|  |  |
| --- | --- |
|  | **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 |  |  |  | А.А. Горенков |
|  | (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 | | | |  | | | | | | | |
|  | | | | Горенков Александр Александрович | | | | | | | | | | | |
| 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** |  |  |  | А.А. Горенков |
|  |  | (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 20](#_Toc197287984)

[4 ФОРМИРОВАНИЕ ОБУЧАЮЩЕЙ И ТЕСТОВОЙ ВЫБОРОК 22](#_Toc197287985)

[5 ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ 23](#_Toc197287986)

[6 ПОДБОР ГИПЕРПАРАМЕТРОВ 25](#_Toc197287987)

[7 СРАВНЕНИЕ РЕШЕНИЙ 28](#_Toc197287988)

[ЗАКЛЮЧЕНИЕ 34](#_Toc197287990)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 35](#_Toc197287991)

# ВВЕДЕНИЕ

В современном мире проблема психического здоровья студентов приобретает все большую актуальность, поскольку эмоциональное благополучие напрямую влияет на их академическую успеваемость, социальную адаптацию и общее качество жизни. В связи с этим, выявление факторов, способствующих возникновению симптомов депрессии у студентов, является критически важным для своевременной диагностики, поддержки и разработки эффективных профилактических мер.

Целью данного исследования является построение и оценка моделей машинного обучения для классификации студентов на тех, кто проявляет симптомы депрессии, и тех, у кого они отсутствуют. Исследование основано на анализе датасета, содержащего информацию о различных аспектах жизни студентов, включая академические, социальные, экономические и личностные характеристики. Важным аспектом работы является применение методов предварительной обработки данных, таких как разведочный анализ, обработка (при необходимости) пропущенных значений, кодирование категориальных признаков (например, с использованием Weight of Evidence) и масштабирование числовых данных, что позволяет повысить точность и эффективность последующей классификации.

В рамках данного исследования мы проводим разведочный анализ данных (EDA) для идентификации основных статистических характеристик выборки, выявления взаимосвязей и аномалий. Дальнейший этап включает в себя тщательный отбор наиболее информативных признаков (например, с использованием Somers' D) и обучение различных моделей машинного обучения с последующим подбором их гиперпараметров для достижения наилучшей предсказательной способности.

# 1 ПОСТАНОВКА ЗАДАЧИ

Целью исследования является построение и оценка моделей машинного обучения для прогнозирования зарплаты человека(больше или меньше 50 000 долларов в год) на основании признаков, представленных в Таблице 1. Целевой переменной в решаемой задаче бинарной классификации является больше ли зарплата, чем 50 000 долларов в год.

Таблица 1 – Описание полей датасета

| Поле | Описание | Тип данных |
| --- | --- | --- |
| age | Возраст человека | object |
| workclass | Тип занятости (например, Private, Self-emp-not-inc, Local-gov, ?) | object |
| fnlwgt | "Final weight"; оценка того, сколько людей в популяции представляет данная запись (создается Бюро переписи) | int64 |
| education | Уровень образования в текстовом виде (например, Bachelors, HS-grad, 11th). | object |
| education-num | Уровень образования в числовом, упорядоченном виде (соответствует education) | int64 |
| marital-status | Семейное положение (например, Married-civ-spouse, Never-married, Divorced). | object |
| relationship | Родственная связь в семье (например, Husband, Not-in-family, Own-child, Wife). | object |
| race | Расовая принадлежность (например, White, Black, Asian-Pac-Islander). | object |
| sex | Пол (Male, Female). | object |
| capital-gain | Прирост капитала (доход от инвестиций и т.п., не связанный с зарплатой). | int64 |
| capital-loss | Убыток капитала. | int64 |
| hours-per-week | Количество рабочих часов в неделю. | int64 |
| native-country | Родная страна (например, United-States, Mexico, Philippines, ?). | object |
| occupation | Профессиональная сфера деятельности (например, Tech-support, Craft-repair, Other-service, ?). | object |
| income | Целевая переменная: уровень годового дохода (<=50K или >50K). | object |

