|  |  |
| --- | --- |
|  | **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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

*\_\_\_\_\_\_\_\_\_\_\_Предсказание выживаемости\_\_\_\_\_\_\_\_\_\_*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_на корабле “Титаник”\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

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

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

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

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

Студент \_\_\_ИУ5-64Б\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Харитонов А.А.\_\_\_**

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

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_ Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 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 \_\_\_\_Предсказание выживаемости на корабле титаник\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Студент группы \_\_ИУ5-64Б\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Стельмах Яна Сергеевна\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследоваткльская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 \_\_32\_\_ листах формата А4.

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

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

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

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

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

Дата выдачи задания « 13 » февраля 2023 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Харитонов А.А.\_\_\_**

(Подпись, дата) (И.О.Фамилия)

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

**СОДЕРЖАНИЕ**

[**Введение** 4](#_Toc133806658)

[**Постановка задачи** 6](#_Toc133806659)

[**Выполнение работы** 7](#_Toc133806660)

[**Создание веб-приложения** 29](#_Toc133806661)

[**Заключение** 31](#_Toc133806662)

[**Список использованной литературы** 32](#_Toc133806663)

# **Введение**

**Титаник**— британский трансатлантический пассажирский пароход, второй лайнер класса Олимпик. Крупнейшее судно в мировой истории начала XX века. Во время первого рейса затонул в Северной Атлантике, столкнувшись с айсбергом в ночь с 14 на 15 апреля 1912 года.

Каюты и общественные помещения «Титаника» разделялись на три класса. К услугам пассажиров первого класса были представлены плавательный бассейн, корт для игры в сквош, ресторан «À La Carte», два кафе, гимнастический зал. Во всех классах имелись обеденные и курительные салоны, открытые и закрытые променады. Наиболее роскошными и изысканными были интерьеры первого класса, выполненные в различных художественных стилях с использованием дорогих материалов, таких как красное дерево, позолота, витражное стекло, шёлк и прочие. Каюты и салоны третьего класса оформлялись максимально просто: стальные стены окрашивались в белый цвет либо обшивались деревянными панелями.

10 апреля 1912 года «Титаник» отправился из Саутгемптона в Нью-Йорк, в свой первый и единственный рейс. Совершив остановки во французском Шербуре и ирландском Квинстауне, лайнер вышел в Атлантический с 1317 пассажирами и 908 членами экипажа на борту. Планировалось прибыть в Нью-Йорк 17 апреля. 14 апреля радиостанция «Титаника» приняла семь ледовых предупреждений, однако лайнер продолжал двигаться почти на предельной скорости. Чтобы избежать встречи с плавучими льдами, капитан приказал идти чуть южнее привычного маршрута.

14 апреля в 23:39 на капитанский мостики было доложено об айсберге прямо по курсу. Меньше чем через минуту произошло столкновение. Получив несколько пробоин, пароход начал тонуть. В шлюпки сажали в первую очередь женщин и детей. В 2:20 15 апреля, разломившись на две части, «Титаник» затонул, унеся жизни 1496 человек. 712 спасшихся человек подобрал пароход «Карпатия».

В ходе данной исследовательской работы был произведён анализ, какие критерии благоприятствовали выживаемости пассажиров судна. Для этого использовался датасет, в котором обозначены все члены судна и информация о них. Такое исследование может быть полезным для выявления закономерностей при кораблекрушениях и будущему их будущему учёту.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии или классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.
7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

# **Постановка задачи**

Данная работа по машинному обучению направлена на решение задачи классификации, а именно, предсказание риска возникновения инсульта у пациентов.

Имеются данные о пассажирах парохода. Каждый пассажир может быть классифицирован как человек, который имеет большие или меньшие шансы на выживание.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания выживаемости человека на корабле. Для этого мы будем использовать различные алгоритмы классификации, такие как K ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

# **Выполнение работы**

Для решения задачи классификации был выбран набор данных содержащий информацию о пассажирах корабля Титаник.

