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:

Предсказание ядовитости грибов с применением машинного обучения

Студент __ИУ5-62Б_____ _____ **Д.О. Щепетов**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Ю.Е. Гапанюк**
(Подпись, дата) (И.О.Фамилия)

2024 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 ___ИУ5____
(Индекс)
_____ ___В.И. Терехов__
(И.О.Фамилия)
«_07_» ____февраля_____ 2024 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме ____Предсказание ядовитости грибов с применением машинного обучения_____

Студент группы __ИУ5-62Б_____

_____Щепетов Дмитрий Олегович_____
(Фамилия, имя, отчество)

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., 100% к ___ нед.

Техническое задание _____
_____Исследовать методы машинного обучения для решения задачи классификации_____
_____
_____
_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __25__ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
_____
_____
_____

Дата выдачи задания « 07 » февраля 2024 г.

Руководитель НИР _____ ___Ю.Е. Гапанюк_____
(Подпись, дата) (И.О.Фамилия)

Студент _____ ____Д.О. Щепетов_____
(Подпись, дата) (И.О.Фамилия)

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

Содержание

Введение

Грибы играют важную роль в экосистеме и служат значимым источником пищи для людей. Однако среди множества видов грибов некоторые могут быть смертельно ядовитыми. Возможность точно и быстро классифицировать грибы на съедобные и ядовитые является критически важной задачей для предотвращения пищевых отравлений и обеспечения безопасности потребителей. В современных условиях, с развитием технологий, машинное обучение предоставляет мощные инструменты для решения задач классификации и предсказания.

Данная работа направлена на разработку и оптимизацию моделей машинного обучения для классификации ядовитости грибов на основе их морфологических характеристик. Для этого используется очищенный набор данных грибов из библиотеки UCI, включающий девять признаков: диаметр крышки, форма крышки, жаберное прикрепление, цвет жабр, высота штока, ширина штока, цвет стебля, время года и целевой класс (съедобно или ядовито).

Целью данной работы является разработка и оптимизация моделей машинного обучения для точной классификации грибов на съедобные и ядовитые на основе их морфологических признаков. Исследование направлено на сравнение эффективности различных алгоритмов классификации, а также на определение оптимальных гиперпараметров для каждой модели. Результаты исследования помогут в создании надежной системы классификации грибов, которая может быть использована для повышения безопасности потребления грибов и предотвращения случаев отравления.

В исследовании используются различные алгоритмы машинного обучения, включая K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest и Gradient Boosting. Для оценки качества моделей применяются метрики точности (accuracy), полноты (recall), F1-скора, точности (precision) и ROC AUC. Особое внимание уделяется подбору гиперпараметров с использованием методов кросс-валидации для достижения оптимальных результатов.

Предлагаемое исследование сочетает в себе передовые методы машинного обучения и современные подходы к обработке данных, что позволяет получить значимые результаты в области классификации грибов. Ожидается, что разработанные модели смогут эффективно различать съедобные и ядовитые грибы, обеспечивая тем самым дополнительный уровень безопасности для потребителей.

Постановка задачи

Данная работа по машинному обучению направлена на решение задачи классификации, а именно, предсказание ядовитости грибов.

Данная работа направлена на создание моделей машинного обучения для классификации грибов на съедобные и ядовитые на основе их морфологических признаков. Исходный набор данных был тщательно обработан, включая удаление пропущенных значений, преобразование категориальных признаков и нормализацию числовых данных.

Основная цель исследования заключается в разработке и сравнении различных моделей, таких как K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest и Gradient Boosting, с целью определения наиболее эффективного алгоритма для данной задачи. Кроме того, будет проведен подбор оптимальных гиперпараметров для каждой модели с использованием методов кросс-валидации.

Оценка качества моделей будет проводиться с использованием стандартных метрик классификации, таких как точность (accuracy), полнота (recall), F1-скор и ROC AUC. Важным аспектом работы является не только достижение высокой точности предсказаний, но и понимание влияния различных параметров на производительность моделей.

Исследование планируется завершить с анализом результатов и выработкой рекомендаций по использованию наиболее подходящей модели для практических задач, связанных с классификацией грибов на основе их морфологических признаков.

Выполнение работы

Для решения задачи классификации был выбран набор данных содержащий информацию о грибах.

В наборе данных присутствуют следующие столбцы:

- Cap Diameter: Диаметр шляпки
- Cap Shape: Форма шляпки
- Gill Attachment: Крепление жабр
- Gill Color: Цвет жабр
- Stem Height: Высота стебля
- Stem Width: Ширина стебля
- Stem Color: Цвет стебля
- Season: Сезон
- Target Class: целевой класс, является ли гриб ядовитым или нет (1 – гриб ядовит, 0 – гриб не ядовит и съедобен)

Загружаем данные, получаем общую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 54035 строк и 9 столбцов, из которых 7 типа int64 и 2 типа float.

Пропусков не было обнаружено.

Строим график pairplot для визуализации распределения данных попарно для множества колонок.

