МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования   
«Пермский государственный национальный исследовательский университет»

Физико-математический институт

Отчёт по дисциплине

«Технологии машинного обучения и нейронные сети для решения прикладных задач»

|  |  |  |
| --- | --- | --- |
|  |  | **Выполнили:**  **студенты 4 курса физико-математического института**  **Косберг А.Д.**  **Кузнецов А.П.** |
|  |  |  |
|  |  |  |

Пермь 2024 год

Введение

Задача:

1. Выполнить анализ проблемы, обосновать ее актуальность.
2. Осуществить загрузку данных и подготовку их к анализу.
3. Выполнить предварительный анализ данных, в том числе корреляционный анализ.
4. Осуществить моделирование зависимости целевого признака от остальных с помощью различных классификаторов, в том числе подобрать наилучшую модель, оценить ее качество и выполнить прогнозирование.
5. Сделать выводы о значимости различных факторов на целевую переменную.

Выбрать лучшую модель и сделать выводы о достижении цели

**Анализ проблемы исследования**

В современном мире, где здоровье и качество жизни становятся все более важными аспектами, проблема ожирения приобретает глобальный характер. Ожирение не только влияет на физическое состояние человека, но и связано с риском развития множества заболеваний, таких как диабет, сердечно-сосудистые заболевания и другие хронические патологии. В связи с этим, актуальность исследования и прогнозирования уровня ожирения на основе различных факторов становится очевидной. Это обусловлено следующими ключевыми аспектами:

1. **Рост распространенности ожирения**  
   За последние десятилетия уровень ожирения среди населения значительно вырос. Это связано с изменением образа жизни, снижением физической активности, увеличением потребления высококалорийной пищи и другими социально-экономическими факторами. Прогнозирование уровня ожирения позволяет выявить группы риска и разработать профилактические меры для предотвращения развития заболевания.
2. **Влияние ожирения на здоровье и качество жизни**  
   Ожирение является одним из основных факторов, способствующих развитию хронических заболеваний. Оно также влияет на психологическое состояние человека, снижая самооценку и увеличивая риск депрессии. Прогнозирование уровня ожирения на основе данных о питании, физической активности и других факторов может помочь в разработке индивидуальных рекомендаций для улучшения здоровья и качества жизни.
3. **Роль образа жизни и привычек**  
   На уровень ожирения влияют множество факторов, включая пищевые привычки, уровень физической активности, генетическую предрасположенность и образ жизни. Анализ данных, таких как частота употребления высококалорийной пищи, потребление овощей, уровень физической активности и другие, позволяет выявить ключевые факторы, способствующие развитию ожирения, и предложить меры для их коррекции.
4. **Развитие технологий и методов анализа данных**  
   Современные методы машинного обучения предоставляют уникальные возможности для анализа больших объемов данных и выявления скрытых закономерностей. Прогнозирование уровня ожирения на основе данных о человеке (возраст, пол, пищевые привычки, физическая активность и т.д.) может стать важным инструментом для врачей, диетологов и специалистов по здоровому образу жизни. Это позволит не только улучшить диагностику, но и разработать персонализированные подходы к профилактике и лечению ожирения.

Таким образом, актуальность исследования прогнозирования уровня ожирения на основе данных о человеке обусловлена необходимостью предотвращения роста заболеваемости, улучшения качества жизни и разработки эффективных профилактических мер. Использование методов машинного обучения для анализа данных о питании, физической активности и других факторах может стать важным шагом в борьбе с ожирением и его последствиями. Разработка моделей, способных точно предсказывать уровень ожирения, позволит не только улучшить понимание проблемы, но и предложить конкретные рекомендации для снижения рисков, связанных с этим заболеванием.

Исходные данные

В дальнейшей работе используется набор данных из датасета: **«ObesityDataSet\_raw\_and\_data\_sinthetic.csv»**

Датасет содержит следующий набор фич:

- Gender [Male|Female] – Пол (мужской или женский);

- Age [<int>] – Возраст человека (в годах);

- Height [<float>] – Рост человека (в метрах);

- Weight [<float>] – Вес человека (в килограммах);

- family\_history\_with\_overweight [yes|no] – Есть ли у человека семейная история избыточного веса (да/нет);

- FAVC [yes|no] – Часто ли человек употребляет высококалорийную пищу (да/нет);

- FCVC [1.0|2.0|3.0] – Частота потребления овощей (шкала от 1 до 3);

- NCP [<float>] – Количество основных приемов пищи в день;

- CAEC [no|Sometimes|Frequently|Always] – Частота употребления пищи между приемами пищи (никогда, иногда, часто, всегда).

- SMOKE [yes|no] – Курит ли человек (да/нет).

- CH2O [1.0|2.0|3.0] – Суточное потребление воды (шкала от 1 до 3).

- SCC [yes|no] – Следит ли человек за потреблением калорий (да/нет).

- FAF [1.0|2.0|3.0] – Частота физических нагрузок (шкала от 0 до 3).

- TUE [0.0|1.0|2.0|3.0] – Время, затраченное на использование технологии (шкала от 0 до 3).

- CALC [no|Sometimes|Frequently|Always] – частота употребления алкоголя (никогда, иногда, часто, всегда).

- MTRANS [Automobile|Bike|MotorbikeFrequently|Public\_Transportation|Walking] – Основной вид транспорта (автомобиль, велосипед, мотоцикл, общественный транспорт, ходьба).

Целевая переменная:

- NObeyesdad [Insufficient\_Weight|Normal\_Weight|Overweight\_Level\_I|Overweight\_Level\_II|Obesity\_Type\_I|Obesity\_Type\_II|Obesity\_Type\_III] – уровень ожирения (недостаточный вес, нормальный вес, избыточный вес I уровня, избыточный вес II уровня, ожирение I типа, ожирение II типа, ожирение III типа).

Реализация проекта

1. Загрузка данных в датафрейм:

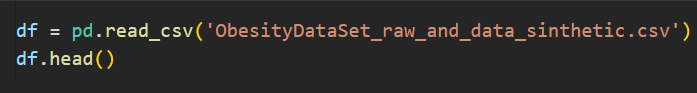


Рисунок 1. Код загрузки данных в проект

Информация о датасете:

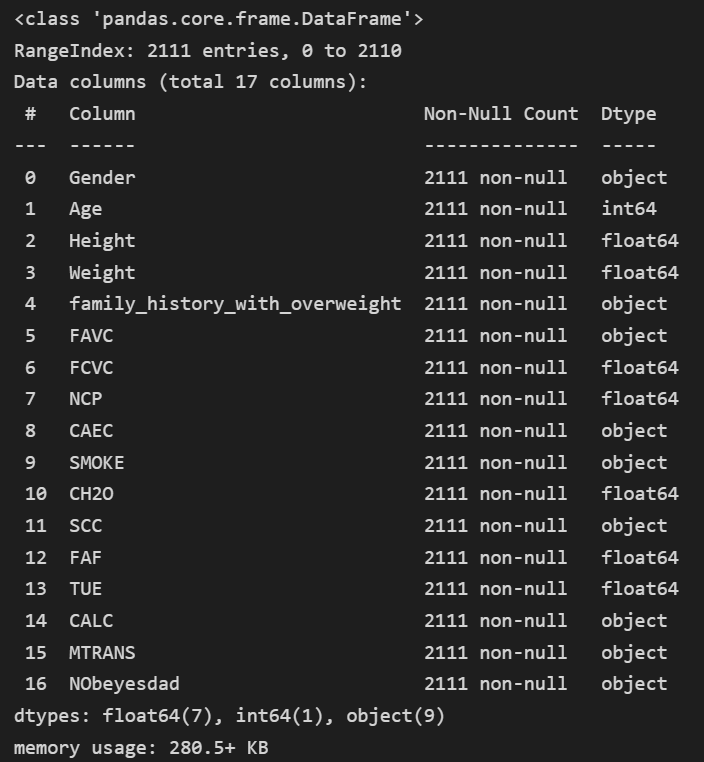


Рисунок 2. Информация о датасете.

Вывод информации о датасете показывает, что пропущенные данные отсутствуют.

Статистика по числовым значениям:



Рисунок 3. Статистика по числовым значениям.

Статистика по категориальным значениям:

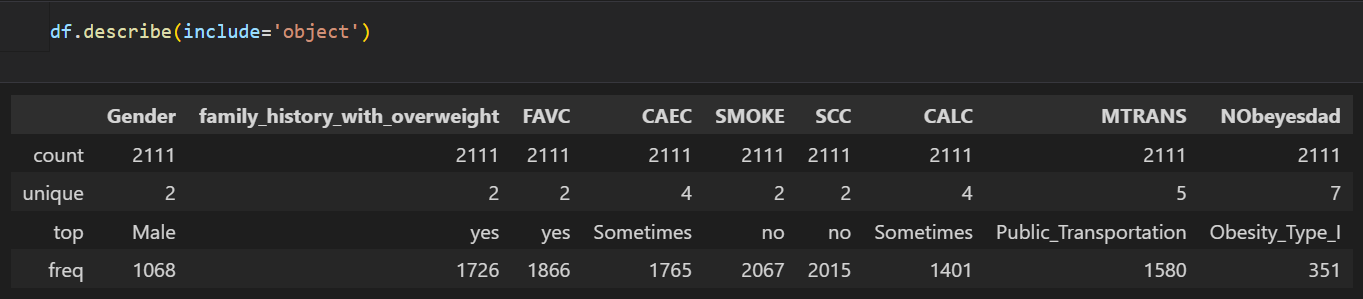


Рисунок 4. Статистика по категориальным значениям.

Проверим категориальные признаки на соответствие описанию датасета:



Рисунок 5. Вывод программы, отражающий соответствие категориальных признаков на соответствие описанию датасета.

**Проверка сбалансированности датасета:**

Количество элементов в каждой категории «GENDR»

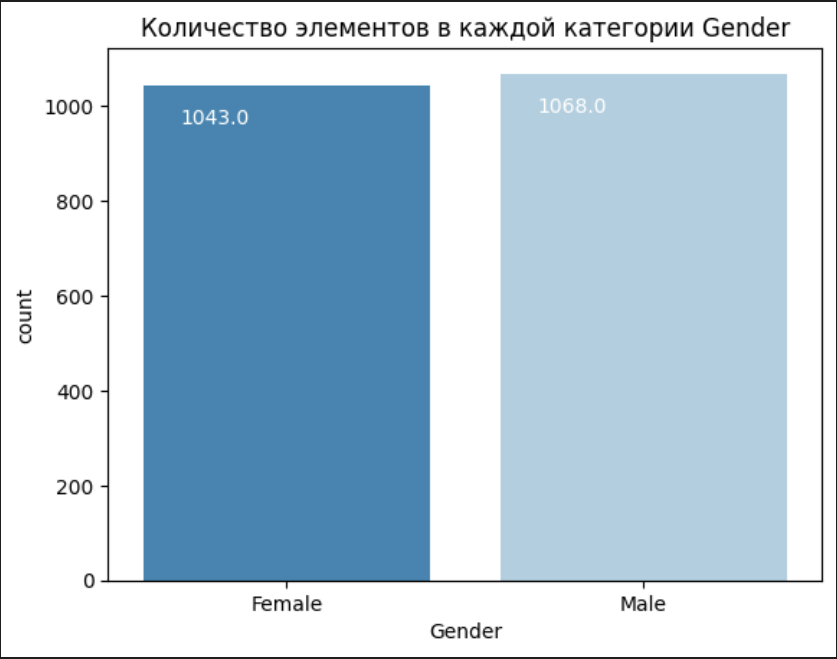


Рисунок 6. Количество объектов, разделенных по категории «Пол»

На рисунке NNNNNNN показано количество элементов в категории Gender. Баланс по полу респондента хороший.