# 2 АНАЛИЗ ДАННЫХ

**Основные характеристики датасета**

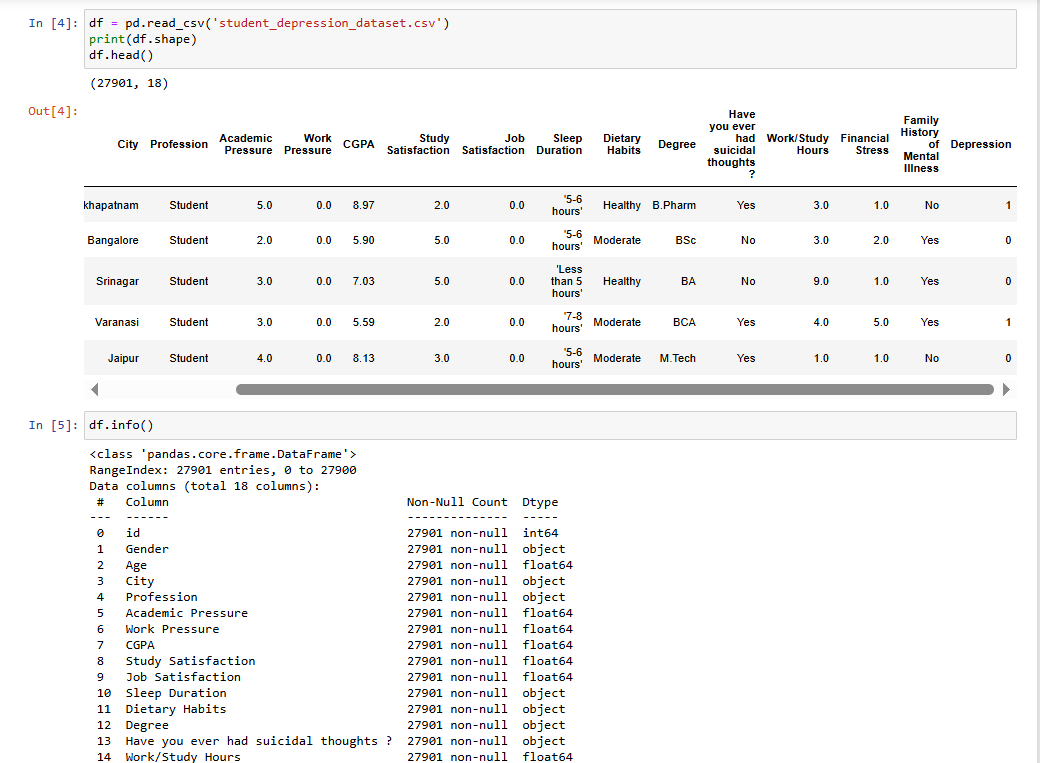
Целью настоящего исследования является разработка и оценка эффективности различных моделей машинного обучения, предназначенных для прогнозирования наличия симптомов депрессии у студентов. Задача формулируется как бинарная классификация: определение, присутствуют ли у студента симптомы депрессии ("Yes") или отсутствуют ("No"). 

Рисунок 1. Основные характеристики датасета

В анализируемом наборе данных нет пропусков.

На Рисунке 3 представлены основные характеристики числовых столбцов датасета.

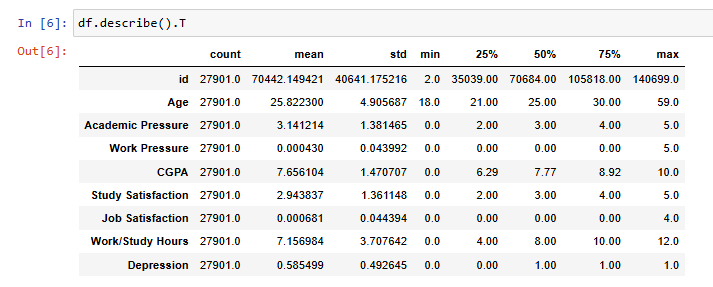


Рисунок 3. Основные характеристики числовых столбцов датасета

**Баланс классов**

Из описания набора данных мы знаем, что присутствует дизбаланс классов, в наборе данных 58% принадлежат к классу «Yes» и 42% – к классу «No».

