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)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«13377. Архитектор данных (Data Architect Pro)»

Слушатель Новиков М.М.

Москва, 2024

**Содержание**

[**Введение** 3](#_Toc1)

[Аналитическая часть 5](#_Toc2)

[1. Постановка задачи 5](#_Toc3)

[2. Описание используемых методов 8](#_Toc4)

[3. Разведочный анализ данных 10](#_Toc5)

[Практическая часть 11](#_Toc6)

[1. Предобработка данных 11](#_Toc7)

[2. Разработка и обучение модели 16](#_Toc8)

[3. Тестирование модели 18](#_Toc9)

[4. Разработка приложения 22](#_Toc10)

[5. Создание удаленного репозитория и загрузка результатов 30](#_Toc11)

[**Заключение** 31](#_Toc12)

[**Список использованной литературы** 32](#_Toc13)

# **Введение**

Современный образ жизни и доступность широкого ассортимента продуктов питания привели к тому, что с каждым годом повышается внимание к здоровому питанию. Правильное питание может играть решающую роль в поддержании здоровья, профилактике заболеваний и обеспечении нормального функционирования организма. В то же время, избыточное потребление насыщенных жиров, сахаров и соли может привести к ожирению, сердечно-сосудистым заболеваниям и другим проблемам со здоровьем. В связи с этим, важным направлением становится анализ питательных свойств потребляемой пищи. Это позволит не только лучше понимать свойства продуктов, но и разрабатывать персональные рекомендации, способствующие улучшению здоровья населения.

Целью данной работы является построение архитектуры данных и разработка рекомендацией по здоровому питанию на основе анализа данных о составе различных продуктов питания. С помощью методов анализа данных и машинного обучения планируется выявить ключевые закономерности в свойствах продуктов и определить оптимальные комбинации для достижения сбалансированного рациона.

В работе основное внимание будет уделено таким макро- и микроэлементам, как белки, жиры, углеводы, клетчатка, сахара и натрий. Для сбора данных воспользуемся базой данных Web-сайта, на котором представлена информация о пищевой ценности различных продуктов FoodData Central. На основе полученных данных, планируется выявить необходимые зависимости, которые позволят создать рекомендации по здоровому питанию.

**Архитектура данных здорового питания** - это комплексный показатель учитывающий многие факторы. Такой образ жизни – это не запреты, а основа существования, уважения и любви к себе, своему телу, концепция жизнедеятельности, направленная на улучшение и сохранение здоровья, профилактики болезней.

**Хороший сон** критически важен для психического и физического здоровья.

**Физическая активность** нужна не только для снижения или поддержания веса. Занятия спортом, фитнесом или хотя бы регулярная физическая активность в виде зарядки и прогулок поддерживают наш организм, мышцы и суставы в «рабочем» состоянии, снижают риск развития хронических заболеваний и в целом продлевают жизнь.

**Гигиена** включает в себя личную гигиену, уход за телом, волосами и зубами, вещами и обувью, ношение одежды с натуральным составом, закаливание, уборку жилища.

Известно, что кожа человека способна защищать весь организм от негативных воздействий окружающей среды. Кроме защитной функции она выполняет еще и терморегулирующую, иммунную, дыхательную и иные функции.

**Безопасная окружающая среда** оказывает существенное влияние на качество питания. Нужно обращать внимание на материалы для строительства и ремонта, мебель, бытовую химию, посуду – это всё влияет на наше здоровье. Эксперты утверждают, что лишь 25% строительных и отделочных материалов являются экологически безопасными. Источником опасности может стать бетон, штукатурка, шпатлевка, ковролин, утеплитель, клеи, различные краски и лаки. Все эти фенолы, толуолы, бензолы, ксилолы в составе строительных материалов не самые полезные вещества для нервной системы и репродуктивных органов. Моющие средства с ядовитыми компонентами типа хлора и продуктов нефтепереработки уже запрещены в большинстве стран мира. Средства используемые при изготовлении посуды, вредны для здоровья людей и животных. Здоровая и безопасная посуда не должна содержать вредных химических веществ.