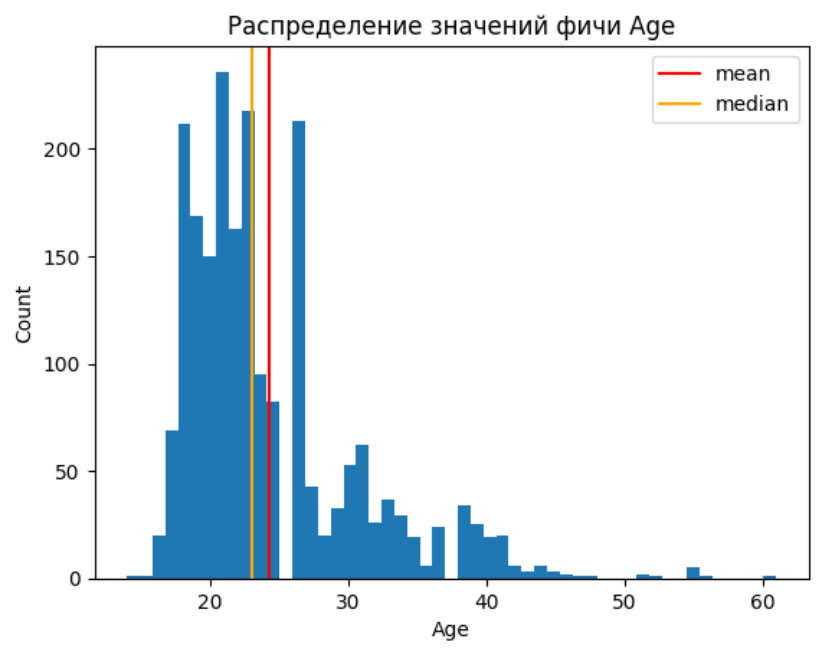
Распределение значение фичи «Age»

Рисунок 7. Распределение значений параметра «Age»(Возраст)

Нормального распределения не наблюдается - гистограмма перевешивает влево относительно медианы и среднего

Распределение значение фичи «Рост»

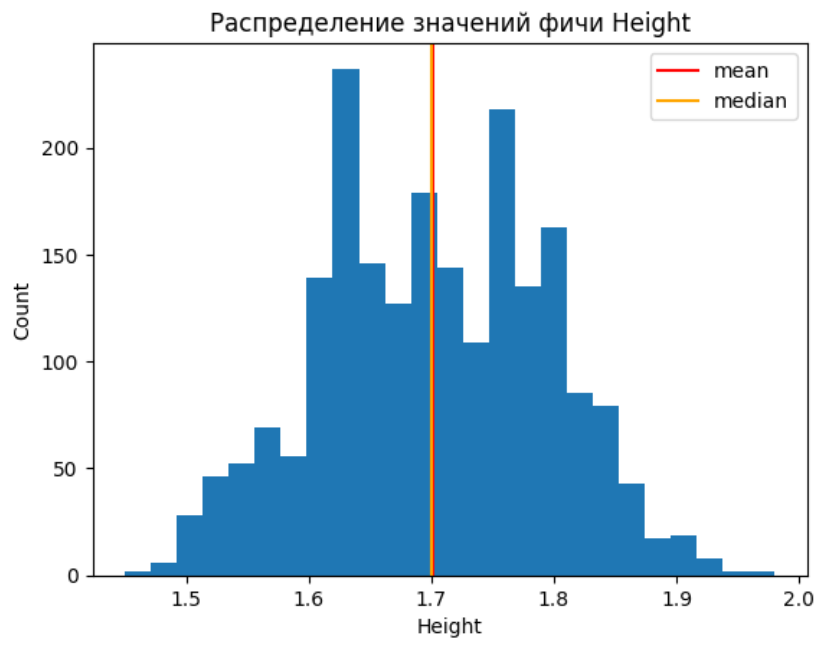


Рисунок 8. Распределение роста объектов.

Большинство людей имеют рост 1,60 м - 1,85 м. И среднее, и медианное значения составляют около 1,70. Тем не менее, значения роста, не распределены по нормальному закону.

Распределение значений фичи «Вес»

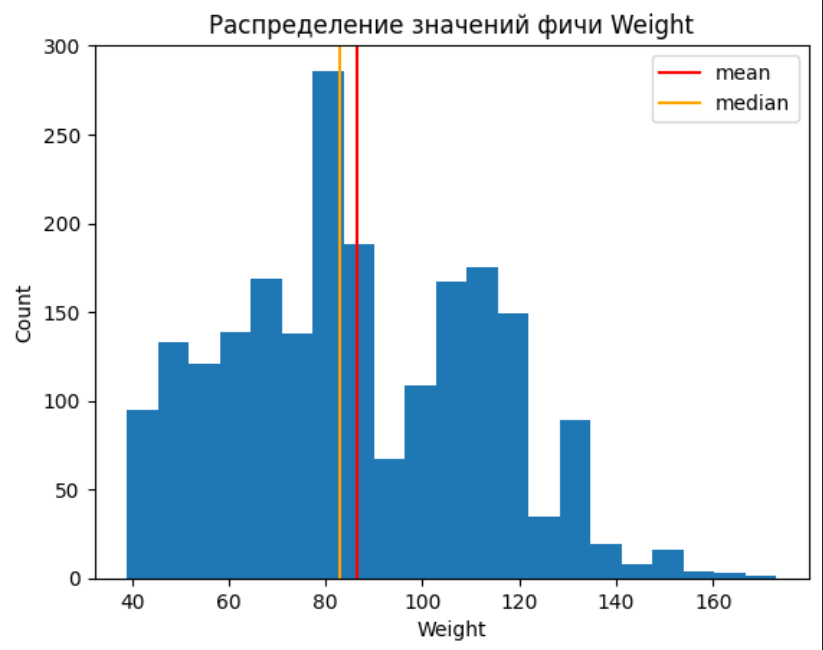


Рисунок 9. Распределение значений веса объектов.

Распределение значений веса не предлагает интересных наблюдений.

Интересно изучить связь между фичами «Рост» и «Вес», поскольку обе метрики используются при вычислении «Индекса массы тела» - которая напрямую указывает на степень ожирения. График, отображающий эту связь:

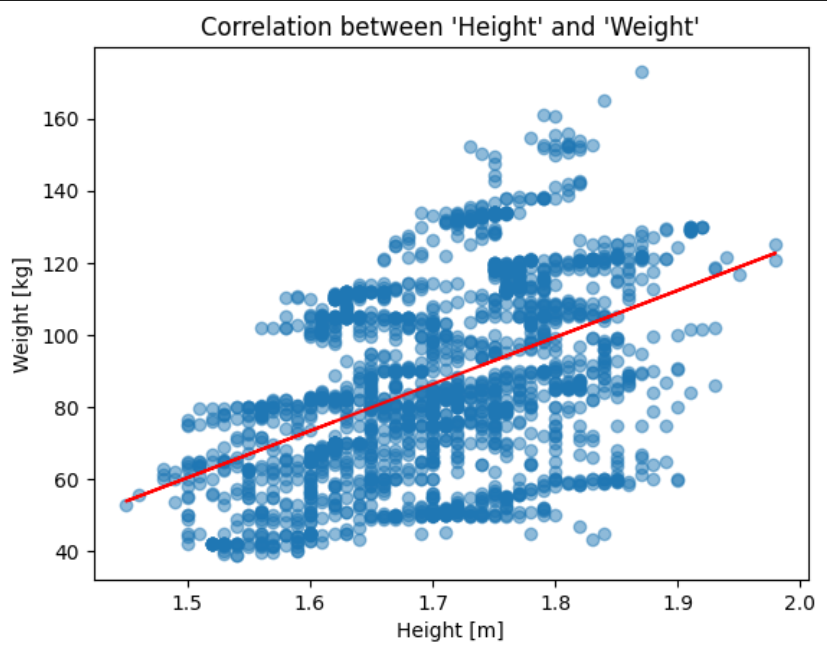


Рисунок 10. Визуализация зависимости веса объектов от роста.

Красная линия показывает, что между фичами «Рост» и «Вес» существует положительная корреляция, что означает, что увеличение одной переменной приводит к увеличению другой. Другими словами, более высокие люди с большей вероятностью будут весить больше.

Количество элементов в каждой категории фичи «Наследственная зависимость»:

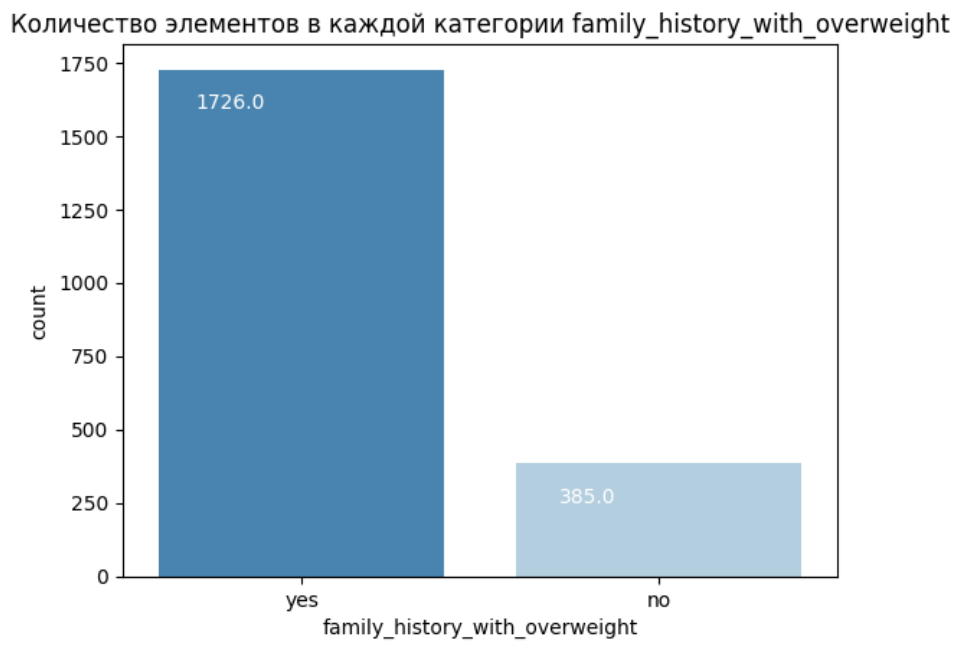


Рисунок 11. Количество элементов в каждой категории фичи «Наследственная зависимость»