**Визуальный анализ датасета**

Распределение значений целевой переменной проиллюстрировано на Рисунке 4. Распределение значений числовых переменных представлено на Рисунке 5. Распределение значений категориальных переменных представлено на Рисунке 6. Стоит отметить, что распределения значений числовой переменной близки к нормальному.

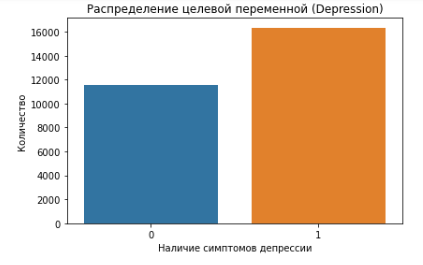


Рисунок 4. Распределение значений целевой переменной

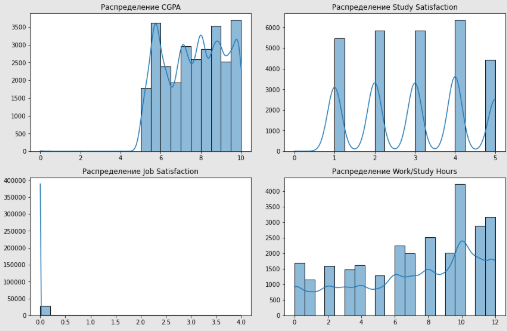


Рисунок 5. Распределение значений числовых переменных



Рисунок 6. Распределение значений категориальных переменных

**Преобразования категориальных признаков**

Поскольку большинство признаков являются категориальными и содержат множество значений, было решено произвести WoE-трансформацию для всех категориальных признаков (см. Рисунок 9), кроме целевой переменной, для ее преобразования был использован LabelEncoder (см. Рисунок 8). После преобразования классу «Yes» соответствует значение 1, а классу «No» - значение 0.