**Психоэмоциональное самочувствие** также влияет на качество жизни. Во время стресса происходит абсолютный гормональный взрыв: выброс адреналина и кортизола, в ответ снижаются «полезные» функции организма. Существует прямая связь между эмоциональным состоянием и такими недугами, как сахарный диабет, дисфункция репродуктивной системы, сердечно-сосудистые заболевания, различные зависимости. Стресс – причина серьезной проблемы – хронической усталости.

**Правильное (рациональное) питание** – это особо важный элемент жизни.

Оно чрезвычайно важно для здоровья! Питание должно быть полноценным и сбалансированным, нужно исключить еду «быстрого приготовления» с пищевыми добавками, усилителями вкуса и подсластителями.

Сбалансированное питание человека – это продуманный рацион, который предполагает поступление в организм белков, жиров, углеводов, жирных кислот, аминокислот и витаминов в правильном соотношении с учетом индивидуальных потребностей.

Питание должно быть также энергетически сбалансировано, организм должен не только расходовать, но и потреблять питательные вещества, в противном случае, возможно истощение и нарушение внутренних процессов и высок риск серьезных проблем со здоровьем.

**Важно не только количество пищи, но и её качество**!

# Аналитическая часть

1. Постановка задачи

Целью данной ВКР является построение архитектуры данных здорового питания на основе анализа данных о составе продуктов питания с применением моделей машинного обучения и методов классификации. Для достижения поставленной цели необходимо разработать модель, которая будет не только анализировать свойства продуктов, но и предлагать персональные рекомендации. Использование алгоритмов классификации и регрессии позволит создать точные рекомендации по выбору продуктов питания для улучшения здоровья.

В рамках выполнения работы планируется последовательно реализовать следующие шаги:

1. Провести сбор данных. Для этого планируется использовать базу данных сайта FoodData Central, которая содержит информацию о пищевой ценности различных продуктов. Процесс сбора включает несколько этапов:

* Получение данных через API, который предоставляется сайтом: <https://fdc.nal.usda.gov/api-guide.html.>
* Проведение первоначальной фильтрации данных с целью выделения необходимых значений, которые будут полезны в работе и подготовке модели.

1. Обработка полученных данных. Данный шаг играет очень важную роль и включает следующие этапы:

* Удаление или корректировка пропусков и аномальных значений для обеспечения чистоты данных.
* Приведение данных к единой шкале для обеспечения корректного сравнения между различными параметрами.
* Удаление лишних данных, которые могут нести шум или не имеют значительного влияния на результаты анализа.

1. Применение методов машинного обучения. На данном шаге будут применены методы для анализа данных и выявлены значимые закономерности с помощью методов машинного обучения. Процесс включает в себя следующие этапы:

* Определение подходящих методов, которые лучше всего подойдут для решения поставленной задачи.
* Обучение модели на подготовленных данных.
* Оценка производительности модели с использованием тестового набора данных и метрики выявления точности.

1. Сохранение полученной модели.
2. Применение полученной модели. На данном шаге полученная модель будет включена в приложение, которое даёт оценку о продуктах питания.

Исходные данные, которые получены с сайта (https://fdc.nal.usda.gov/api-guide.html) представлены на Рис. 1 и сохранены в формате CSV.