Количество элементов в каждой категории фичи «Потребление высококалорийной пищи:

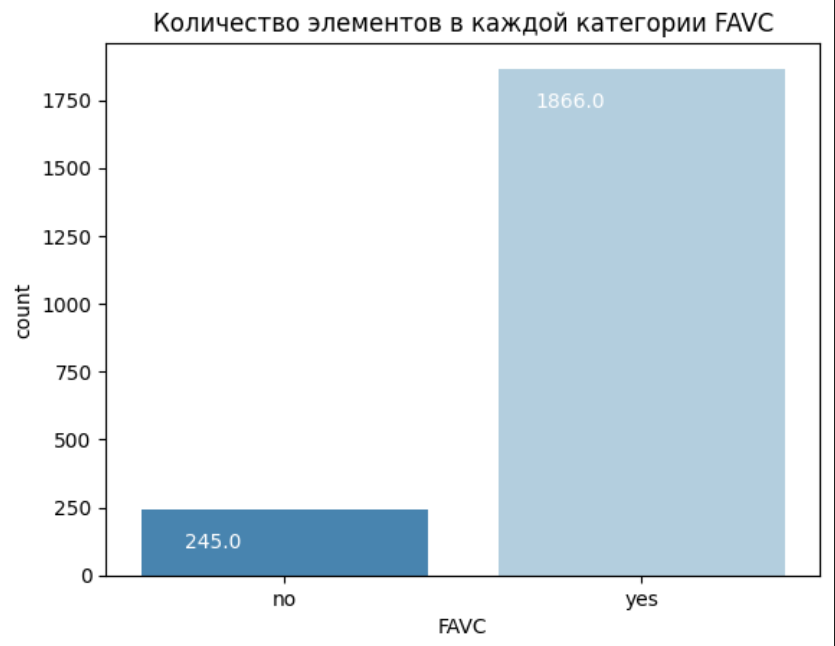


Рисунок 12.Количество элементов в каждой категории фичи «Потребление высококалорийной пищи»

Распределение значений фичи «Потребление овощей»

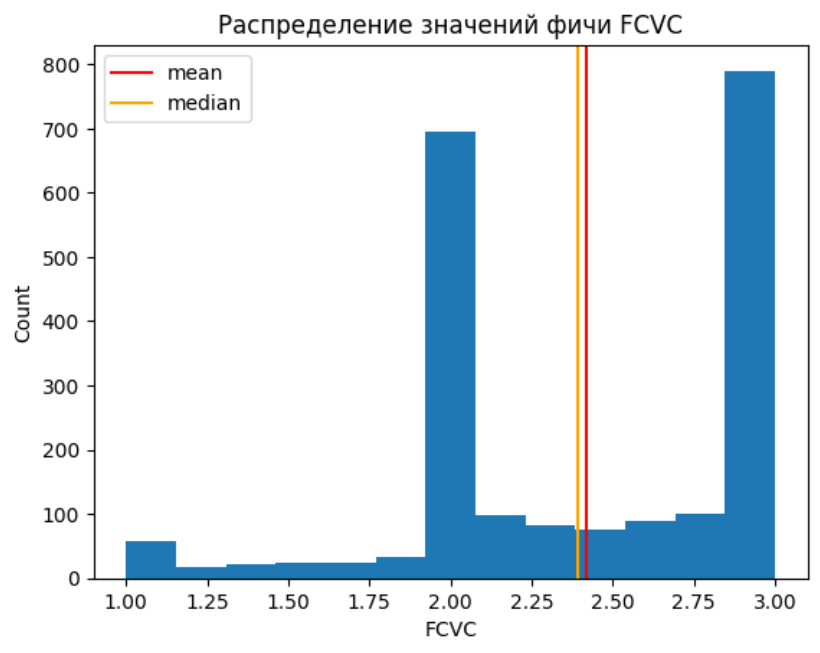


Рисунок 13.Количество элементов в каждой категории фичи «Потребление овощей»

Распределение значений фичи «Количество приемов пищи в день»:

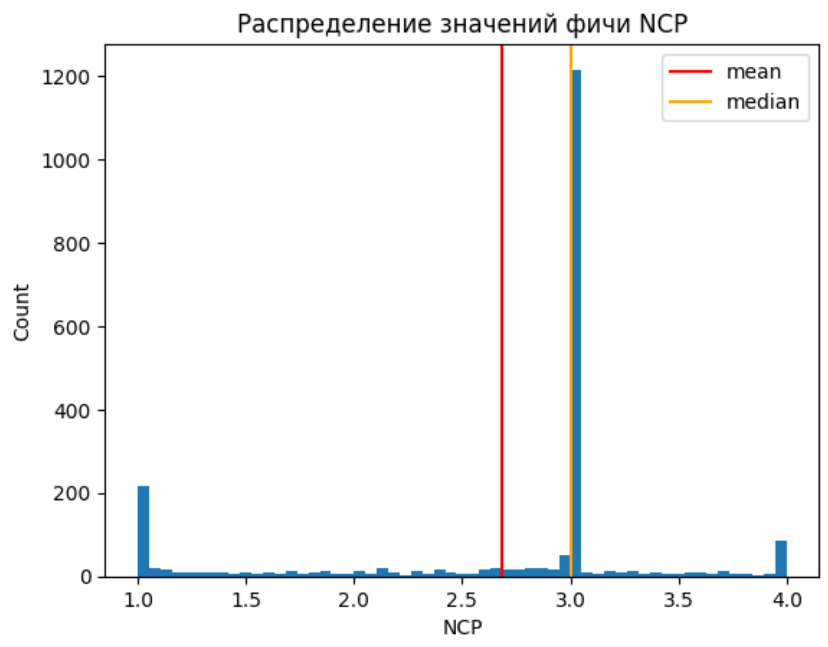


Рисунок 14. Распределение значений фичи «Количество приёмов пищи в день»

Наиболее часто встречающееся значение – 3.

Распределение значений фичи «Перекусы»:

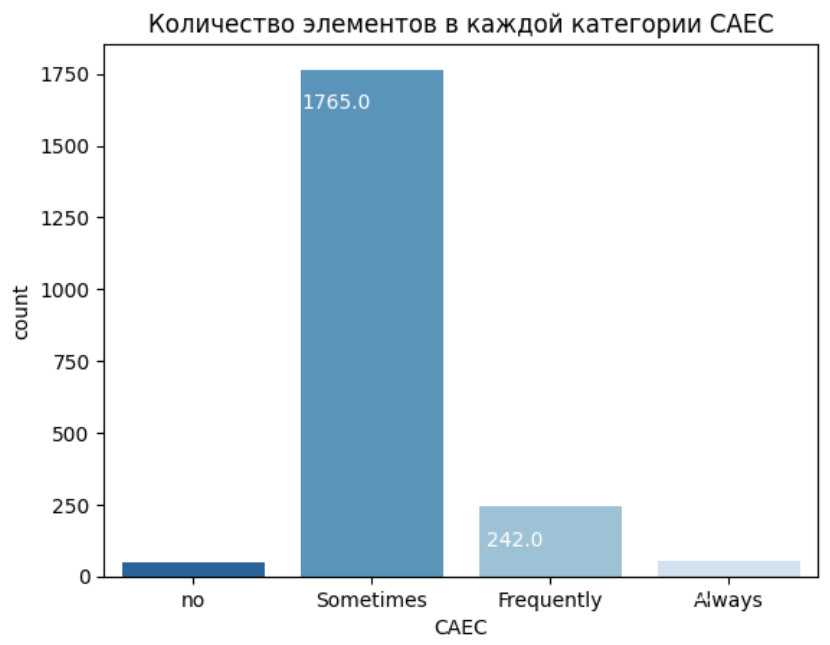


Рисунок 15. Распределение значений фичи «Перекусы»

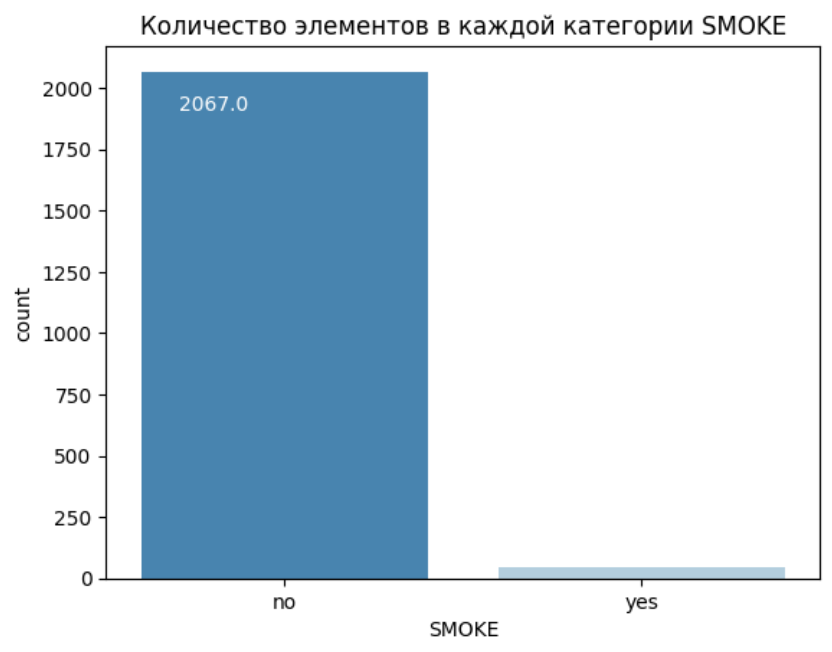
Количество элементов в каждой категории «Курение»  


Рисунок 16. Распределение значений фичи «Курение»

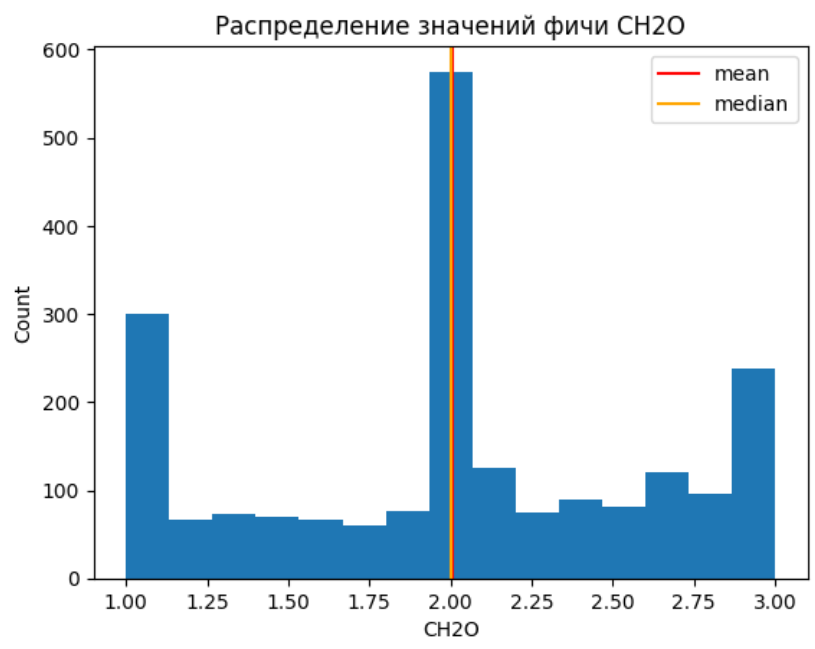
Распределение значений фичи «Потребление воды»  