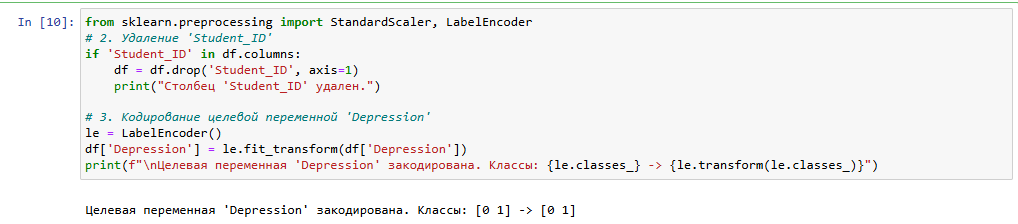


Рисунок 8. Кодирование целевой переменной

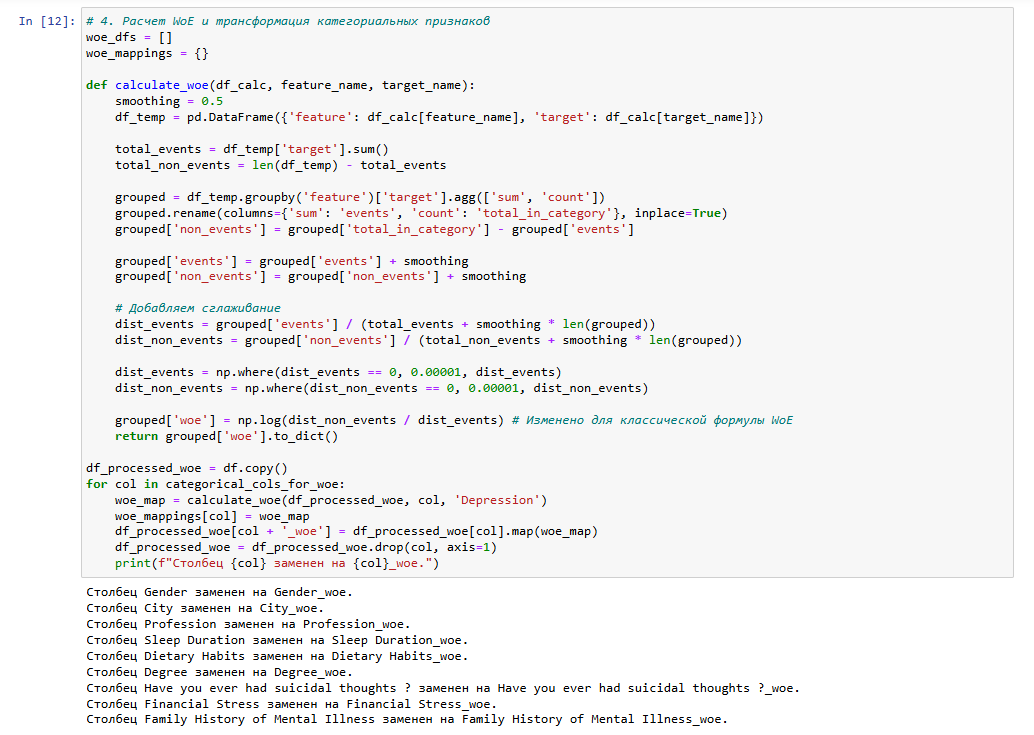


Рисунок 9. Код WoE-трансформации

**Масштабирование данных**

Для числовых данных было произведено масштабирование (см. Рисунок 10).

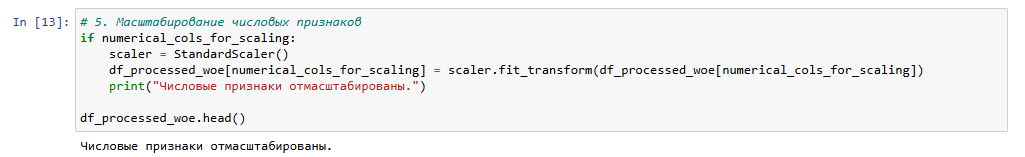


Рисунок 10. Масштабирования числовых признаков

**Однофакторный анализ**

Был произведен однофакторный анализ между всеми признаками и целевой переменной. В качестве метрики отбора был выбран коэффициент Somers’D, при значениях модуля которого <0,1 признак признавался недостаточно информативным и отбрасывался, и не был использован в качестве финального признака (см. Рисунок 11, Рисунок 12).

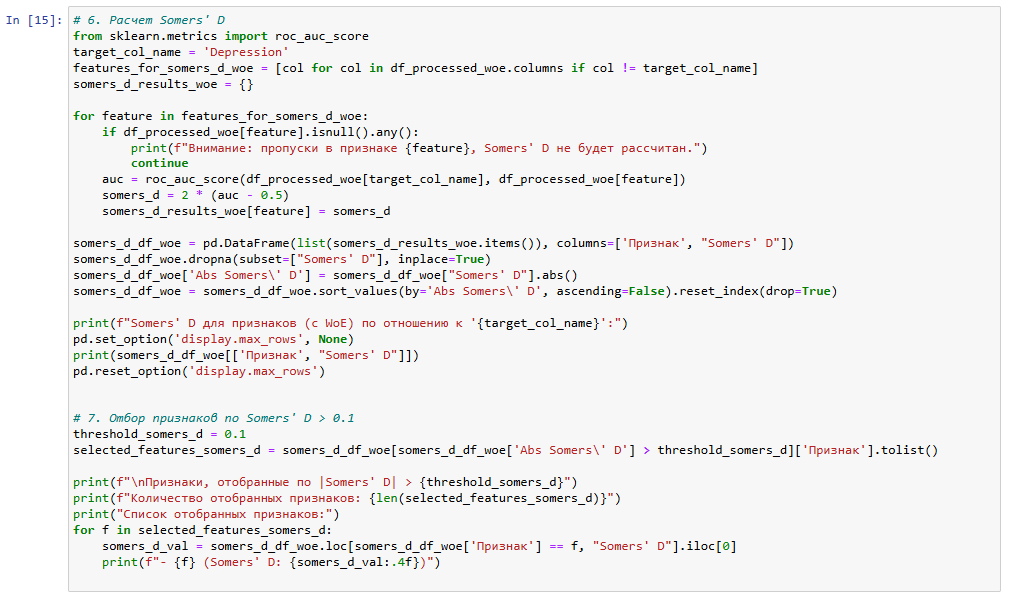
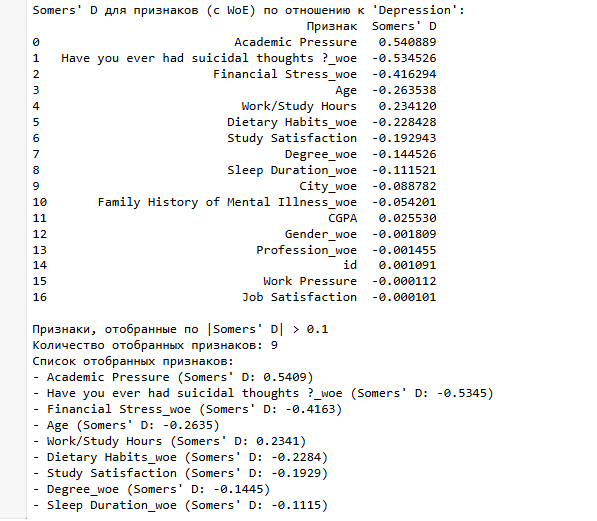