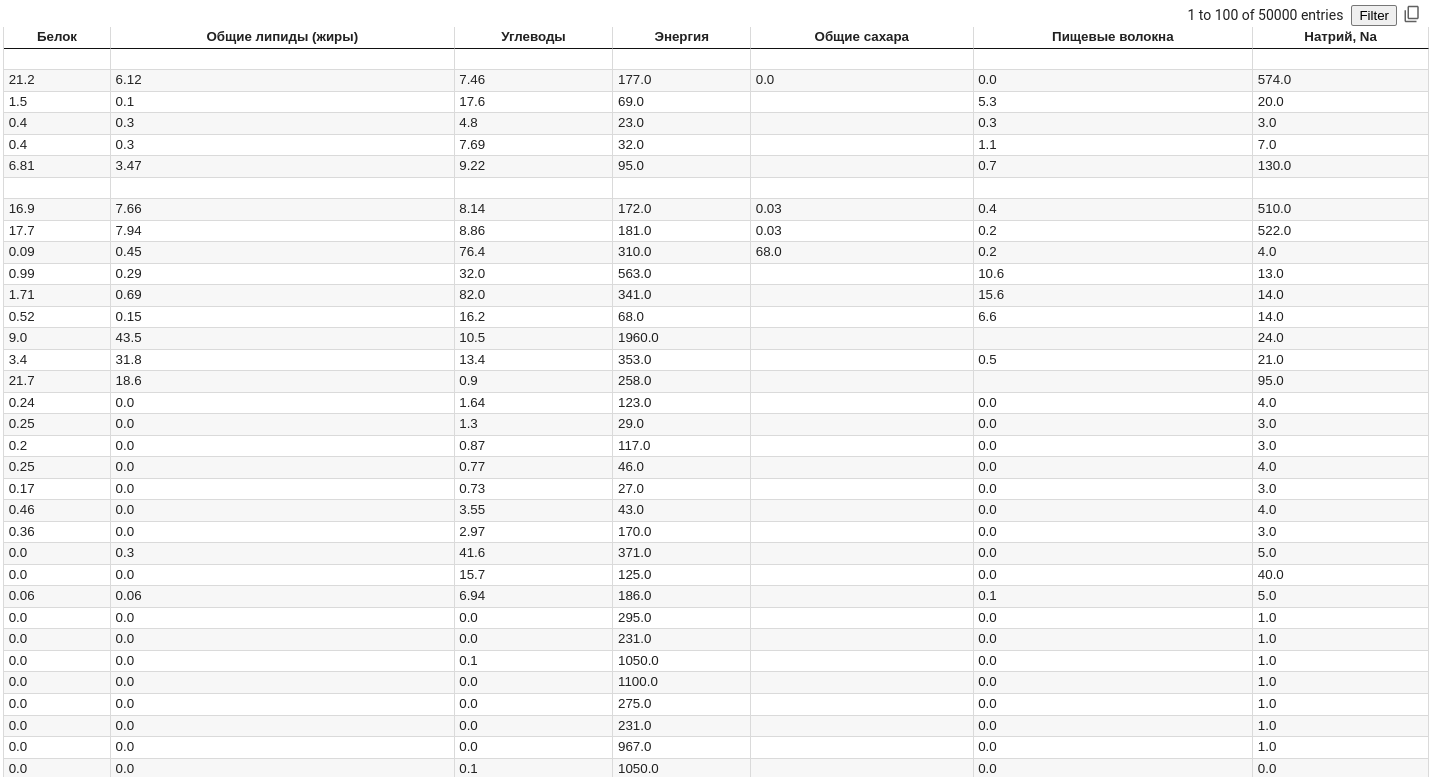


Рис. 1. Исходные данные с сайта FoodData Central.

1. Описание используемых методов

Для решения поставленной задачи предполагается использовать следующие методы и подходы:

1. Собрать и обработать необходимые данные. Для получения данных воспользуемся библиотекой «aiohttp». Причины выбора данной библиотеки следующие: необходим сбор большого количества данных с нескольких страниц и скорость получения данных. После получения данных предварительная обработка и итоговый результат загружается в файле формата CSV.
2. Анализ распределения данных. Для анализа данных используются гистограммы и распределения (KDE) для визуализации ключевых питательных веществ, таких как углеводы, белок и клетчатка, с целью выявления закономерностей и особенностей распределения данных.

* KDE – процесс нахождения оценочной функции, которая пытается вывести характеристики генеральной совокупности на основе набора данных. Кривая рассчитывается путем взвешивания расстояния между всеми точками в каждом конкретном месте распределения. Если есть больше точек, которые будут сгруппированы локально, то оценка будет выше, так как вероятность увидеть точку в этом месте увеличивается.

1. Выявление и обработка выбросов. Для анализа выбросов используется метод интерквартильного размаха (IQR) и анализ выбросов. Метод IQR позволяет обнаружить и удалить аномальные значения, которые помогли улучшить качество модели и сделать ее более устойчивой к экстремальным данным.

* Метод интерквартильного размаха. С помощью верхней и нижней квартилей определяется важная мера рассеивания набора чисел – интеквартильный размах. Данный метод представляет собой разность между Q3 и Q1: IQR = Q3 - Q1. Интерквартильный размах показывает разброс в пределах 50% от «центральных» значений рассматриваемого набора чисел.