Рисунок 17. Распределение значений фичи «Потребление воды»

Распределение значений фичи «Отслеживание потребления калорий»



Рисунок 18. Распределение значений фичи «Отслеживание потребления калорий»

Распределение значений фичи «Физическая активность»

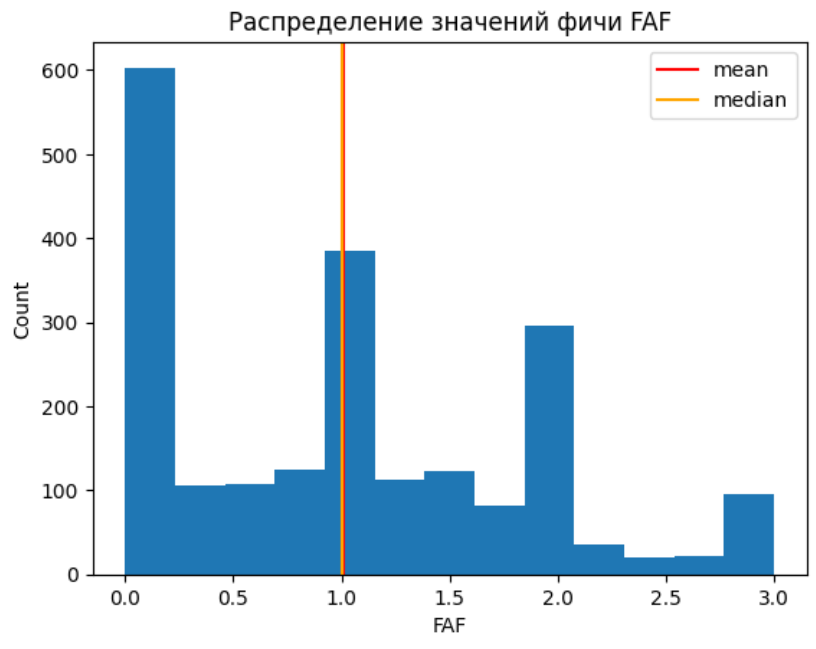


Рисунок 19. Распределение значений фичи «Физическая активность»

Распределение значений фичи «Использование гаджетов»

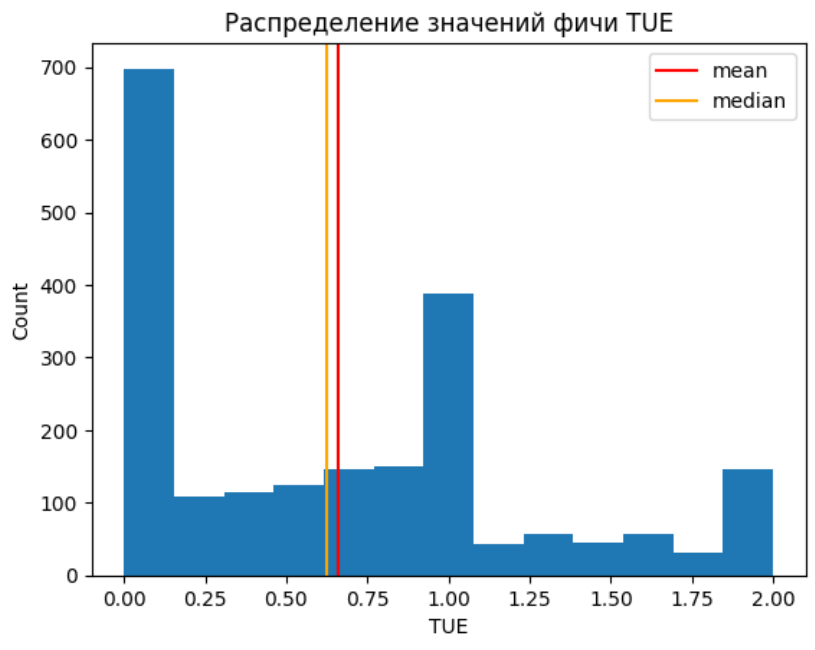


Рисунок 20. Распределение значений фичи «Использование каджетов»

Распределение значений фичи «Потребление алкоголя»

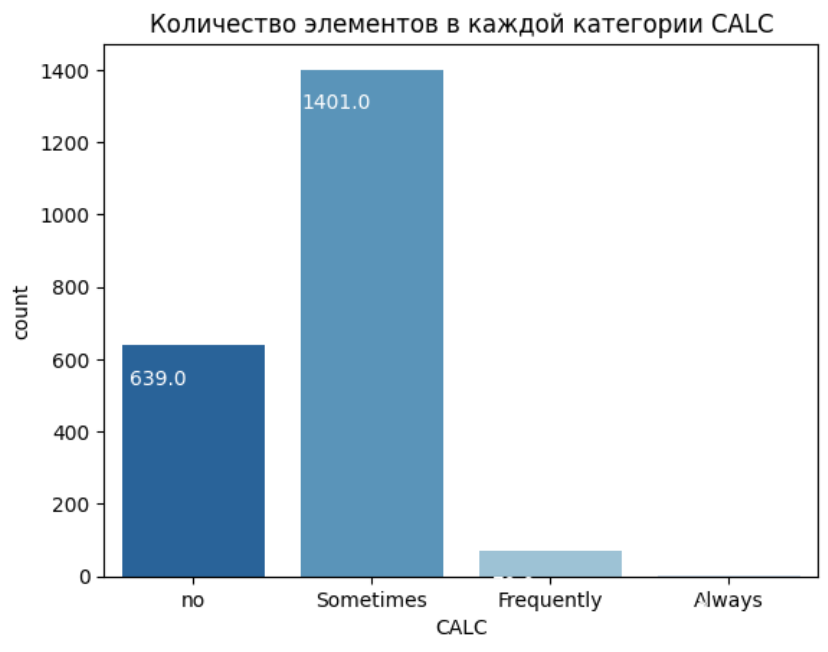


Рисунок 21. Распределение значений фичи «Потребление алкоголя»

Распределение значений фичи «Вид передвижения»

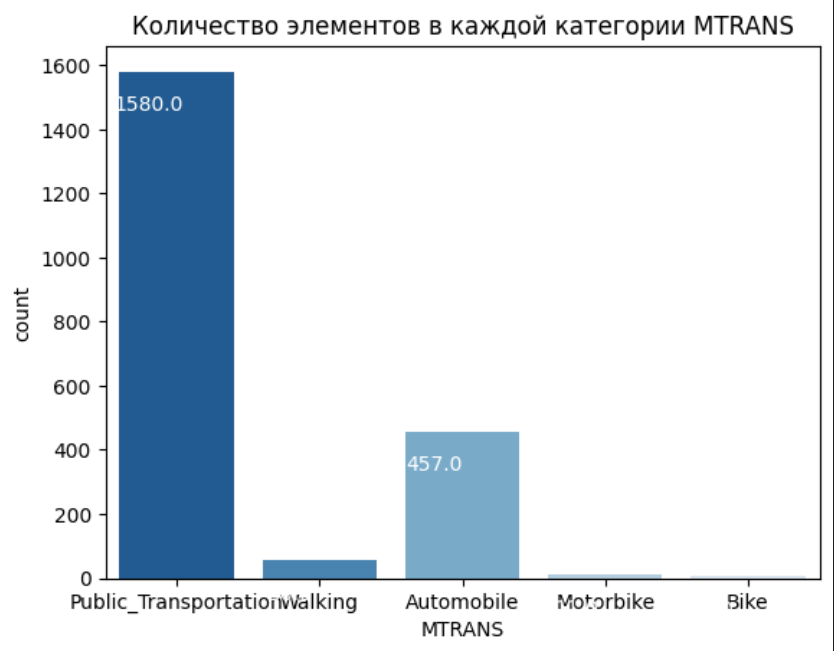


Рисунок 22. Распределение значений фичи «Вид передвижения»

1. Индекс массы тела:

Люди, в соответствии с их `индексом массы тела (ИМТ)`, классифицируются как:

\* Недостаточный вес, если ИМТ < 18,5

\* Нормальный, если ИМТ 18,5 - 24,9

\* Избыточный вес, если ИМТ 25,0 - 29,9

\* Ожирение I, если ИМТ 30,0 - 34,9

\* Ожирение II, если ИМТ от 35,0 до 39,9

\* Ожирение III, если ИМТ > 40

Количество людей в каждой категории отображается ниже (примечание: категории упорядочены логически).

График показывает, что набор данных сбалансирован; только класс "Ожирение типа I" немного превосходит другие категории.

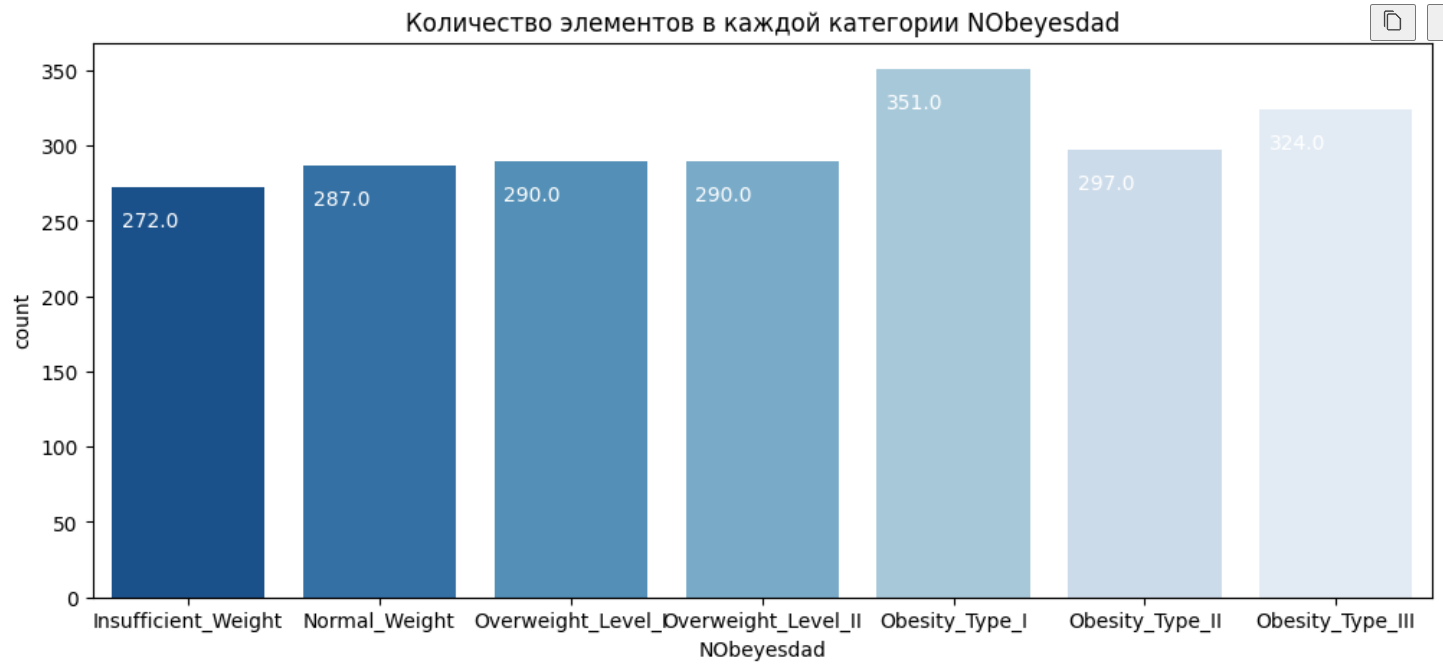


Рисунок 23. Распределение количества элементов в каждой категории веса

1. Зависимости Фич:

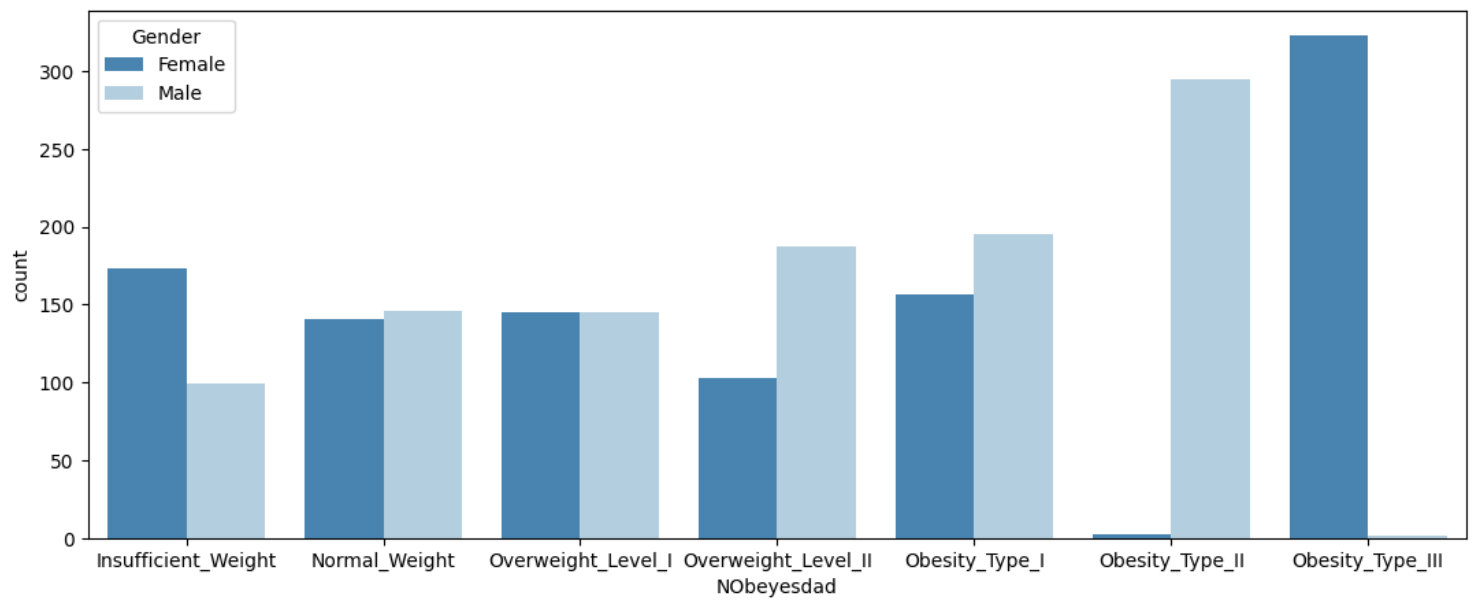


Рисунок 24. Зависимость лишнего веса от пола

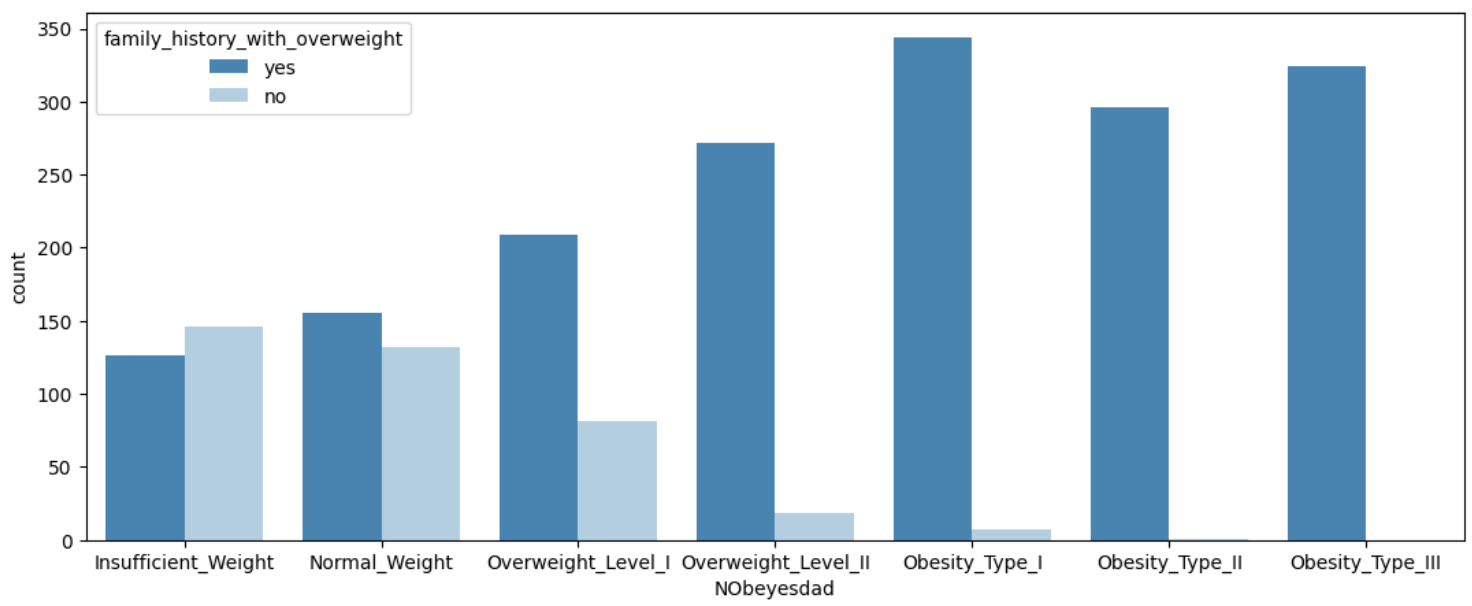


Рисунок 25. Зависимость лишнего веса наследственности

У всех, кто был отнесен к категории с избыточным весом или ожирением, были родственники, страдающие от проблем с весом.