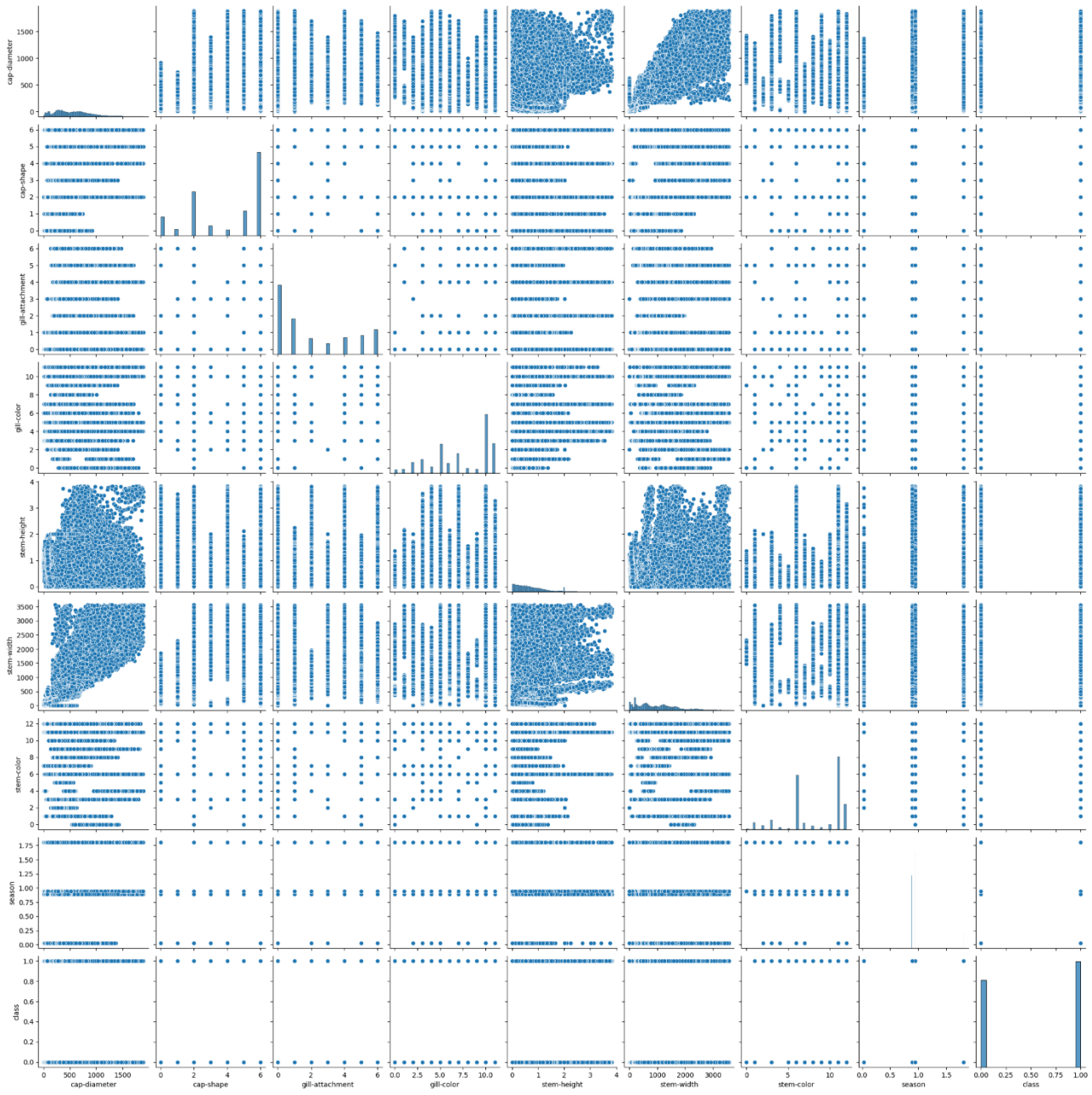

Рисунок 1 - Визуализация распределения данных попарно для множества колонок

Проверяем сбалансированы ли классы в нашем наборе данных. Получаем следующую гистограмму:

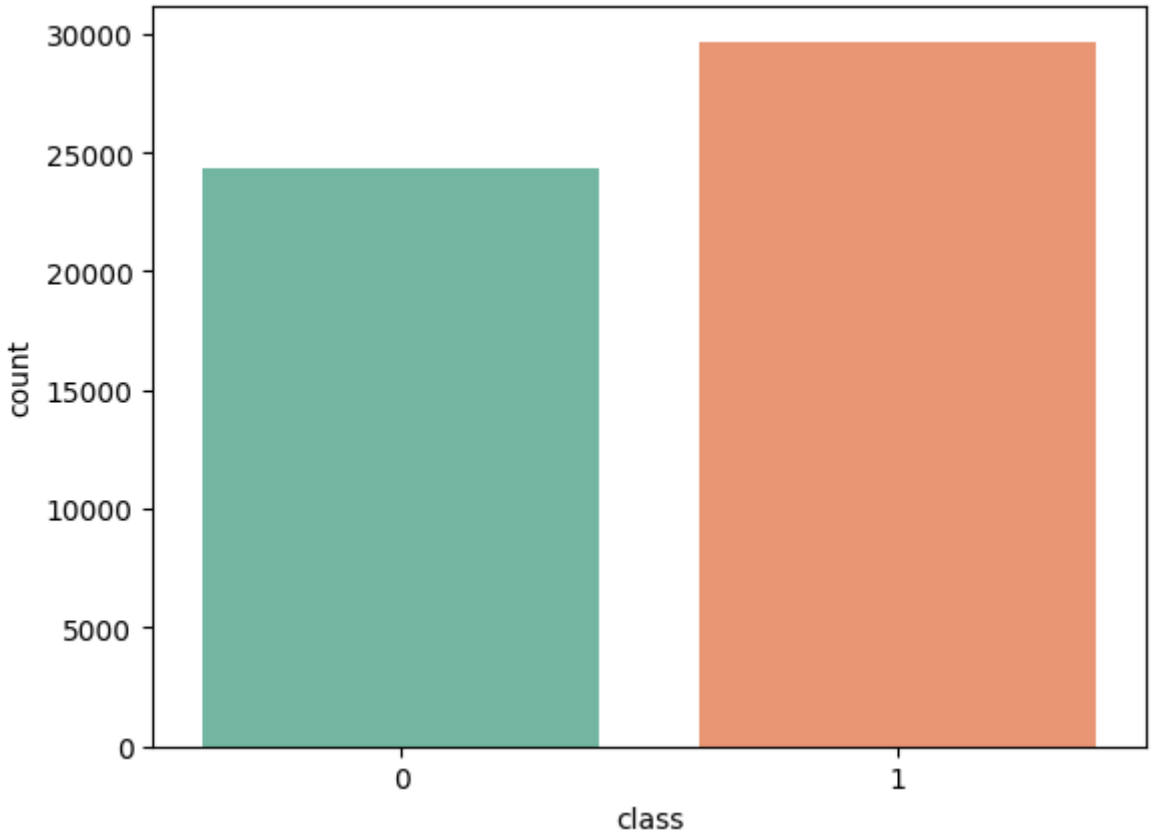

Рисунок 2 - Гистограмма классов

Видим, что классы немножко не сбалансированы.

Строим таблицу средних значений с группировкой по целевому признаку и делаем следующие предположения:

- У ядовитых грибов ширина стебля меньше
- У ядовитых грибов есть жабры типа 0
- У ядовитых грибов больше высота стебля
- У ядовитых грибов диаметр шляпки меньше

Подтвердим наши предположения графиками.

Строим гистограмму с важностью признаков для целевого признака.