Рисунок 11. Рассчет Somers’D



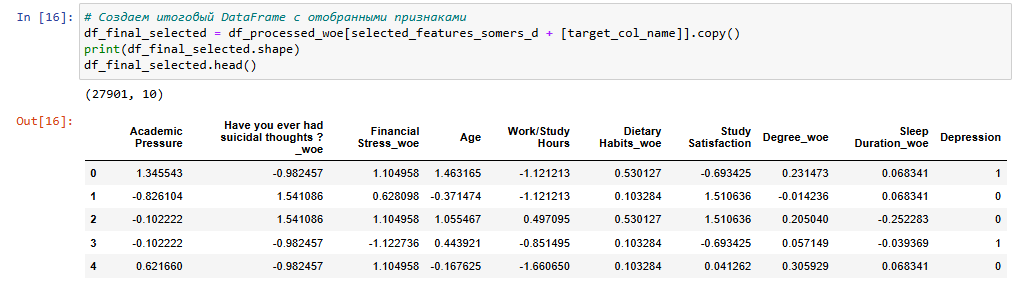


Рисунок 12. Отбор предфинального набора признаков

**Многофакторный анализ**

Для оставшихся признаков был подсчитан коэффициент корреляции Пирсона и была построена heat-map (см. Рисунок 13).Сильной корреляции не было замечено, все признаки будут использоваться для моделей. (см. Рисунок 14).



Рисунок 13. 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 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. ***Recall***

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

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.

*В качестве основной метрики оценки качества была выбрана метрика ROC AUC.*

**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 после всех преобразований содержит только числовые признаки;

выберем модели:

* Логистическая регрессия;
* Дерево решений;
* Алгоритм KNN для классификации;
* Случайный лес;
* XGBoost(один из представителей GBM).

# 4 ФОРМИРОВАНИЕ ОБУЧАЮЩЕЙ И ТЕСТОВОЙ ВЫБОРОК

Разделим выборку на обучающую и тестовую с помощью train\_test\_split, учтем дисбаланс классов (см. Рисунок 15).