1. Нормализация данных. Для приведения к единой шкале данных используется нормализация, которая помогает упростить обучение модели и улучшить ее производительность.
2. Создание и обучение модели. Для классификации данных используется многослойный персептрон (MLP). Данная модель строится с использованием библиотек «TensorFlow» и «Keras», которая состоит из трех слоев: входного, скрытого и выходного.

* Многослойный персептрон – это класс искусственных нейронных сетей прямого расположения, который состоит как минимум из трех слоев: входного, скрытого и выходного. За исключение входных, все нейроны используют нелинейную функцию активации. При обучении MLP используется обучение с учителем и алгоритм обратного распространения ошибки.
* Алгоритм обратного распространения ошибки является популярным алгоритмом обучения плоскослоистых нейронных сетей прямого назначения. В основе идеи алгоритма лежит использование выходной ошибки нейронной сети для вычисления величин коррекции весов нейронов в ее скрытых слоях.

1. Оценка и валидация модели. Для оценки производительности модели были применены метрики точности и матрица ошибок. Данные метрики помогают определить насколько хорошо модель выполняет поставленную задачу и выявить в ней слабые места.

1. Разведочный анализ данных

Как уже было указано в п. 1.1 исходный датасет получается из базы данных Web-сайта FoodData Central, который содержит большое количество ошибочных данных, выявленных при первоначальном разведочном анализе.

В ходе предварительной обработки все пропуски были уже заполнены нулями. Данное решения принято с связи с тем, что оно могло упростить анализ и дальнейшую обработку данных, так как пропуски свидетельствуют об отсутствии информации о составе продукта.

Для исключения из анализа данных, которые могли сыграть отрицательную роль в обучении модели, был проделан шаг с фильтрацией: удаление строк, которые содержат большое количество показателей с нулевыми значениями.

В ходе визуализации распределения ключевых питательных веществ, были выявлены аномалии в данных и основное направление в решении поставленной задачи. Выявление выбросов данных с использованием метода интерквартильного размера позволило очистить данные от экстремальных значений. В последствии, данный подход повысил точность модели.

Для обеспечения равномерного распределения значений и улучшения качества модели, был предпринят шаг с нормализацией, то есть приведение к единой шкале и уменьшение влияния отдельных компонентов на модель.

Этап разведочного анализа позволил выявить важные закономерности и очистить данные от шумов. На данном этапе были подготовлены данные, с которыми можно проводить дальнейший анализ и строить модель машинного обучения.

# Практическая часть

1. Предобработка данных

Предобработка данных для последующего их анализа проведена в несколько шагов:

1. Получение ключа к API сайта FoodData Central и создание функций на Python для загрузки данных, получения количества страниц, обработка загруженных данных (Рис. 2 и Рис. 3).

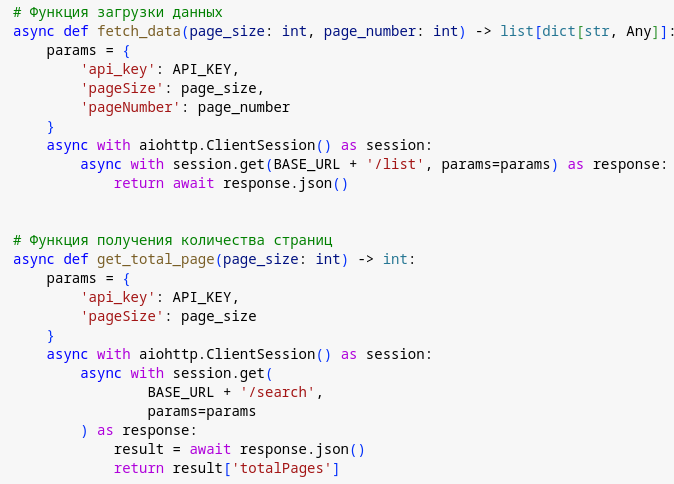


Рис. 2. Функции загрузки данных и получения количества страниц.

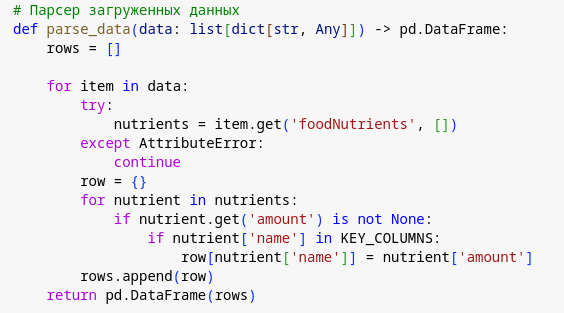


Рис. 3. Функция с обработкой данных.