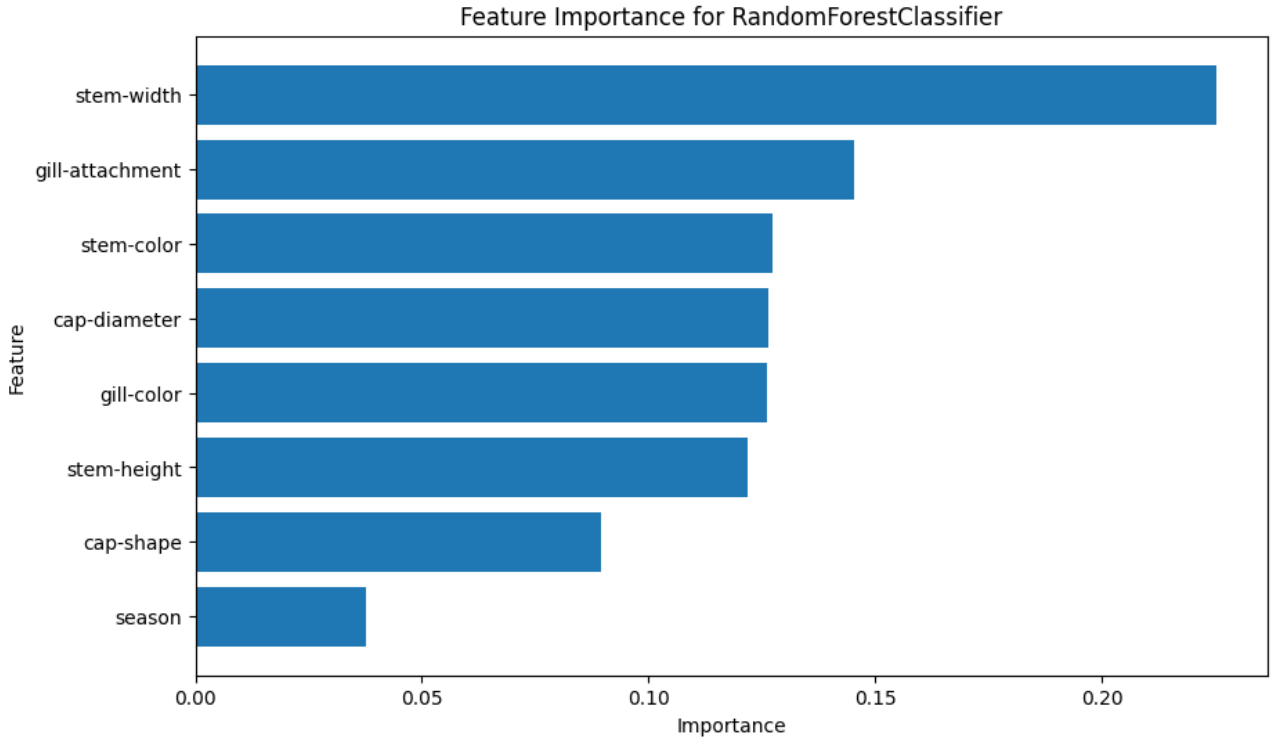

Рисунок 3 - Гистограмма важности признаков для целевого признака

Можно заметить, что ширина стебля и крепление жабр наиболее важны для целевого признака.

Далее приведем данные к нужному формату. Сначала масштабируем численные признаки методом MinMaxScaler, который преобразует каждый признак таким образом, чтобы он имел среднее значение равное 0 и стандартное отклонение равное 1. Посмотрим на распределения колонок до и после масштабирования.

Распределение не изменилось.

Проводим корреляционный анализ данных. Строим тепловую карту корреляций.

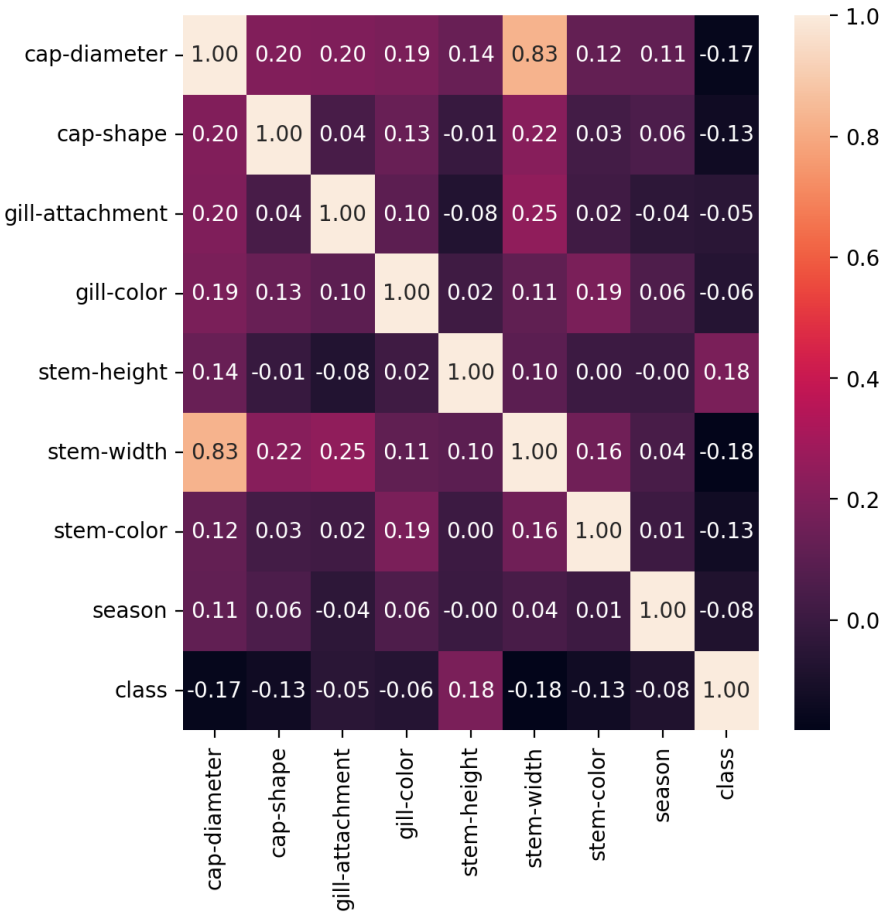

Рисунок 44 - Тепловая карта корреляций

Выберем метрики для оценки качества модели:

- $Precision = \frac{TP}{TP+FP}$ - показывает, какую долю объектов, которые модель предсказала как положительные, действительно являются положительными.
- $F_1 = \frac{TP}{TP+FN}$ - показывает, какую долю положительных объектов модель способна обнаружить.
- $F_1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$ - среднее гармоническое precision и recall. Другими словами, это средневзвешенное значение точности и отзыва. [2]
- $ROC\ AUC$ - основана на вычислении следующих характеристик: $TPR = \frac{TP}{TP+FN}$ - True Positive Rate, откладывается по оси ординат. Совпадает с recall. $FPR = \frac{FP}{FP+TN}$ - False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно. 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. [3]

Выберем модели для решения задачи классификации:

- KNN;
- SVC;
- Дерево решений;
- Случайный лес;
- Градиентный бустинг.

Формируем обучающую и тестовую выборку в соотношении 8:2. Оставляем все колонки, так как они влияют на целевой признак.

Строим базовое решения, выводим значениями метрик и ROC-кривую.