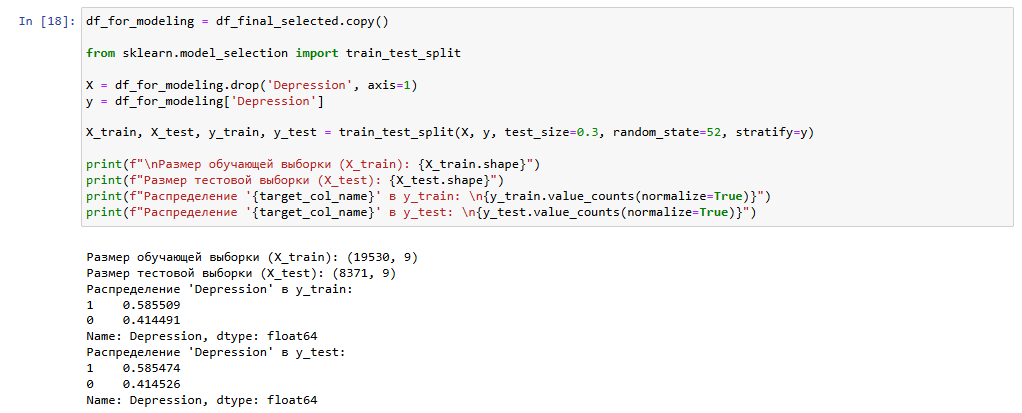


Рисунок 15. Разделение выборки

# 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 16). Используем модели из библиотек scikit-learn и xgboost.

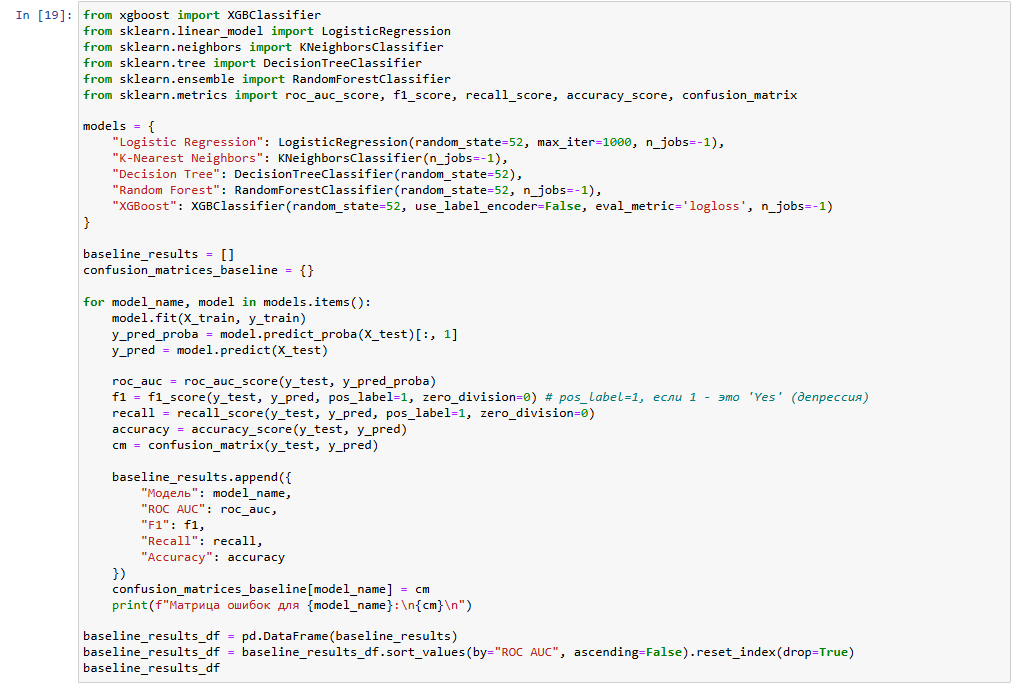


Рисунок 16. Обучение моделей без подбора гиперпараметров

**Оценка качества моделей**

Оценим качество работы моделей. На Рисунке 17 приведены 4 метрики: ROC-AUC, Accuracy, Recall и F1-мера.

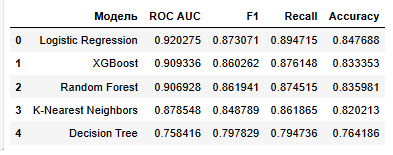


Рисунок 17. Оценка качества работы моделей baseline-решения

Лучшие результаты по основной метрике показала модель Logistic Regression. Худшие результаты у модели Decision Tree для классификации.

# 6 ПОДБОР ГИПЕРПАРАМЕТРОВ

Производится подбор гиперпараметров для выбранных моделей.

**Обучение моделей**

Обучим выбранные модели, подбирая гиперпараметры с помощью GridSearchCV и кросс-валидации по методу StratifiedK-Fold (см. Рисунок 18, Рисунок 19).



Рисунок 18. Обучение моделей с подбором гиперпараметров



Рисунок 19. Обучение моделей с подбором гиперпараметров

**Оценка качества моделей**

Оценим качество работы моделей с подбором гиперпараметров. На Рисунке 20 приведены 4 метрики: ROC-AUC, Accuracy, Recall и F1-мера.