1. Полученные данные были объединены в единые DataFrame с последующим переименованием столбцов на русский язык для удобства анализа. После объединения данные были сохранены в файл формата CSV.
2. Полученные данные необходимо было обработать. Первым этапом производим заполнение пропусков нулями. Данный этап необходим для того, чтоб избежать проблемы при анализе и построении модели, а также для упрощения обработки данных. Пропуски, которых в исходном датасете оказалось много, могут свидетельствовать о том, что полная информация по некоторым продуктам питания отсутствует или отсутствуют некоторые питательные компоненты. На втором этапе произведен анализ количества строк, которые содержат нулевые значения. Данные строки были удалены, чтоб исключить из анализа продукты с недостаточным количеством информации о питательных веществах.



Рис. 4. Количество пропусков по компонентам.

Графики полученные после обработки представлены на рисунках: 5, 6, 7.

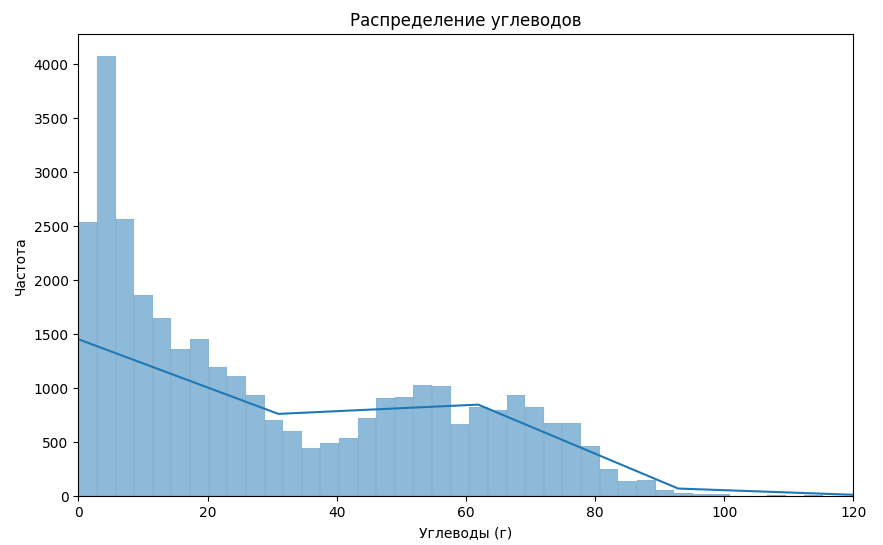


Рис. 5. Распределение углеводов.

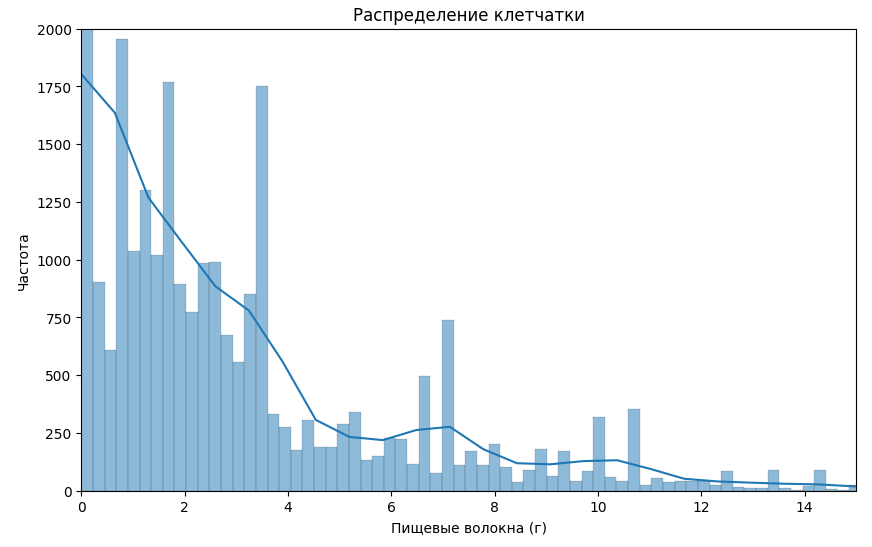


Рис. 6. Распределение клетчатки.