В наборе данных присутствуют следующие столбцы:

* Id пассажира
* Выжил пассажир или нет
* Социальный статус пассажира
* Имя пассажира
* Пол пассажира
* Возраст пассажира
* Количество родственников/супругов на борту
* Количество родителей/детей на борту
* Номер билета
* Стоимость проезда
* Кабина пассажира
* Порт посадки

Загружаем данные, получаем обую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 418 строк и 8 столбцов.

Заполняем пропуски данных усреднёнными значениями страт, к которым принадлежит пассажир. Стоимость билета коррелирует с категориальным признаком социального статуса пассажира, а возраст – с социальным положением, закодированным в имени пассажира. Категории кабины, номера пассажира, его имени и номера билета можно отбросить за ненадобностью.

Проведём кодирование категориальных признаков, очистку от выбросов данных и нормализацию.

После проведения очистки данных произведём визуализацию данных:

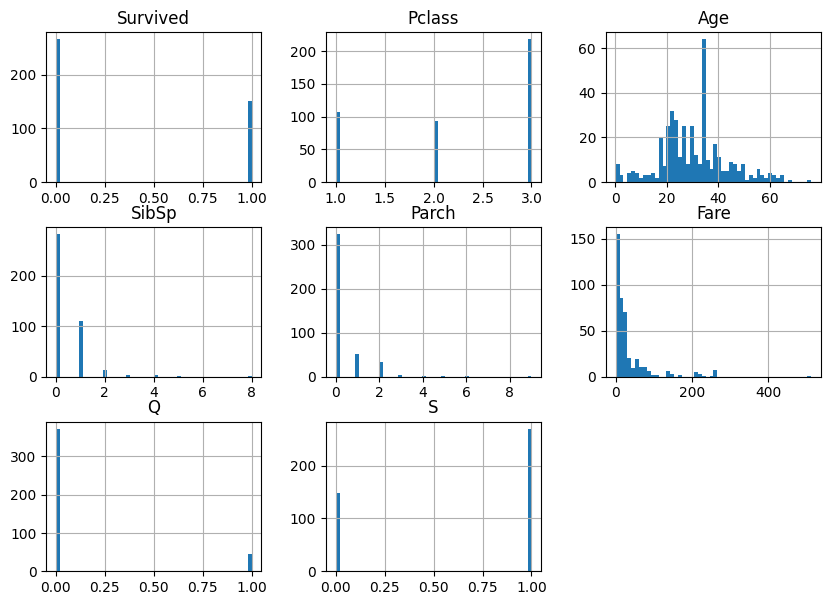