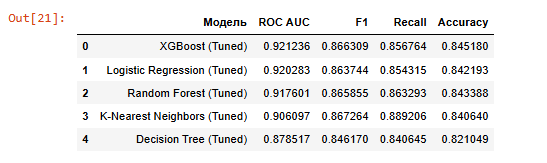


Рисунок 20. Оценка качества работы моделей оптимального решения

Мы видим, что лучшие результаты относительно основной метрики показывает классификатор XGBoost. Подбор гиперпараметров не позволил значительно улучшить качество работы моделей.

Стоить отметить, что после подбора гиперпараметров некоторые метрики упали, пусть и незначительно.

# 7 СРАВНЕНИЕ РЕШЕНИЙ

На Рисунке 21 и Рисунке 22 представлены столбчатые диаграммы позволяющие сравнить AUC ROC и F1-меру моделей базового и оптимального решения.

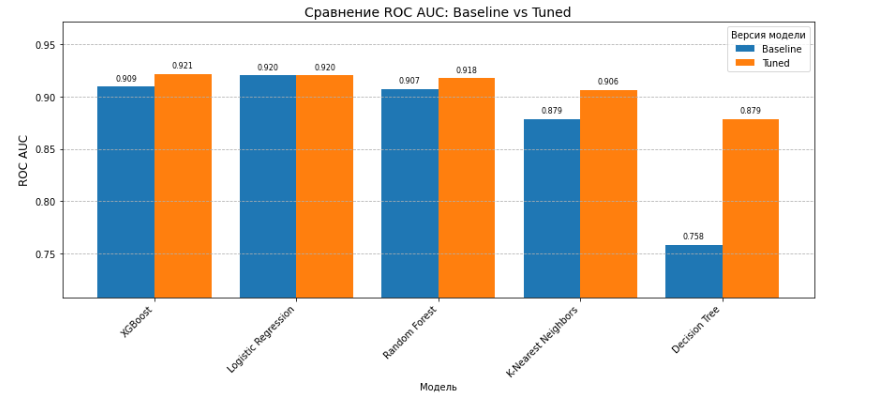


Рисунок 21. Сравнение решений по ROC AUC

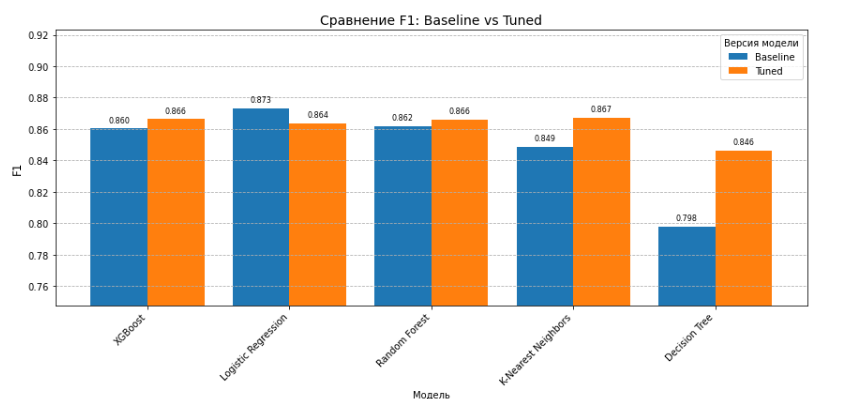


Рисунок 22. Сравнение решений по F1-мере

# ЗАКЛЮЧЕНИЕ

В рамках научно-исследовательской работы была решена задача бинарной классификации – определение наличия депрессии у человека. В ходе исследования были получены два решения: базовое и оптимальное за счет подбора гиперпараметров.

Наилучшего качества классификации по выбранным метрикам достигла модель XGBoost после подбора гиперпараметров.

Построенные модели показывают достаточно хорошие результаты работы, позволяющие с достаточной уверенностью определять людей с депрессией по их данным. Существует некоторое пространство для улучшения, связанное с использованием большего количества признаков и более тщательной настройки гиперпараметров.

# 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).