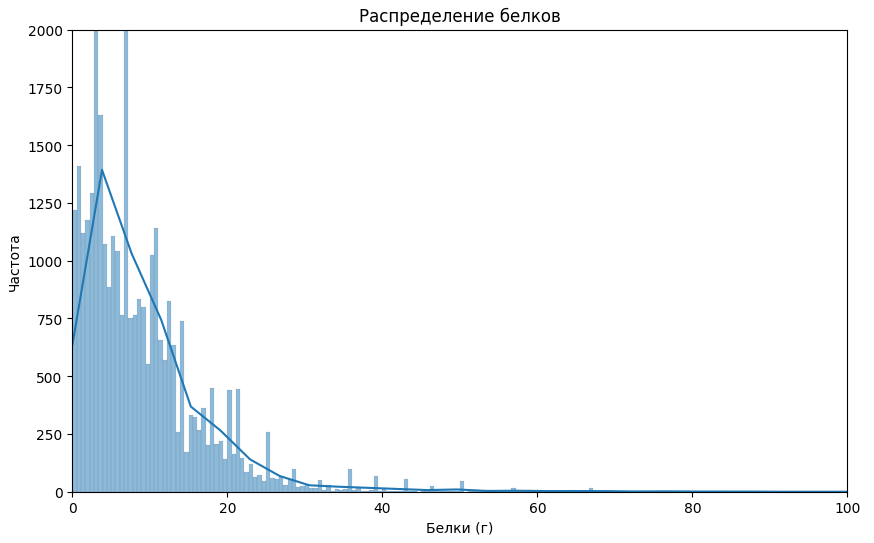


Рис. 7. Распределение белков.

1. Для приведения значений к единой шкале была проведена нормализация. В рамках данной работы была применена функция нормализации, которая преобразует значения каждого показателя в диапазон от 0 до 10. Данный подход обеспечивает равномерное распределение данных и снижает влияние крайних значений.
2. Последним заключительным шагом было применение метода интерквартильного размаха. Выбросы, которые могли содержаться в данных, имели высокий шанс повлиять на результат анализа и качество модели. После данного шага, показатели значений, выходящие за пределы допустимого диапазона, были удалены.
3. Разработка и обучение модели

Перед началом разработки модели был выполнен процесс формирования признаков и целевой переменной. В качестве признаков были выбраны ключевые вещества: белки, жиры, углеводы, клетчатка, сахара, пищевые волокна и натрий. Целевая переменная («оценка здоровья») была рассчитана на основе весовой формулы, которая учитывает вклад каждого из компонентов в общий рейтинг продукта.

Для обеспечения точности и надежности модели, данные были разделены на три части:

1. Обучающая выборка (60% от общего объема) – использовалась для обучения модели.
2. Валидационная выборка (20%) – применялась для настройки гиперпараметров и выбора лучшей модели.
3. Тестовая выборка (20%) – использовалась для окончательной оценки точности модели.

Для разработки модели был выбран многослойный персептрон (MLP). Архитектура модели включает следующие слои:

* Входной слой – принимает данные, размерность которого соответствует числу признаков.
* Два скрытых слоя (64 и 32 нейрона) – используются для моделирования нелинейных зависимостями между признаками.
* Слои Dropout (с долей 50%) – для предотвращения переобучения модели.
* Выходной слой с одним нейроном – для получения вероятности принадлежности к признаку «здоровый» или «нездоровый».

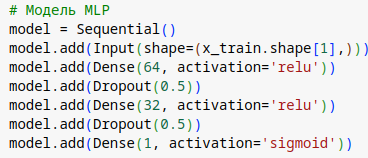


Рис. 8. Модель MLP.

После определения архитектуры модель была скомпилирована с использованием оптимизатора «Adam» и функции потерь «binary\_crossentropy». Обучение модели проводилось на 50 эпохах с размером батча 32. В процессе обучения проводилось отслеживание потерь и точности на обучающей и валидационной выборках.