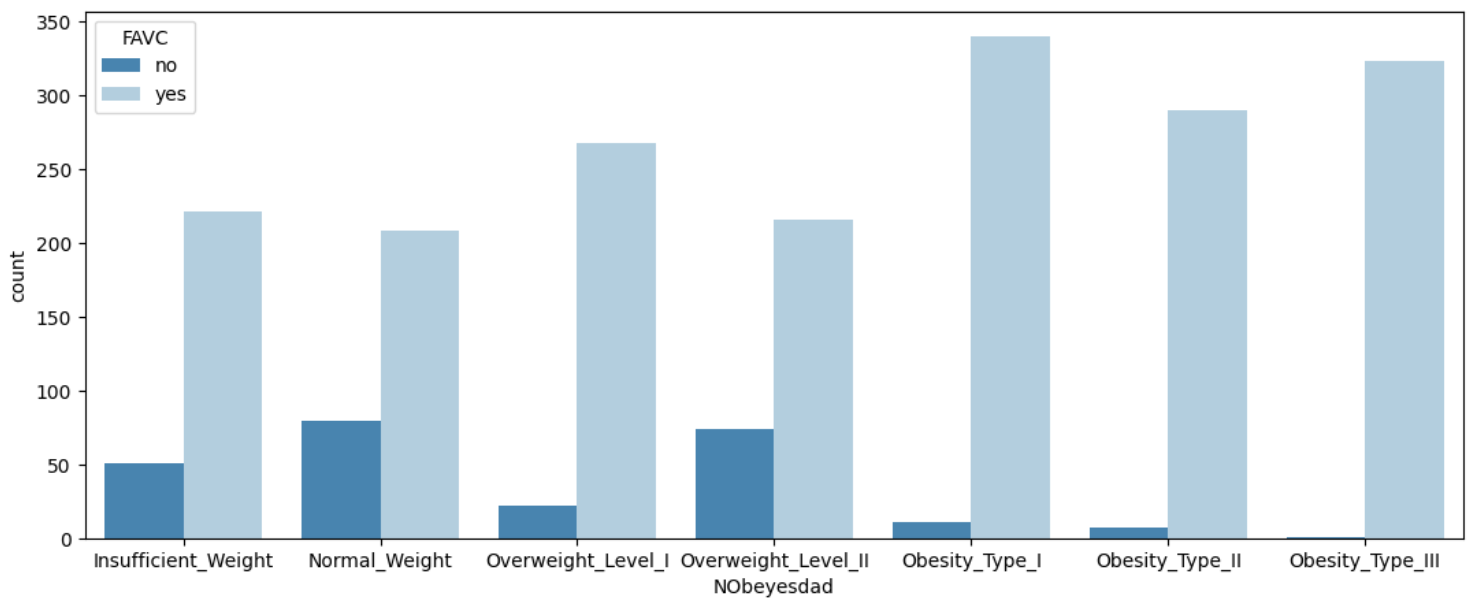


Рисунок 26. Зависимость лишнего веса и калорийности пищи

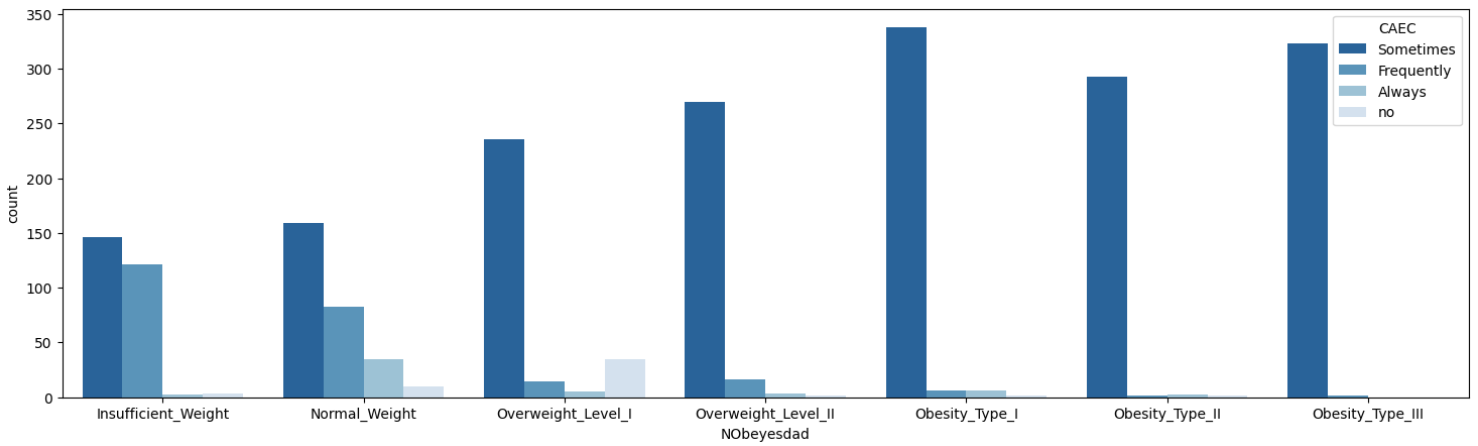


Рисунок 27. Зависимость лишнего веса и частоты перекусов

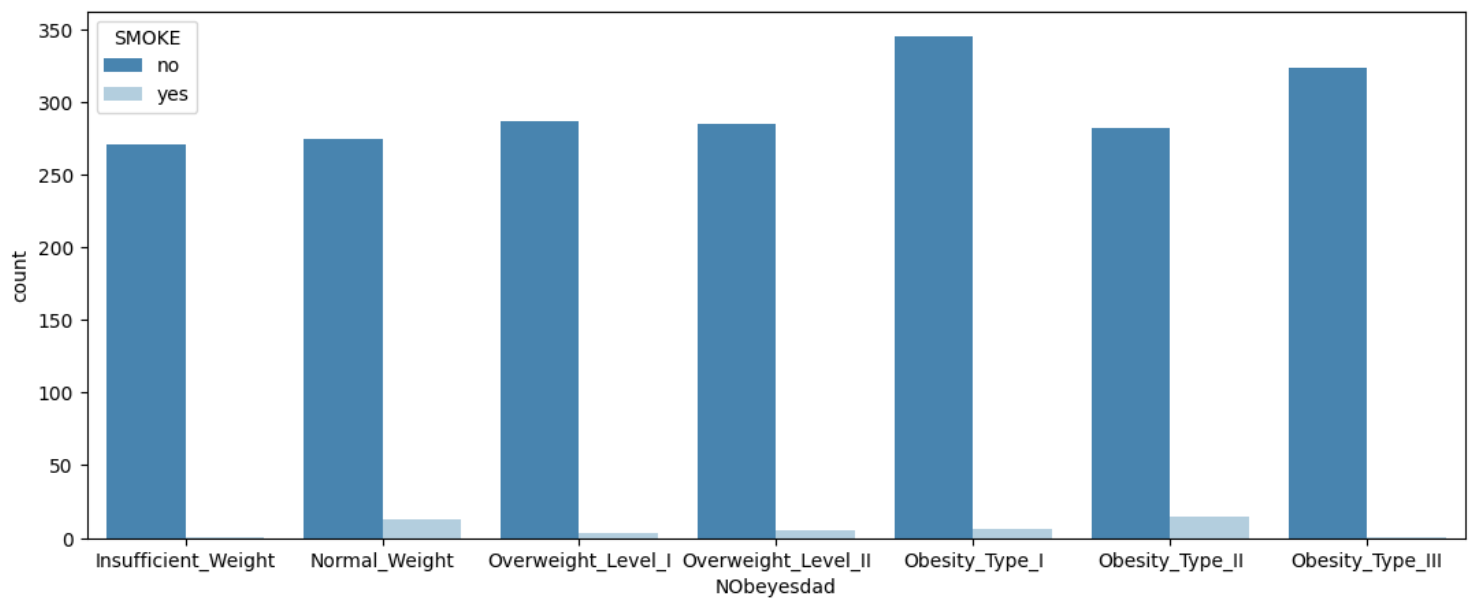


Рисунок 28. Зависимость лишнего веса и курения

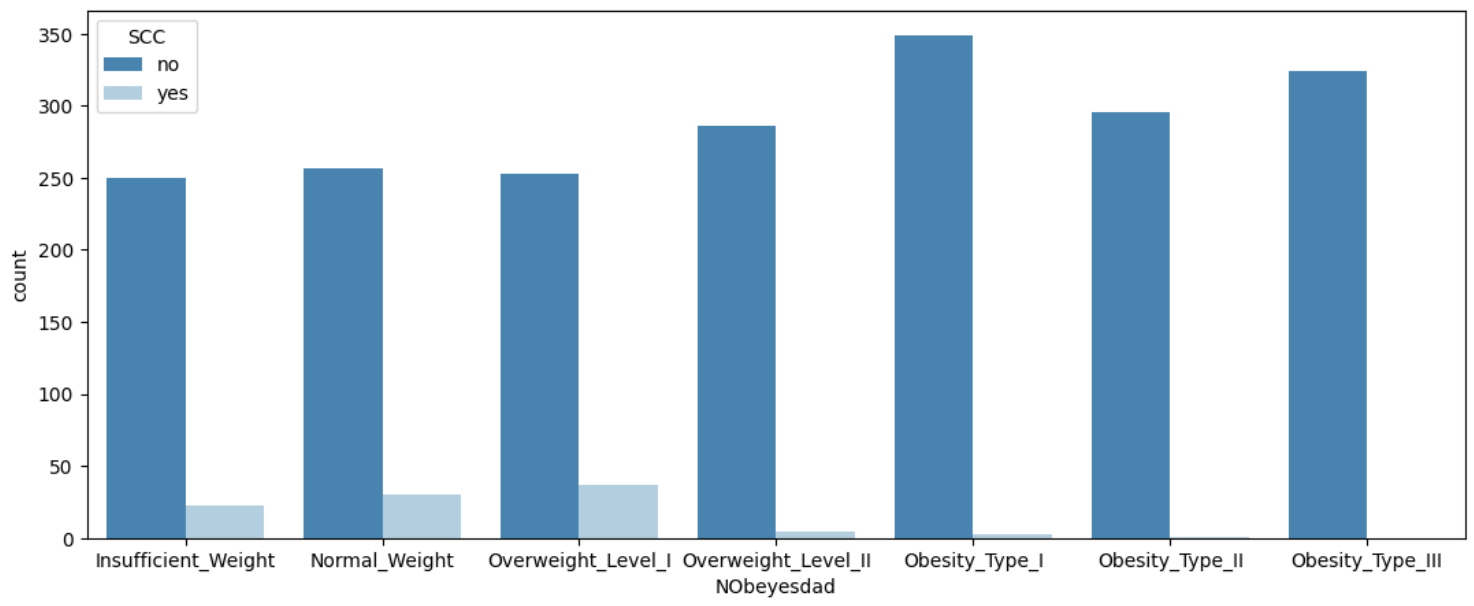


Рисунок 29. Зависимость лишнего веса и Отслеживания калорий

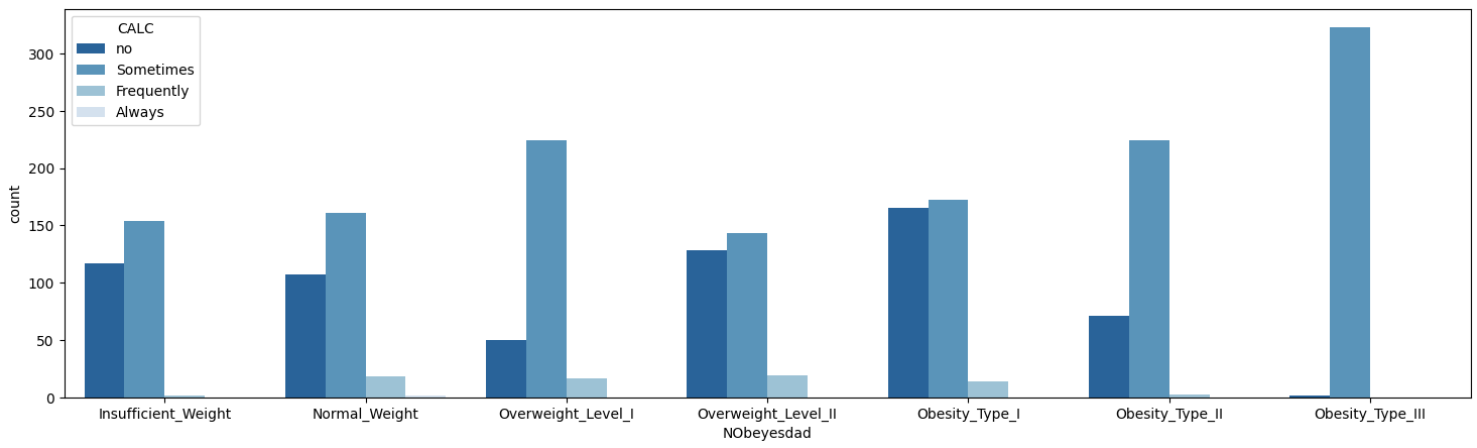


Рисунок 30. Зависимость лишнего веса и употребления алкоголя

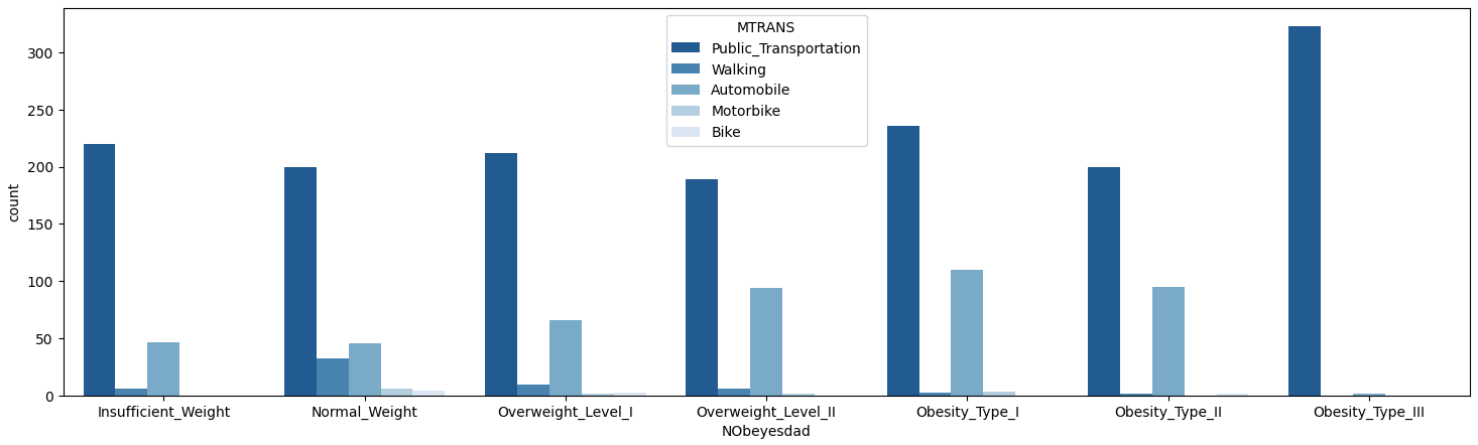


Рисунок 31. Зависимость лишнего веса и средств передвижения

Транспорт, похоже, не (значительно) влияет на вес человека. Общественным транспортом пользуются как худые, так и нормальные и полные люди; все группы также полагаются на автомобили.