Рисунок 1 - Визуализация распределения данных попарно для множества колонок

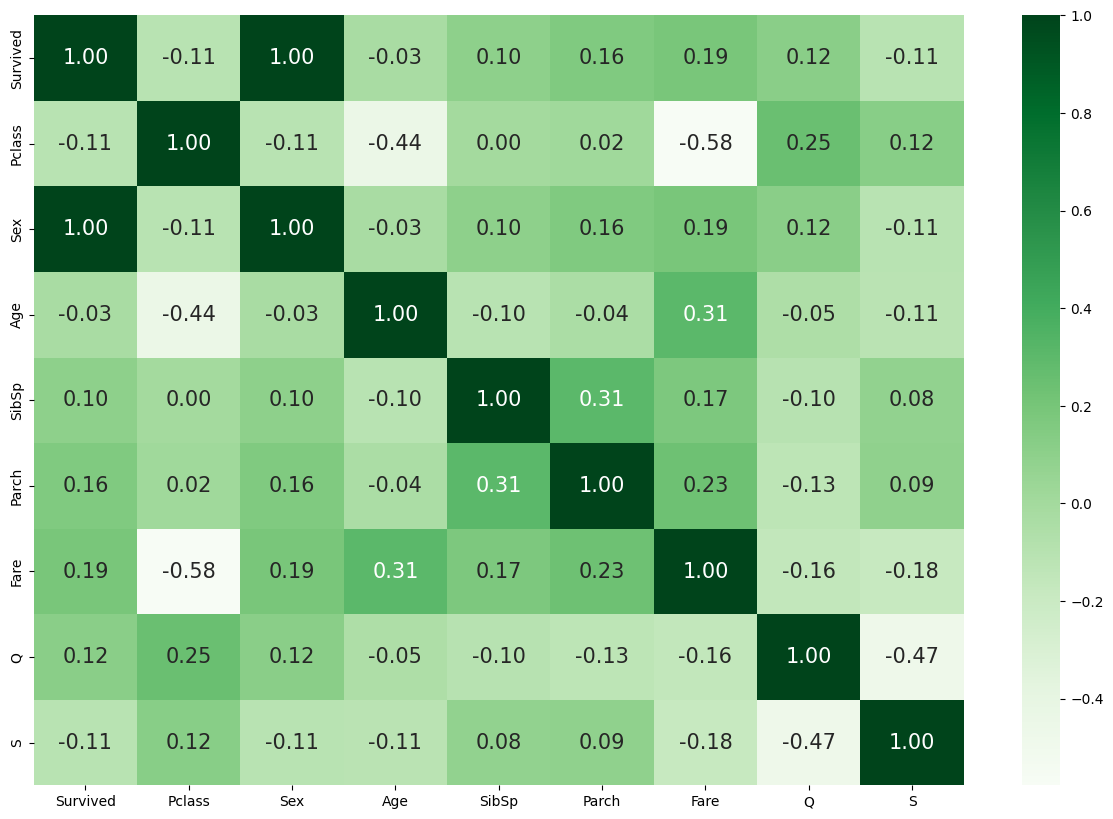


Рисунок 2 – Корреляция данных

Признак половой принадлежности имеет 100% корреляцию с целевым признаком выживаемости. Удалим это поле, чтобы не привести модель к перетренированности

Выберем метрики для оценки качества модели:

* - показывает, какую долю объектов, которые модель предсказала как положительные, действительно являются положительными.
* - показывает, какую долю положительных объектов модель способна обнаружить.
* - среднее гармоническое precision и recall. Другими словами, это средневзвешенное значение точности и отзыва. [2]
* - oснована 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. 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.

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

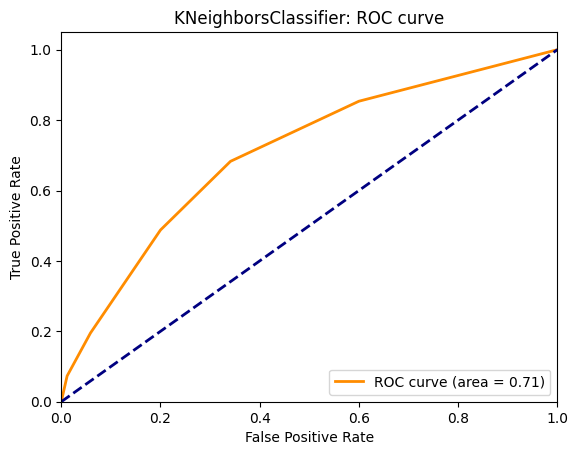
KNeighborsClassifier:

Precision: 0.54

Recall: 0.49

F1-score: 0.51

ROC AUC score: 0.7071736011477762



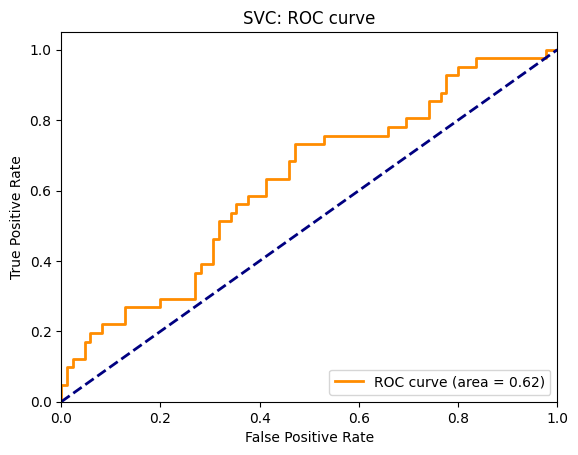
SVC:

Precision: 0.44

Recall: 0.27

F1-score: 0.33

ROC AUC score: 0.6235294117647059



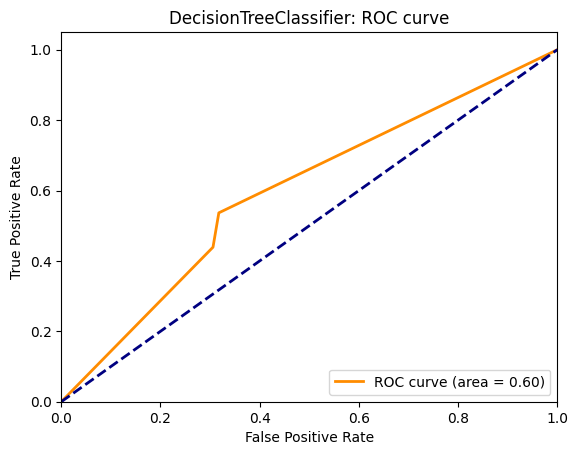
DecisionTreeClassifier:

Precision: 0.41

Recall: 0.44

F1-score: 0.42

ROC AUC score: 0.5971305595408896



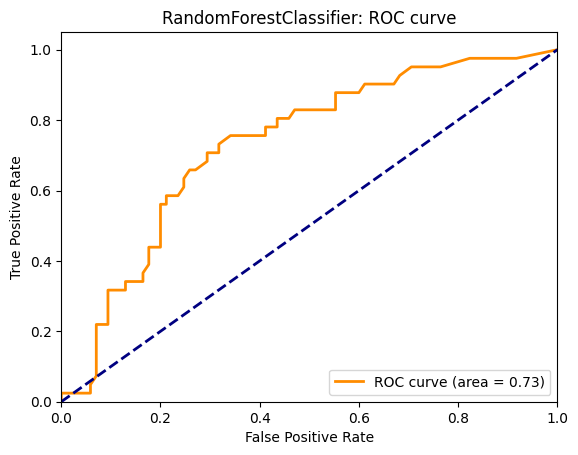
RandomForestClassifier:

Precision: 0.55

Recall: 0.51

F1-score: 0.53

ROC AUC score: 0.73012912482066



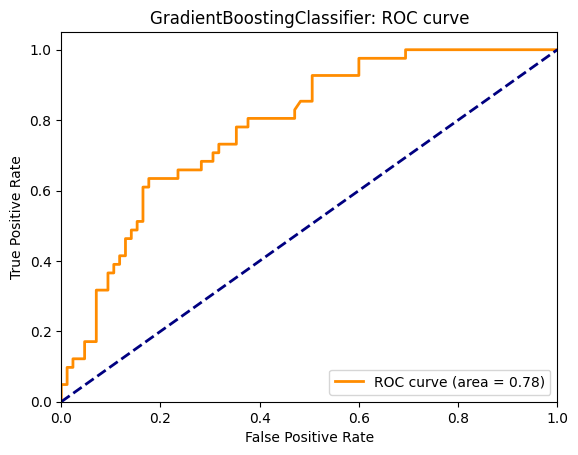
GradientBoostingClassifier:

Precision: 0.61

Recall: 0.49

F1-score: 0.54

ROC AUC score: 0.7817790530846485



Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

KNeighboursClassifier:

Best hyperparameters: {'algorithm': 'auto', 'n\_neighbors': 10, 'weights': 'uniform'}

SVC:

Best hyperparameters: {'C': 10, 'degree': 5, 'gamma': 'scale', 'kernel': 'poly'}

DecisionTreeClassifier:

Best hyperparameters: {'criterion': 'gini', 'max\_depth': 7, 'max\_features': 'log2', 'min\_samples\_leaf': 4, 'min\_samples\_split': 2}

RandomForestClassifier:

Best hyperparameters: {'max\_depth': None, 'max\_features': 'log2', 'min\_samples\_leaf': 1, 'min\_samples\_split': 5, 'n\_estimators': 100}

GradientBoostingClassifier:

Best hyperparameters: {'learning\_rate': 0.2, 'max\_depth': 7, 'max\_features': None, 'min\_samples\_leaf': 2, 'min\_samples\_split': 5}

Используя найденные гиперпараметры, произведём переобучение моделей и сравним их результаты с базовым решением.