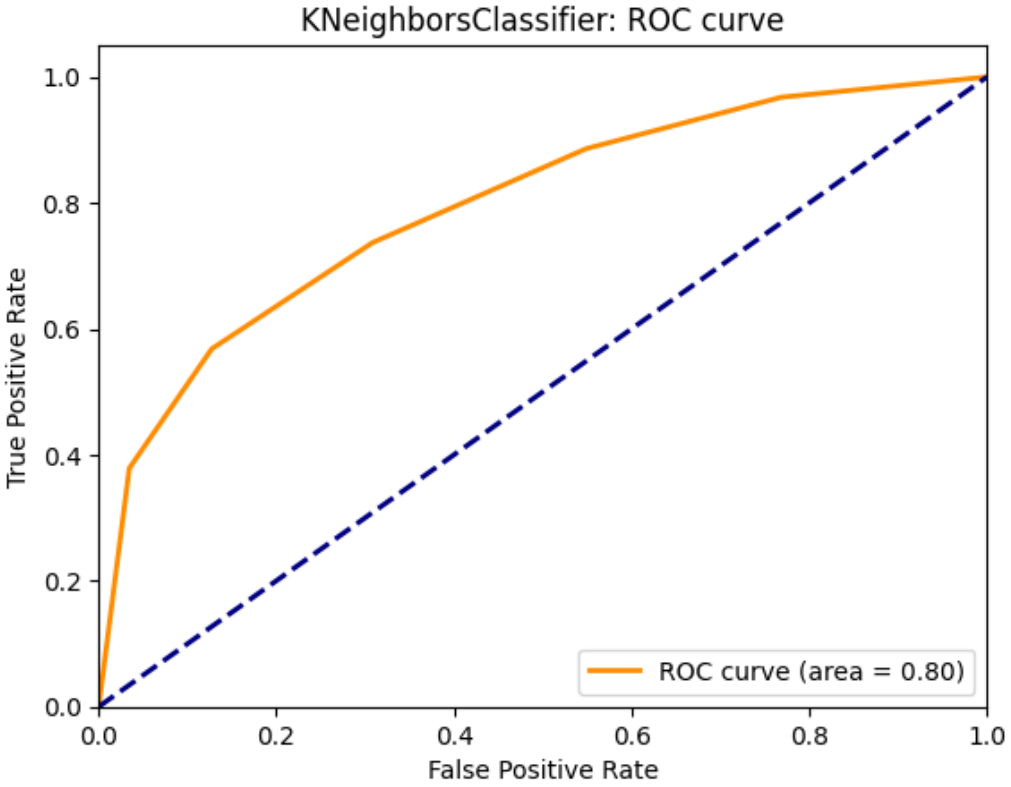

Рисунок 5 - ROC-кривая базовой модели KNN

KNeighborsClassifier:

Accuracy: 0.72

Precision: 0.74

Recall: 0.74

F1-score: 0.74

ROC AUC score: 0.7951912325518053

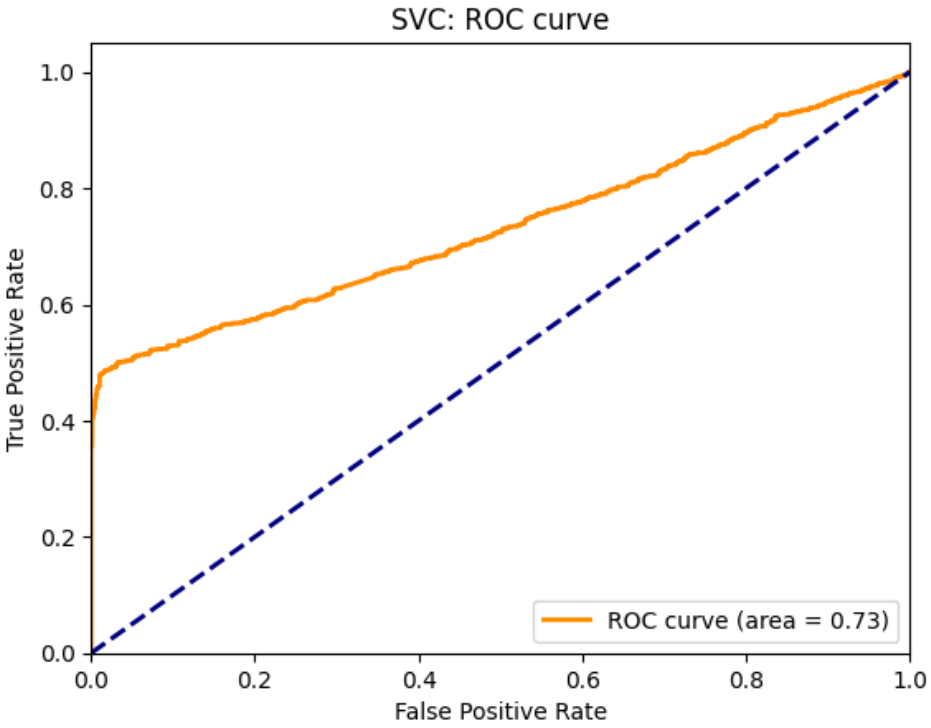

Рисунок 6 5- ROC-кривая базовой модели SVC

SVC:

Accuracy: 0.9

Precision: 0.91

Recall: 0.89

F1-score: 0.74

ROC AUC score: 0.7322453450732926

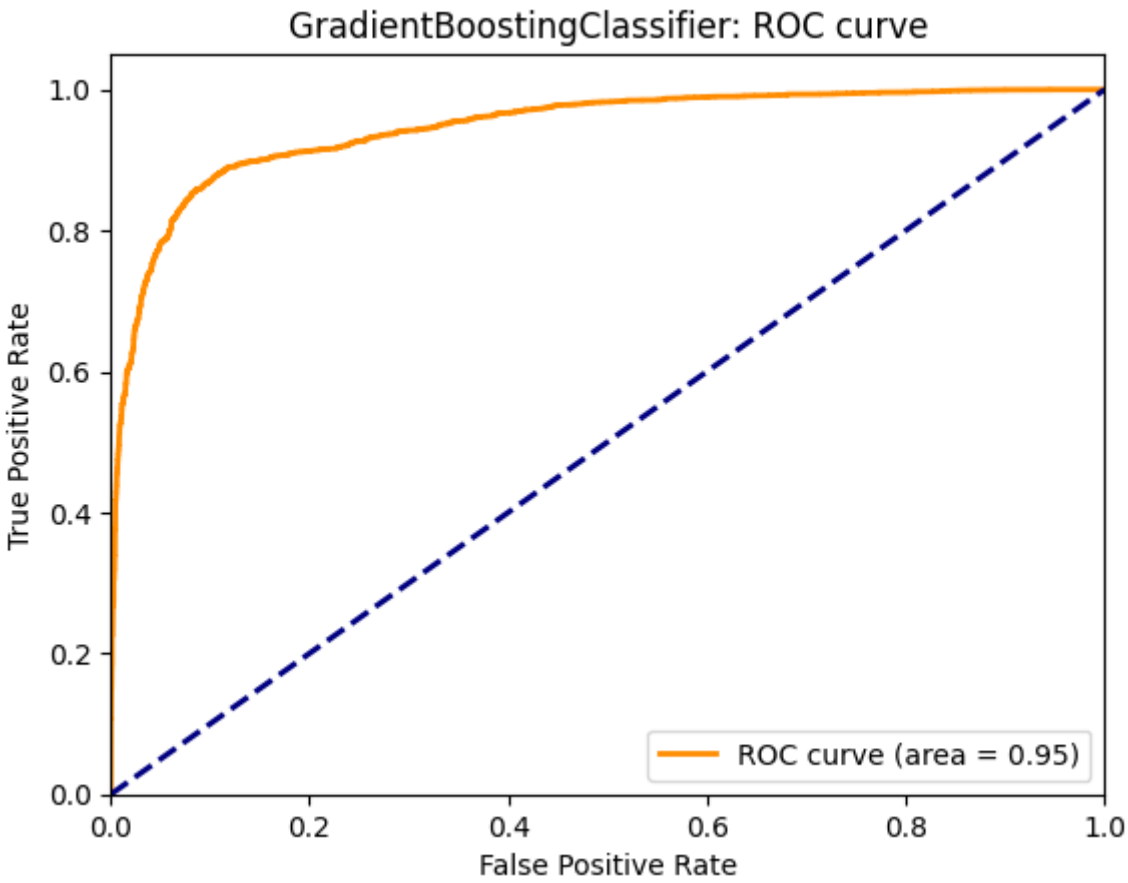

Рисунок 7 - ROC-кривая базовой модели Decision Tree

DecisionTreeClassifier:

Accuracy: 0.88

Precision: 0.9

Recall: 0.89

F1-score: 0.89

ROC AUC score: 0 .9452147255706624

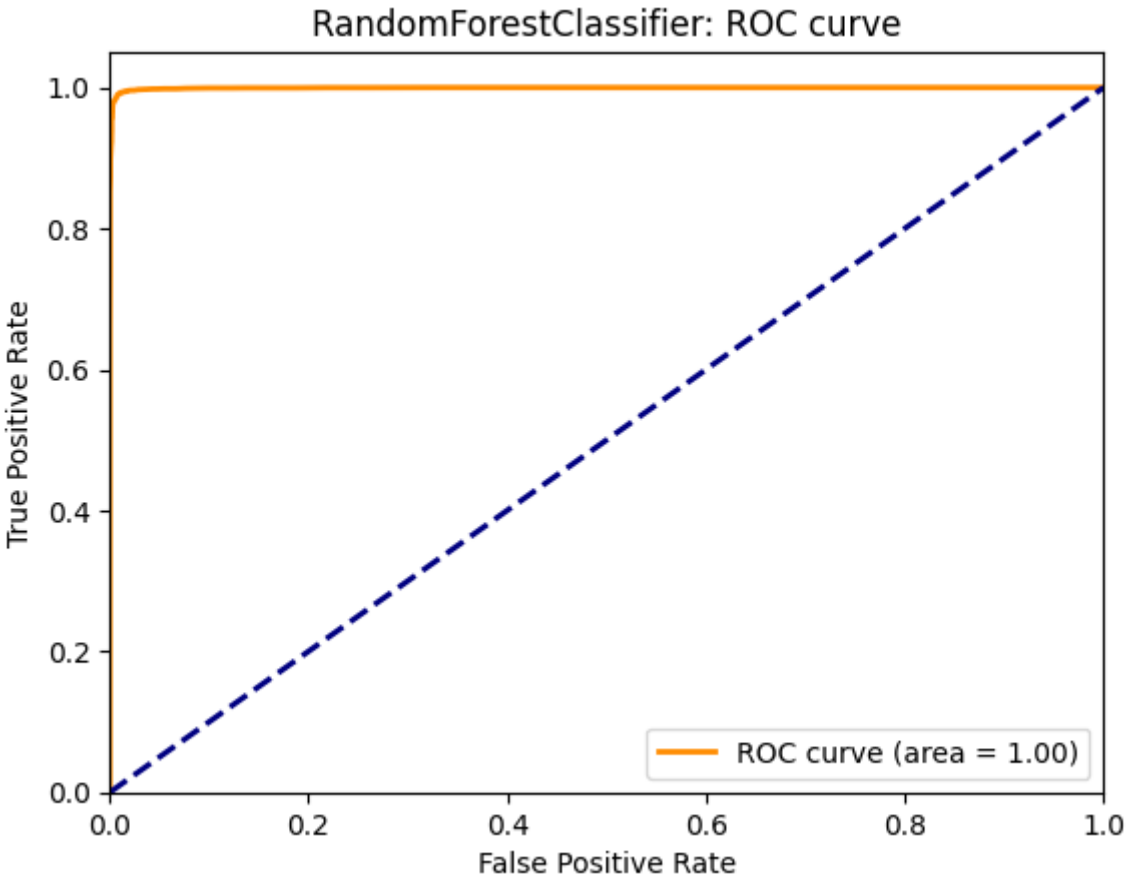

Рисунок 8 - ROC-кривая базовой модели Random Forest

RandomForestClassifier:

Accuracy: 0.99

Precision: 0.99

Recall: 0.99

F1-score: 0.99

ROC AUC score: 0.9993568777453278

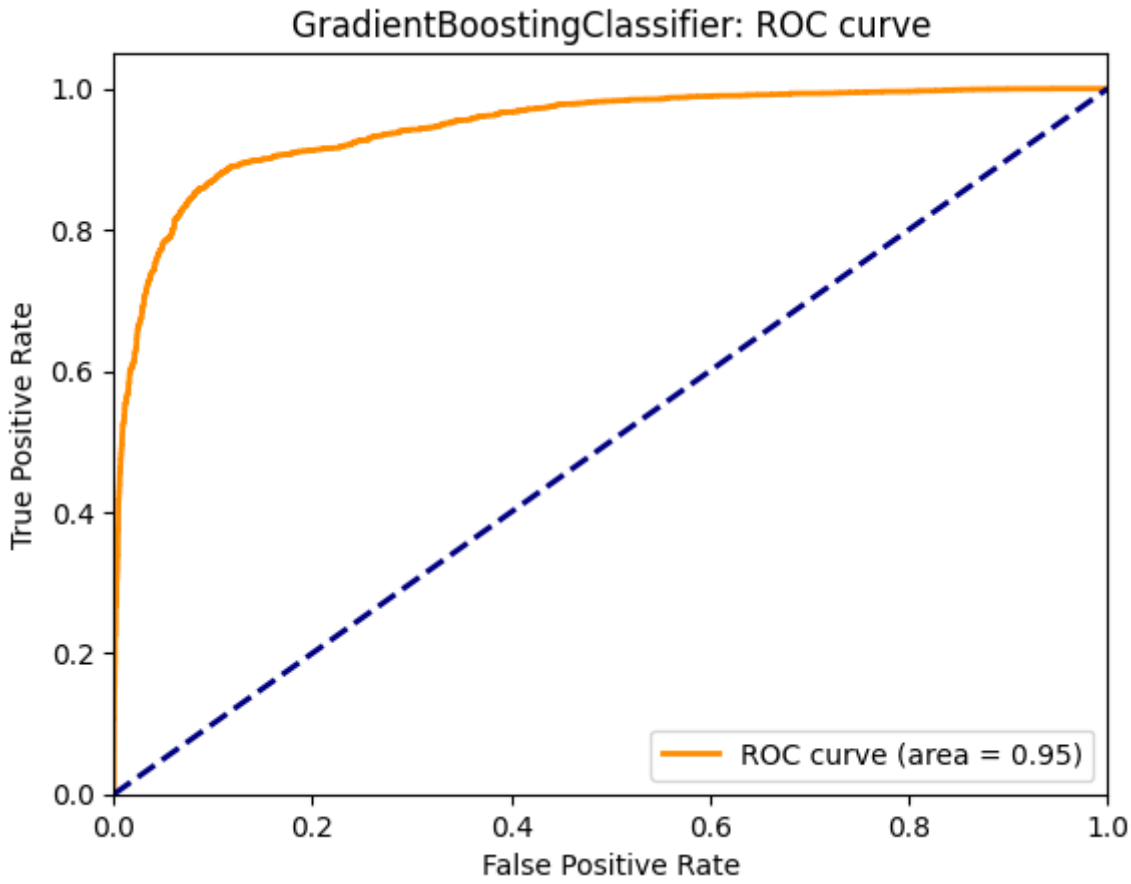

Рисунок 96 - ROC-кривая базовой модели Gradient Boosting

GradientBoostingClassifier:

Accuracy: 0.88

Precision: 0.9

Recall: 0.89

F1-score: 0.89

ROC AUC score: 0.9452147255706624

Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

KNeighboursClassifier:

Best hyperparameters: {'algorithm': 'auto', 'n_neighbors': 5, 'weights': 'distance'}