Корреляция:

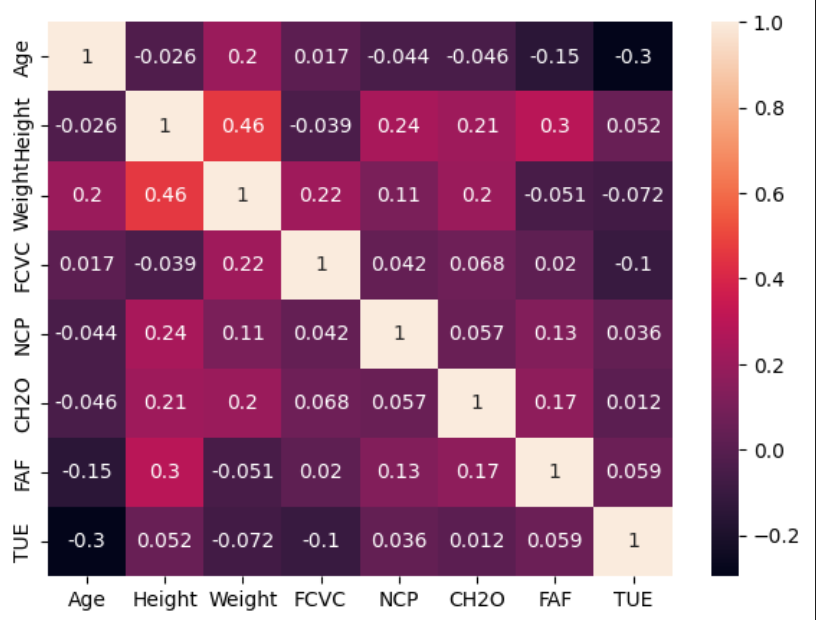


Рисунок 32. Матрица корреляции признаков.

1. Подготовка к применению модели:

* Замена категориальных параметров:

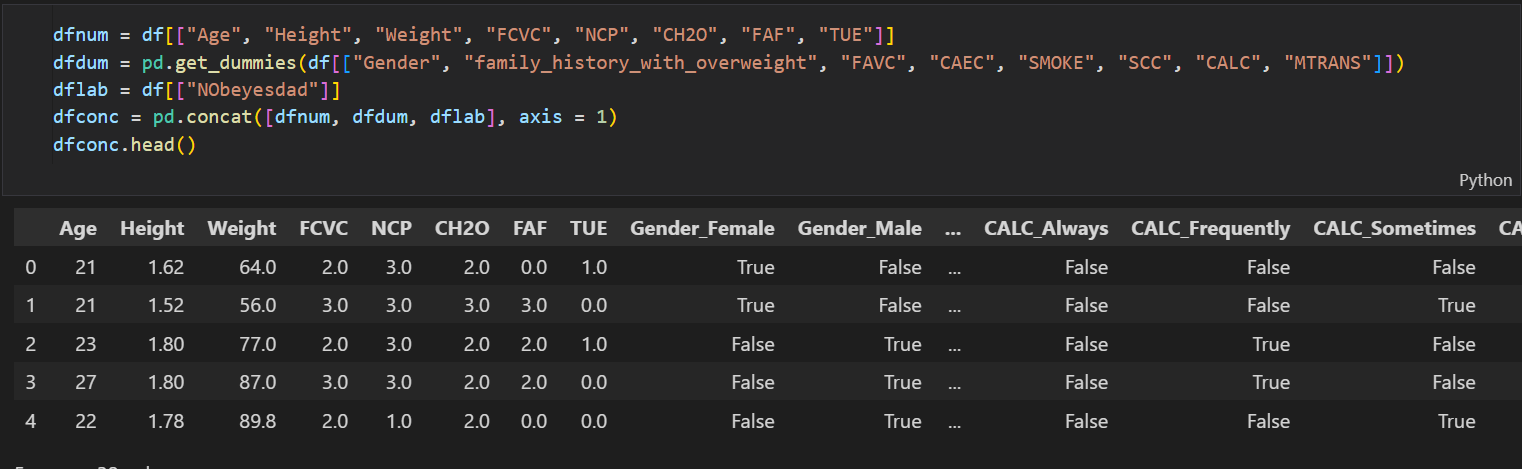


Рисунок 33. Вывод части датасета после замены категориальных параметров.

* Разделение на фичи и целевую переменную:

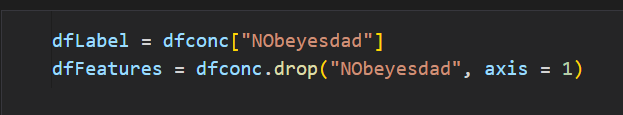


Рисунок 34. Код разделения датасета и целевой переменной

* Переводим все числовые данные в float64

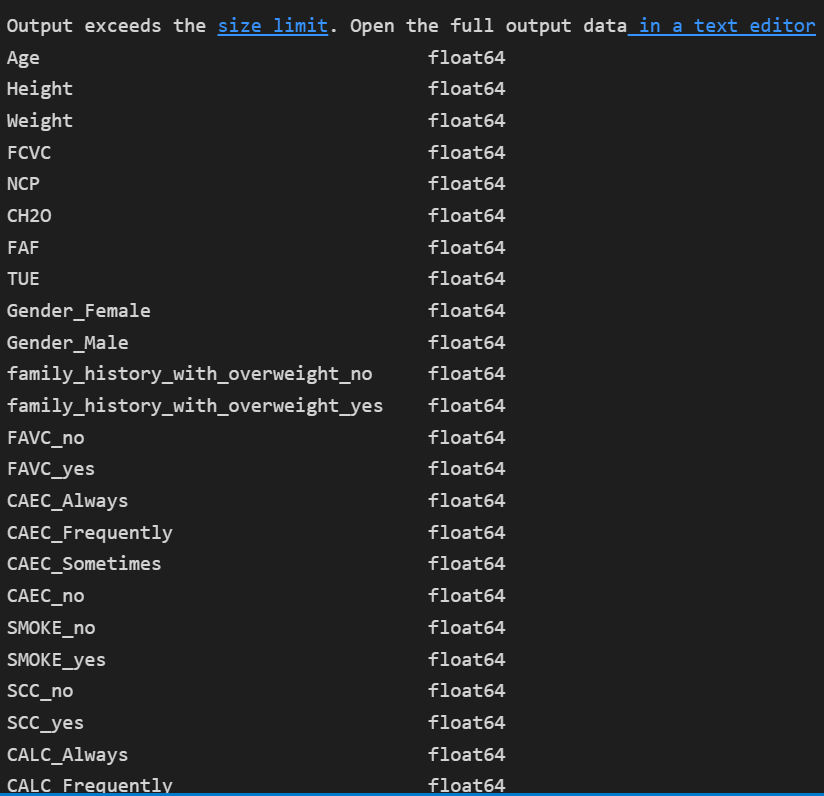


Рисунок 35. Информация о датасете после преобразования.

Производим нормализацию:

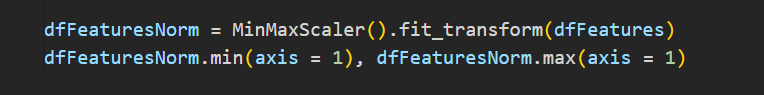


Рисунок 36. Код нормализации.

Кодировка целевой переменной:

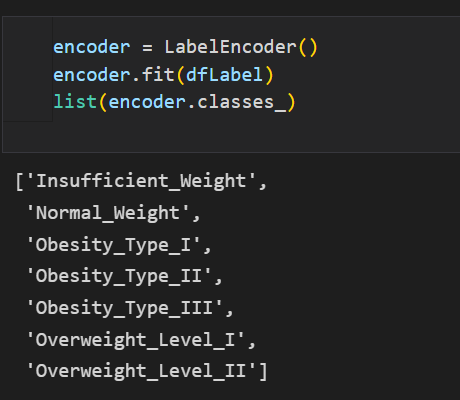


Рисунок 37. Код кодирования целевой переменной.

Для обучения модели датасет был разделен на Train и Test части в пропорции 80/20

1. Выбор модели (задача классификации)

Список испытываемых моделей:

* **LR (Logistic Regression) — Логистическая регрессия**  
  Модель для классификации, предсказывающая вероятность принадлежности к классу. Используется для бинарной и многоклассовой классификации.
* **LDA (Linear Discriminant Analysis) — Линейный дискриминантный анализ**  
  Метод для классификации и снижения размерности данных. Ищет линейные комбинации признаков, которые лучше всего разделяют классы.
* **KNN (K-Nearest Neighbors) — Метод k-ближайших соседей**  
  Модель классификации, которая относит объект к классу, наиболее часто встречающемуся среди его k ближайших соседей.
* **CART (Decision Tree Classifier) — Дерево решений**  
  Модель классификации, которая строит дерево решений, разделяя данные на основе правил, полученных из признаков.
* **NB (Gaussian Naive Bayes) — Наивный байесовский классификатор (гауссовский)**  
  Вероятностная модель, основанная на теореме Байеса. Предполагает независимость признаков и использует гауссовское распределение для непрерывных данных.
* **SVC (Support Vector Classifier) — Метод опорных векторов**  
  Модель классификации, который ищет гиперплоскость, максимально разделяющую классы. Параметр gamma='auto' автоматически настраивает влияние отдельных объектов на границу решения.

1. Оценка метрик разных моделей:

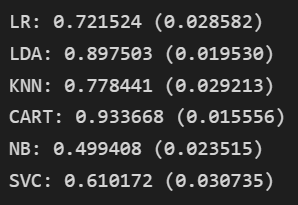


Рисунок 38. Оценка метрик разных моделей при обработке полного датасета.

Модель «Дерево решений»(CART) – показала наивысшую точность ~93%

Метрики модели «Дерево решений»

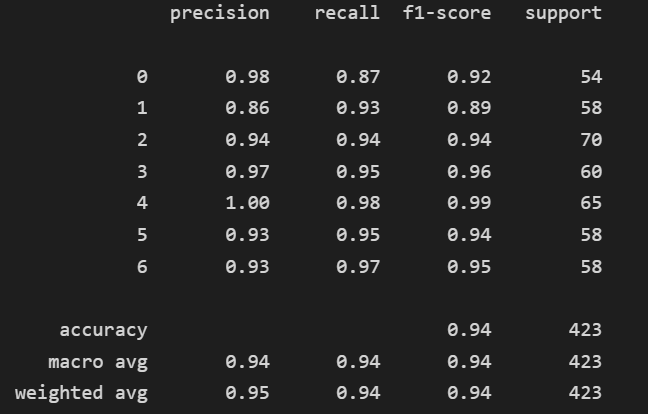


Рисунок 39. Оценка метрик модели «Дерево решений»

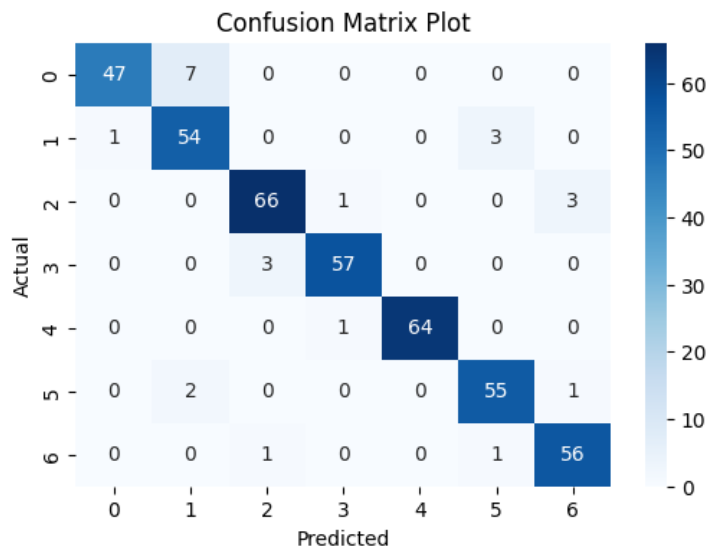


Рисунок 40. Результат обучения модели.

