: 03.05.2025).

4. scikit-learn. Supervised learning [Электронный ресурс]. URL: https://scikit-learn.org/stable/supervised\_learning.html (дата обращения: 03.05.2025).

5. Гапанюк Ю. Е. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр [Электронный ресурс]. URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2025/ (дата обращения: 03.05.2025).

6. Метрики в задачах машинного обучения [Электронный ресурс]. URL: https://habr.com/ru/companies/ods/articles/328372/ (дата обращения: 03.05.2025).