Best score: 0.7475562898671996

SVC:

Best hyperparameters: {'C': 1, 'degree': 4, 'gamma': 'scale', 'kernel': 'rbf'}

DecisionTreeClassifier:

Best hyperparameters: {'criterion': 'entropy', 'max_depth': None, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2}

Best score: 0.9770749620006349

RandomForestClassifier:

Best hyperparameters: {'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}

Best score: 0.9897056900512103

GradientBoostingClassifier:

Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 3, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 2}

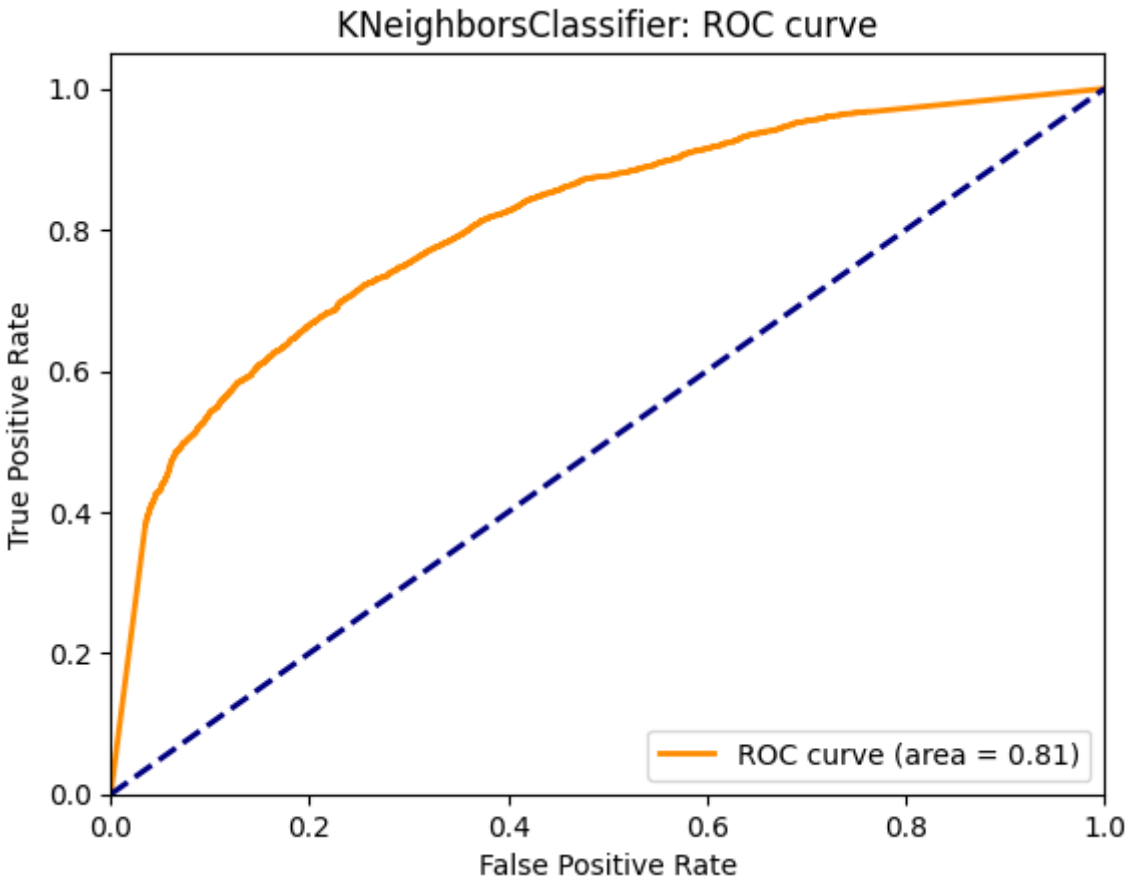

Рисунок 10 - ROC-кривая модели KNN после поиска гиперпараметров

KNeighborsClassifier:

Accuracy: 0.73

Precision: 0.76

Recall: 0.75

F1-score: 0.75

ROC AUC score: 0.8109146313706337

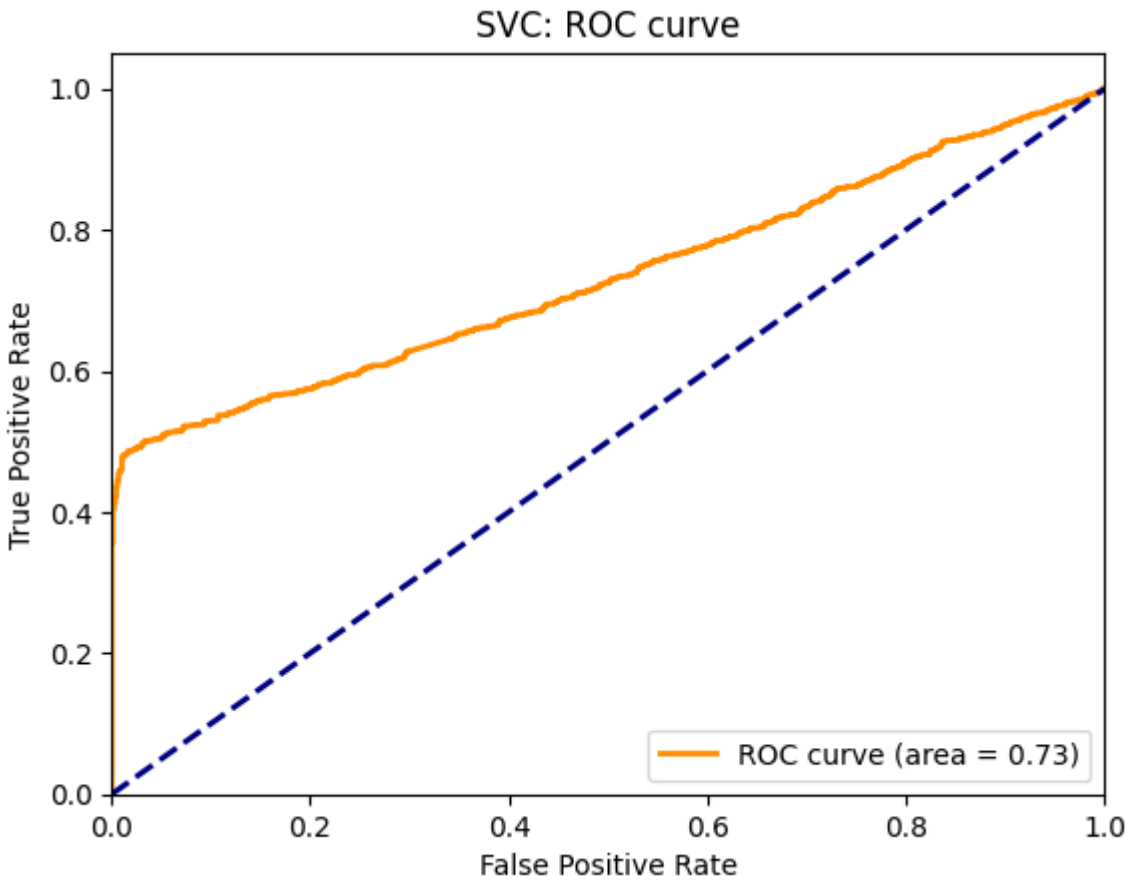

Рисунок 11 - ROC-кривая модели SVC после поиска гиперпараметров

SVC:

Precision: 0.95

Recall: 0.5

F1-score: 0.66

ROC AUC score: 0.7322713031198976

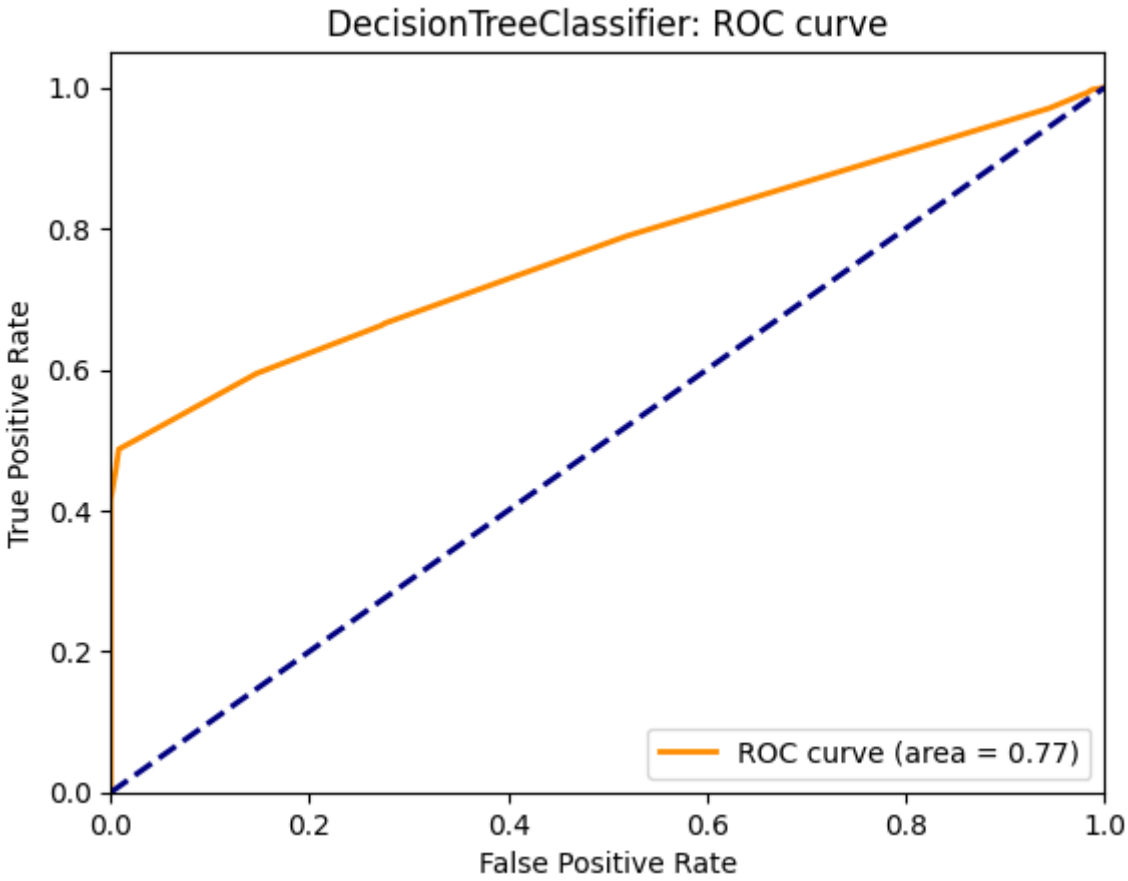

Рисунок 12 - ROC-кривая модели Decision Tree после поиска гиперпараметров

DecisionTreeClassifier:

Precision: 0.97

Recall: 0.5

F1-score: 0.66

ROC AUC score: 0.7658675049385183

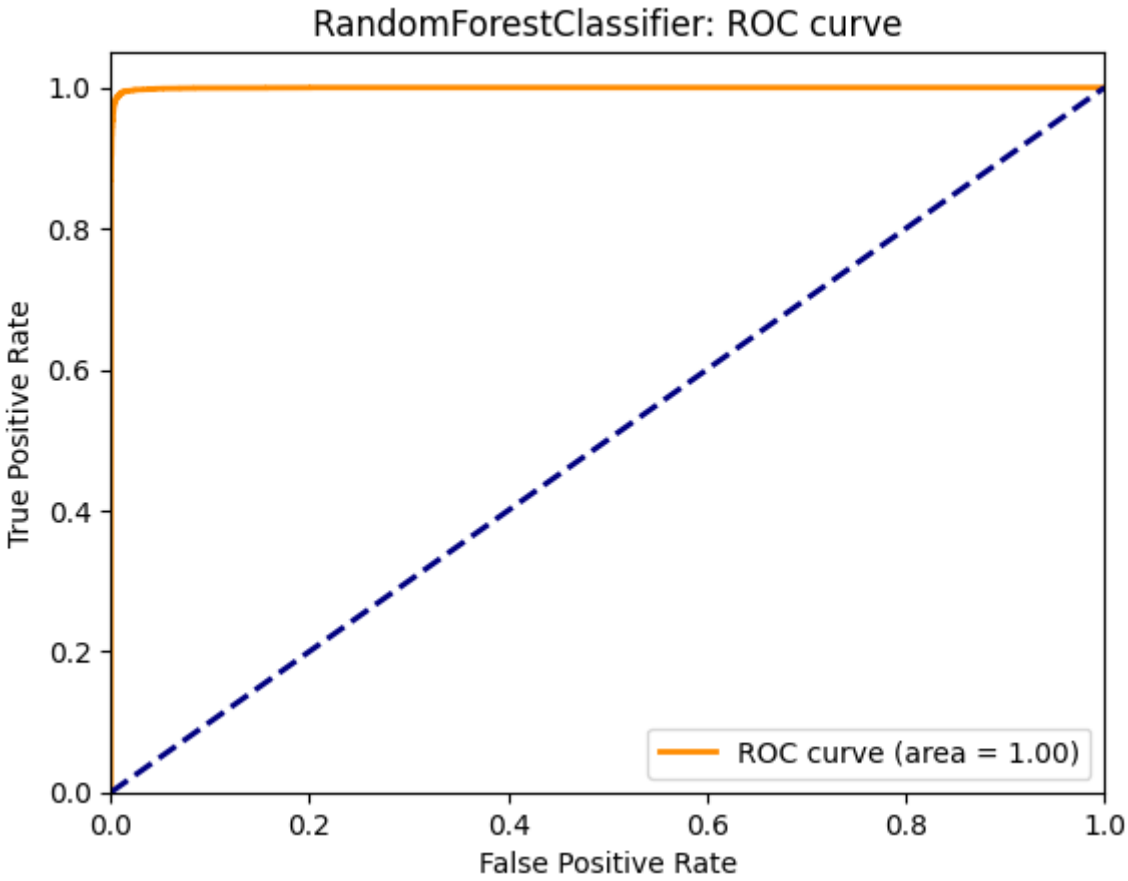

Рисунок 13 - ROC-кривая модели Random Forest после поиска гиперпараметров

RandomForestClassifier:

Precision: 0.95

Recall: 0.51

F1-score: 0.67

ROC AUC score: 0.7622333784138077

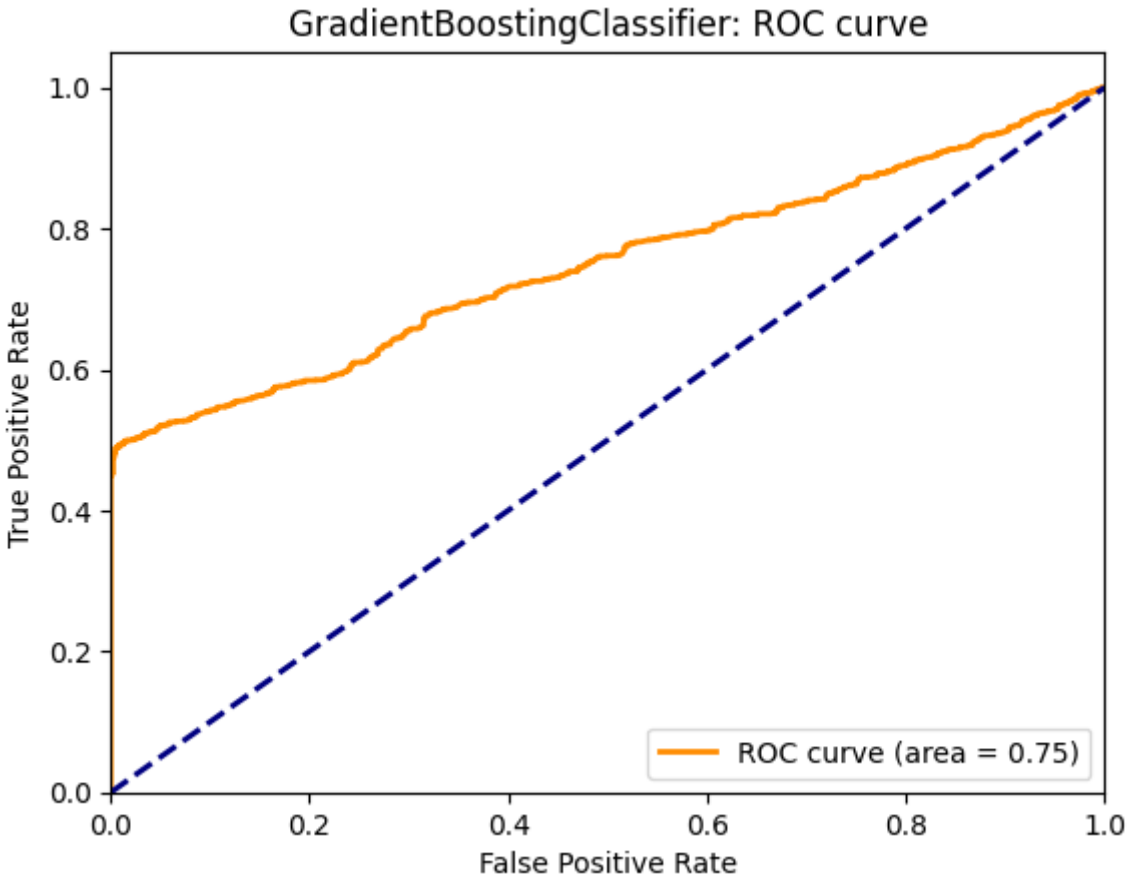

Рисунок 14 - ROC-кривая модели Gradient Boosting после поиска гиперпараметров

GradientBoostingClassifier:

Precision: 0.91

Recall: 0.53

F1-score: 0.67

ROC AUC score: 0.7467165234215128

Таблица 1 - Сравнение базовых моделей с моделями после подбора гиперпараметров по 4 метрикам

Модель	Baseline	GridSearch()
KNN	Precision: 0.74 Recall: 0.68 F1-score: 0.71 ROC AUC score: 0.7403174322992512	Precision: 0.8 Recall: 0.61 F1-score: 0.69 ROC AUC score: 0.7447099664189403
SVC	Precision: 0.95 Recall: 0.5 F1-score: 0.66 ROC AUC score: 0.7322453450732926	Precision: 0.95 Recall: 0.5 F1-score: 0.66 ROC AUC score: 0.7322713031198976

Decision Tree	Precision: 0.72 Recall: 0.71 F1-score: 0.71 ROC AUC score: 0.6374504525785426	Precision: 0.97 Recall: 0.5 F1-score: 0.66 ROC AUC score: 0.7658675049385183
Random forest	Precision: 0.75 Recall: 0.66 F1-score: 0.7 ROC AUC score: 0.743556996515565	Precision: 0.95 Recall: 0.51 F1-score: 0.67 ROC AUC score: 0.7622333784138077
Gradient Boosting	Precision: 0.91 Recall: 0.54 F1-score: 0.67 ROC AUC score: 0.7442457500188195	Precision: 0.91 Recall: 0.53 F1-score: 0.67 ROC AUC score: 0.7467165234215128

На основании трех метрик из четырех лучшими для решения данной задачи классификации оказались модели градиентного бустинга и метод случайного леса.

Заключение

Классификация грибов на съедобные и ядовитые с использованием методов машинного обучения является актуальной и важной задачей в области безопасности продуктов питания. Анализ и обработка данных с помощью алгоритмов машинного обучения могут помочь точно и быстро определить, какие грибы являются ядовитыми, что позволяет предотвращать случаи отравления и повышать безопасность потребления грибов.

В рамках данного исследования была разработана эффективная модель, которая может помочь быстро и точно определить съедобность грибов на основе их морфологических признаков. Исходные данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы машинного обучения, такие как метод ближайших соседей (KNN), метод опорных векторов (SVC), дерево решений, случайный лес и градиентный бустинг.

Результаты исследования показали, что большинство использованных методов достигли хороших результатов в классификации грибов. Однако самыми точными, на основании всех метрик (точность, полнота, F1-скор и ROC AUC), оказались модели градиентного бустинга и случайного леса. Эти модели продемонстрировали наилучшие показатели и могут быть рекомендованы для практического применения в системах автоматической классификации грибов.

В ходе работы также было показано, что оптимизация гиперпараметров с использованием методов кросс-валидации значительно улучшает производительность моделей. Визуализация результатов и анализ влияния различных гиперпараметров на качество моделей помогли глубже понять их поведение и выбрать наилучшие настройки для каждой модели.

Данное исследование вносит значимый вклад в область применения машинного обучения для классификации грибов и может быть использовано для создания надежных систем, повышающих безопасность потребления грибов и предотвращающих случаи отравления. В дальнейшем возможно углубленное изучение дополнительных методов обработки данных и использование более

сложных моделей для достижения еще более высоких показателей точности и надежности предсказаний.

Список использованной литературы

1. T-test на Python для проверки и получения t-статистики // Помощник Python URL: https://pythonpip.ru/osnovy/t-test-na-python

2. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6

3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://nbviewer.org/github/ugapanyuk/courses_current/blob/main/notebooks/ml_project_example/project_classification_regression.ipynb

4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses_current/wiki/COURSE_TMO_SPRING_2024/