Для более удобного представления результатов – представим признаки (фичи) в порядке убывания их важности:

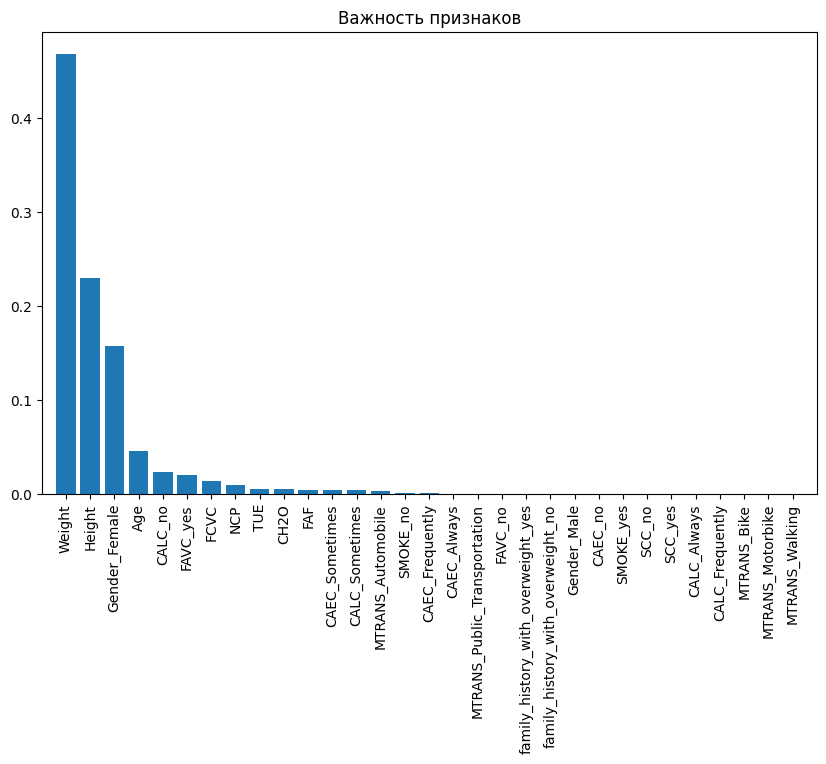


Рисунок 41. Фичи, расположенные по важности в обучении и предсказании.

1. Построение модели без явных признаков.

Так как параметры «Рост» и «Вес» напрямую влияют на степень ожирения человека и по ним можно рассчитать степень лишнего веса напрямую – попробуем построить модель без учета этих фич.

Иная предобработка датасета в данном случае не требуется. Просто повторяем такую же предобработку и обучаем те же модели на «новом» датасете – то есть без фич «Рост» и «Вес»:

Метрики разных моделей для обновленного датасета:

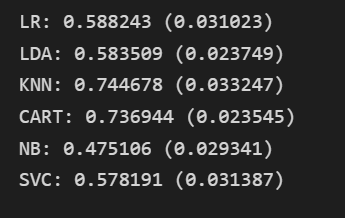


Рисунок 42. Оценка метрик разных моделей при обработке неполного датасета.

В данном случае модель «Дерево решений» также показала лучший результат, но при этом точность снизилась примерно на 20%.

Объяснить это можно тем, что факторы, напрямую влияющие на степень ожирения были самыми значимыми, а так как они отсутствуют в обновленном датасете – модели приходится ориентироваться лишь по косвенным факторам, которые могут быть не совсем точными.

Ниже приведены метрики для модели «Дерево решений»

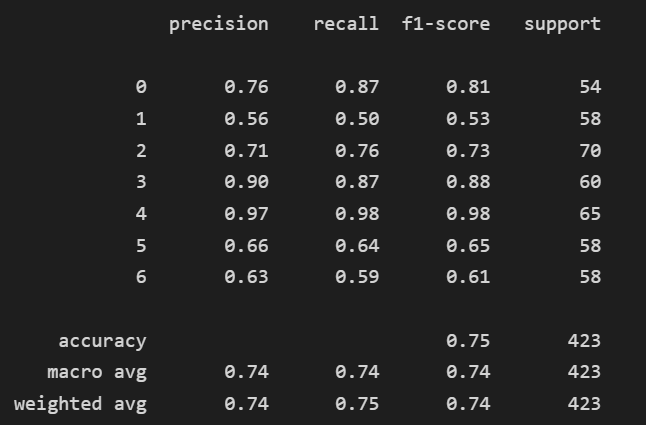


Рисунок 43. Оценка метрик модели «Дерево решений»

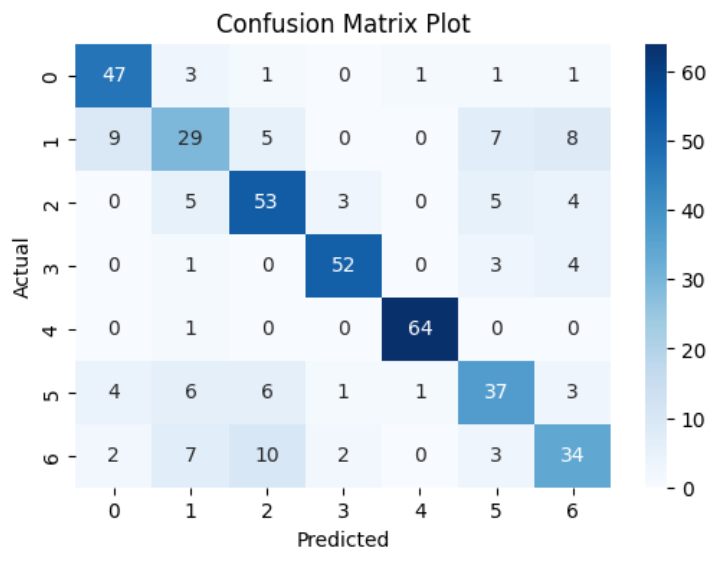


Рисунок 44. Результат обучения модели.

Также как и с прошлым датасетом, отобразим наиболее важные метрики:

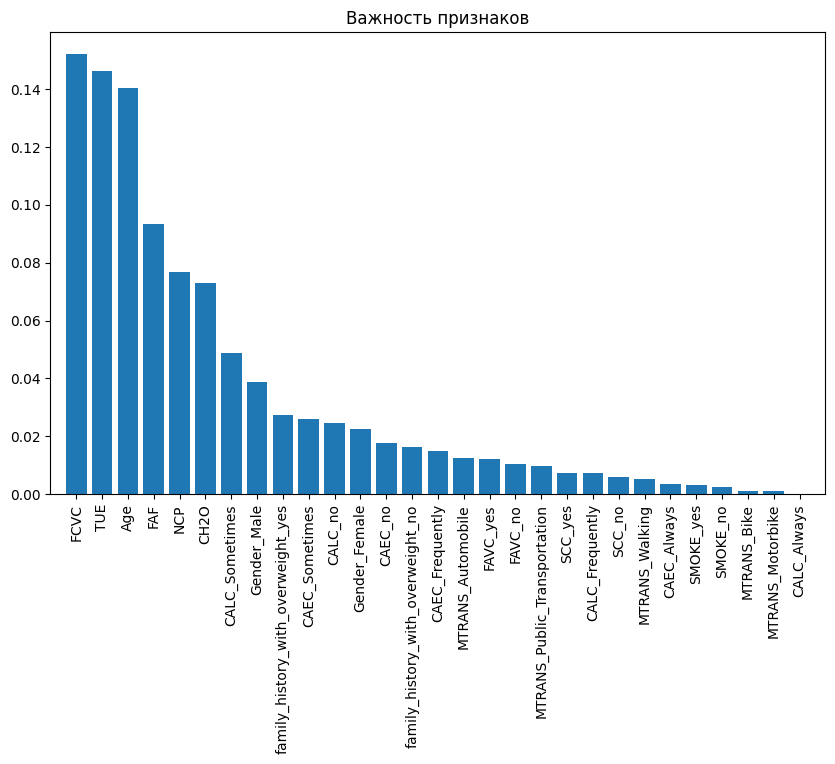


Рисунок 45. Фичи, расположенные по важности в обучении и предсказании.

1. Вывод:

По заданному датасету был произведён подробный разведочный анализ, выделены зависимости целевой переменной от фич. Далее было произведено кодирование категориальных переменных, нормализация и разделение на тренировочную и тестовую выборки. После с помощью кросс-валидационного анализа была выбрана подходящая модель классификации - Дерево Решений. Была построена модель, на которой удалось добиться точности по f1 оценке = 0.94, что является очень хорошим показателем.

После этого была построена визуализация значимости фич для задачи классификации. Самые важные параметры - вес и рост, что логично.

Было выдвинуто предположение, что лишний вес считается напрямую из показателя ИМТ, который зависит только от роста и веса человека. В таком случае – машинное обучение избыточно для данной задачи.

В следствие этого была поставлена следующая задача: научиться определять лишний вес или «Степень ожирения» без использования прямых факторов, которых достаточно для определения степени лишнего веса без машинного обучения. Таким образом, задача стала более прикладной – нам удалось определить косвенные признаки лишнего веса, либо признаки, которые могут приводить к лишнему весу с вероятностью около 75%, что позволяет получить более прикладную пользу. Например: избавиться от привычек, которые могут привести к лишнему весу ещё до того, как он появился.

Таким образом, поставленная цель была достигнута. Была выполнена задача классификации несколькими методами.

Для достижения цели были решены следующие задачи:

1. Выполнен анализ проблемы и обоснована ее актуальность.
2. Осуществлена предобработка данных
3. Осуществлено моделирование зависимости целевого признака от других с помощью различных классификаторов. Так же подобрана наилучшая модель, оценено ее качество и выполнено прогнозирование.
4. Создана и обучена модель модель, оценено её качество, а также выполнено прогнозирование.
5. Выбрана лучшая модель и сделаны выводы.
6. Были найдены как прямые признаки наличия проблемы, так и косвенные.

Построенная модель классификации «дерево решений» достаточно точно прогнозирует степень лишнего веса, и в случае с косвенными признаками – может даже предсказывать его появление.