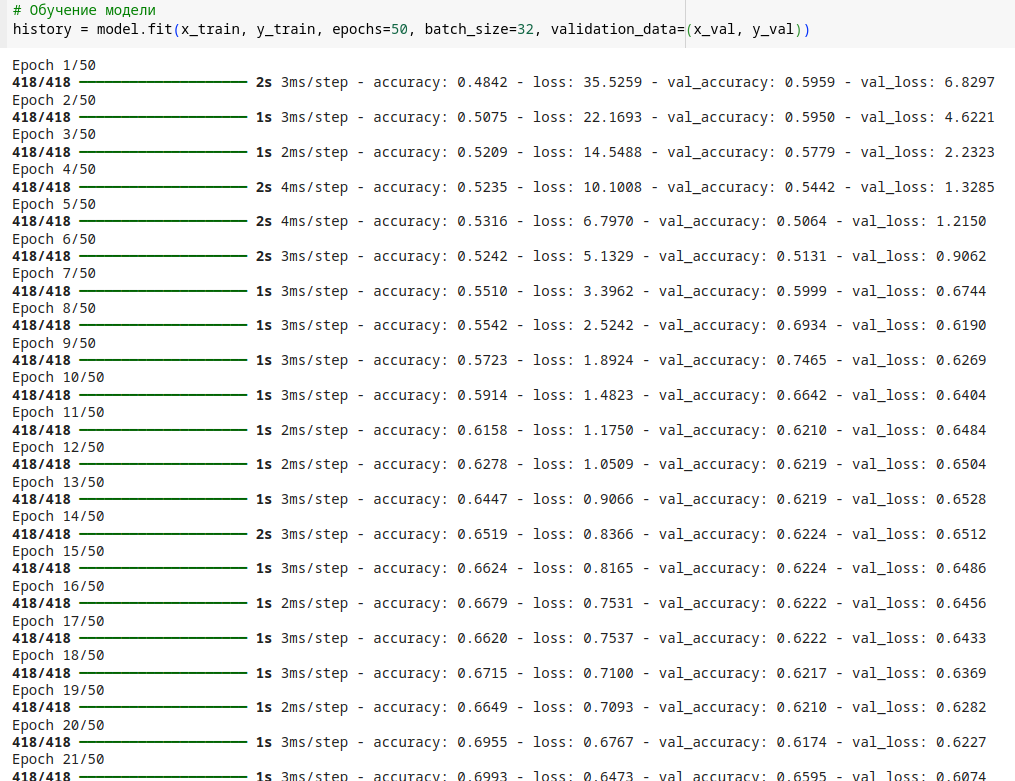


Рис. 9. Обучение модели.

1. Тестирование модели

На этапе тестирования оценивалась эффективность и точность разработанной модели машинного обучения.

После завершения процесса обучения модель была протестирована на отложенной тестовой выборке, которая составляла 20% от всего набора данных. Использование данной выборки позволило объективно оценить, насколько хорошо модель может обобщать информацию и делать предсказания на новых данных.

Основной метрикой для оценки точности модели был выбран показатель «accuracy», который отражает долю правильно классифицированных примеров из общего числа.

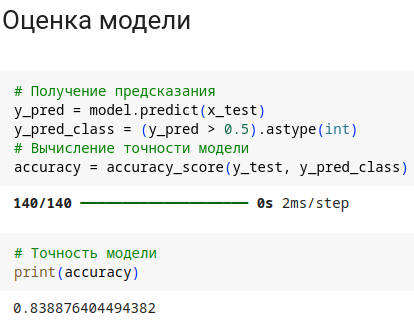


Рис. 10. Оценка точности модели.

Результаты тестирования также были визуализированы для наглядного представления работы модели. Были построены графики, которые отражают работу модели на тестовых данных, а также матрица ошибок, которая позволила оценить распределение ошибок между классами.

Для более детальной оценки работы модели была построена матрица ошибок. Матрица ошибок позволила проанализировать распределение правильных и неправильных предсказаний между двумя классами:

* True Positive (TP) – количество продуктов, которые модель правильно классифицировала как «здоровые».
* True Negative (TN) – количество продуктов, которые модель правильно классифицировала как «нездоровые».
* False Positive (FP) – количество продуктов, которые модель ошибочно классифицировала как «здоровые».
* False Negative (FN) – количество продуктов, которые модель ошибочно классифицировала как «нездоровые».

Анализ матрицы ошибок позволил выявить слабые стороны модели, такие как склонность к ошибкам в пользу одного из классов, и помог в дальнейшем улучшить модель.