KNeighborsClassifier:

Precision: 0.59

Recall: 0.39

F1-score: 0.47

ROC AUC score: 0.7014347202295552

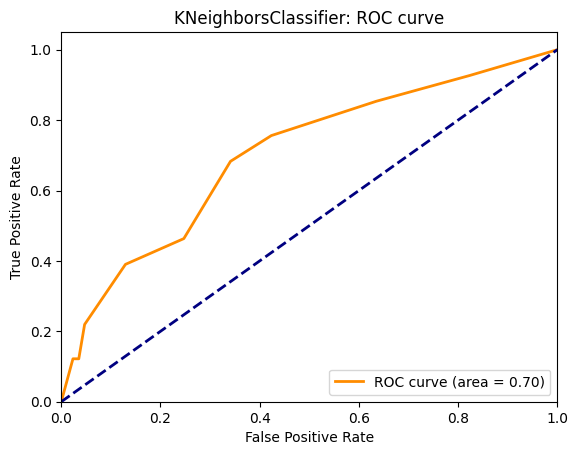


Рисунок 22 - ROC-кривая модели KNN после поиска гиперпараметров

SVC:

Precision: 0.62

Recall: 0.39

F1-score: 0.48

ROC AUC score: 0.6559540889526543

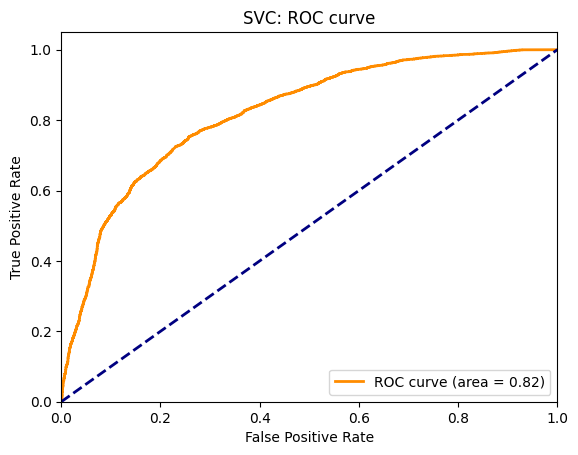


Рисунок 23 - ROC-кривая модели SVC после поиска гиперпараметров

DecisionTreeClassifier:

Precision: 0.52

Recall: 0.41

F1-score: 0.46

ROC AUC score: 0.6370157819225251

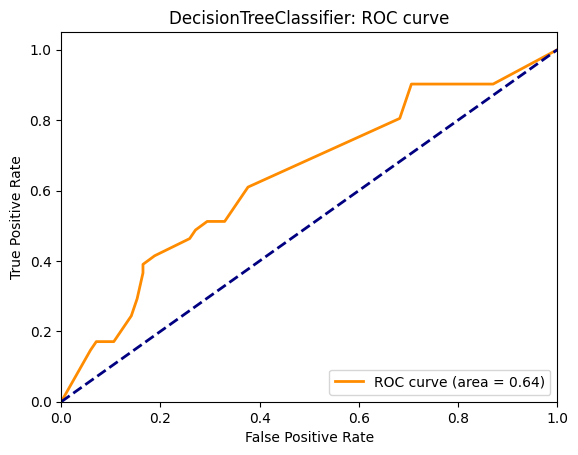


Рисунок 24 - ROC-кривая модели Decision Tree после поиска гиперпараметров

RandomForestClassifier:

Precision: 0.53

Recall: 0.46

F1-score: 0.49

ROC AUC score: 0.7294117647058823

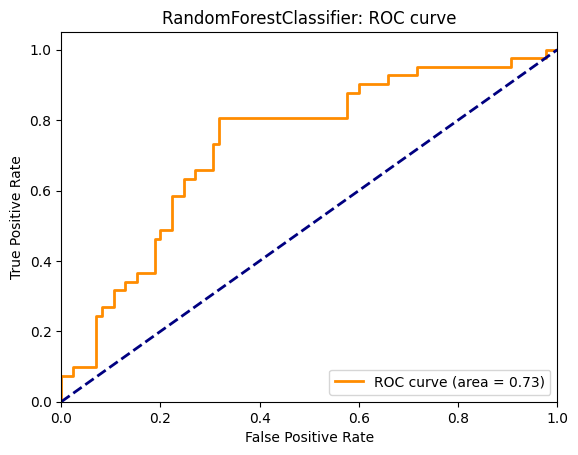


Рисунок 25 - ROC-кривая модели Random Forest после поиска гиперпараметров

GradientBoostingClassifier:

Precision: 0.56

Recall: 0.61

F1-score: 0.58

ROC AUC score: 0.7374461979913917

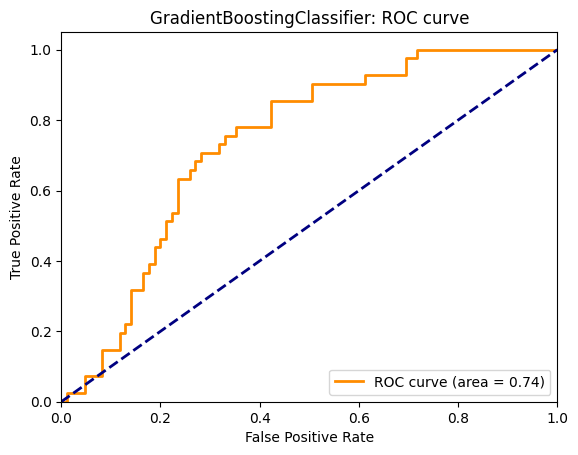


Рисунок 26 - ROC-кривая модели Gradient Boosting после поиска гиперпараметров

Таблица 2 - Сравнение базовых моделей с моделями после подбора гиперпараметров по 4 метрикам

|  |  |  |
| --- | --- | --- |
| Модель | Baseline | GridSearch |
| KNN | Precision: 0.54  Recall: 0.49  F1-score: 0.51  ROC AUC score: 0.7071736011477762 | Precision: 0.59  Recall: 0.39  F1-score: 0.47  ROC AUC score: 0.7014347202295552 |
| SVC | Precision: 0.44  Recall: 0.27  F1-score: 0.33  ROC AUC score: 0.6235294117647059 | Precision: 0.62  Recall: 0.39  F1-score: 0.48  ROC AUC score: 0.6559540889526543 |
| Decision Tree | Precision: 0.41  Recall: 0.44  F1-score: 0.42  ROC AUC score: 0.5971305595408896 | Precision: 0.52  Recall: 0.41  F1-score: 0.46  ROC AUC score: 0.6370157819225251 |
| Random forest | Precision: 0.55  Recall: 0.51  F1-score: 0.53  ROC AUC score: 0.73012912482066 | Precision: 0.53  Recall: 0.46  F1-score: 0.49  ROC AUC score: 0.7294117647058823 |
| Gradient Boosting | Precision: 0.61  Recall: 0.49  F1-score: 0.54  ROC AUC score: 0.7817790530846485 | Precision: 0.56  Recall: 0.61  F1-score: 0.58  ROC AUC score: 0.7374461979913917 |

На основании трех метрик из четырех лучшими для решения данной задачи классификации оказались модели градиентного бустинга и дерева решений.

# **Заключение**

В рамках НИР была рассмотрена задача классификации выживаемости с помощью методов машинного обучения. Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самыми точными на основании трех метрик из четырех оказались модели градиентного бустинга и дерева решений.

# **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 (дата обращения: 30.04.2023).
2. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6 (дата обращения: 01.05.2023).
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://scikit-learn.org/stable/tutorial/statistical\_inference/supervised\_learning.html (дата обращения: 25.04.2023).
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2023/ (дата обращения: 25.04.2023).