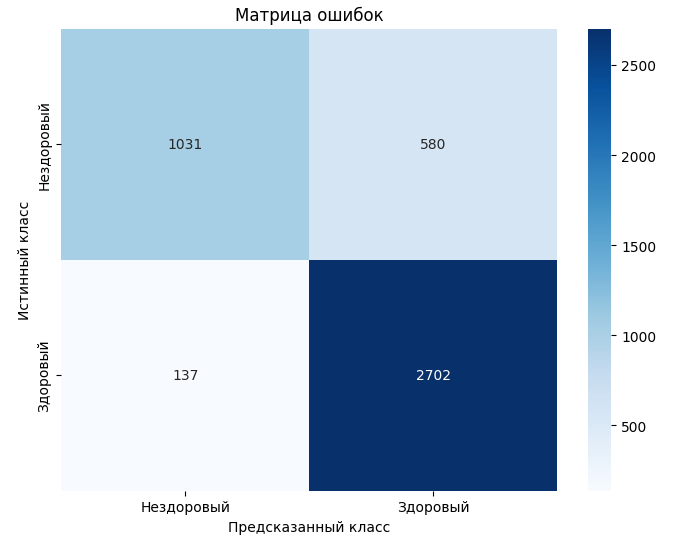


Рис. 11. Матрица ошибок.

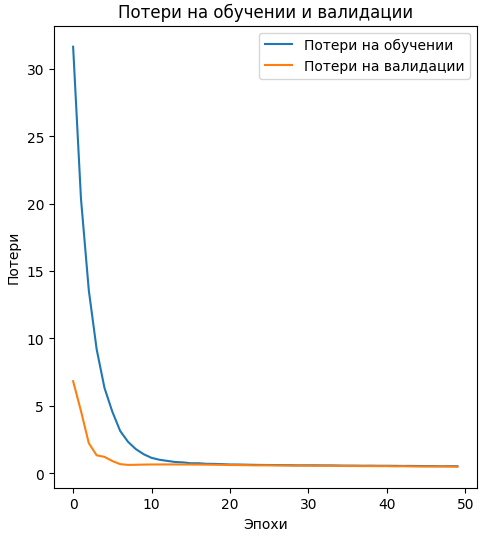


Рис. 12. Потери на обучении и валидации.

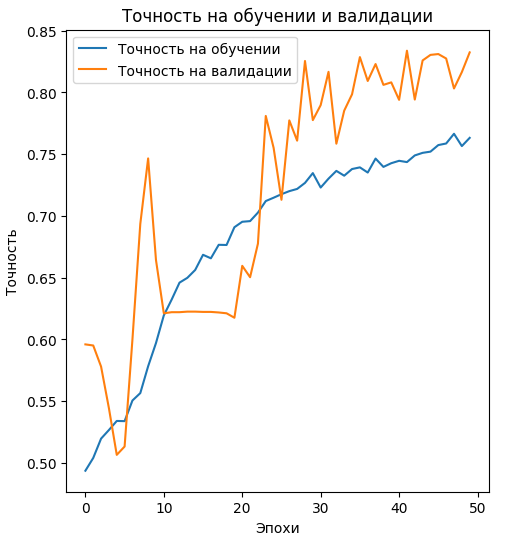


Рис. 13. Точность на обучении и валидации.

Последний график (Рис. 13) может свидетельствовать о том, что модель в промежутке эпох 10-20 могла быть переобучена, а также в промежутке эпох 25-35.

1. Разработка приложения

Для создания приложения, которое могло бы выдать оценку продукта, будет использоваться библиотека «Flask». Основной функционал данного приложения:

* Загрузка данных о продуктах.
* Взаимодействие с моделью машинного обучения.
* Сохранение полученных результатов в базе данных с целью уменьшения обращения к модели.

Для анализа данных о продуктах используется обученная модель, которая реализована с помощью «TensorFlow» и «Keras». Модель загружается в приложение и применяется для прогнозирования оценки продукта питания.

Приложение использует базу данных SQLite для хранения информации о продуктах и их оценках. Реализован функционал, который позволяет проверить наличие записи о продукте в базе данных, что даёт возможность избежать повторного вычисления оценки для уже проверенных продуктов.

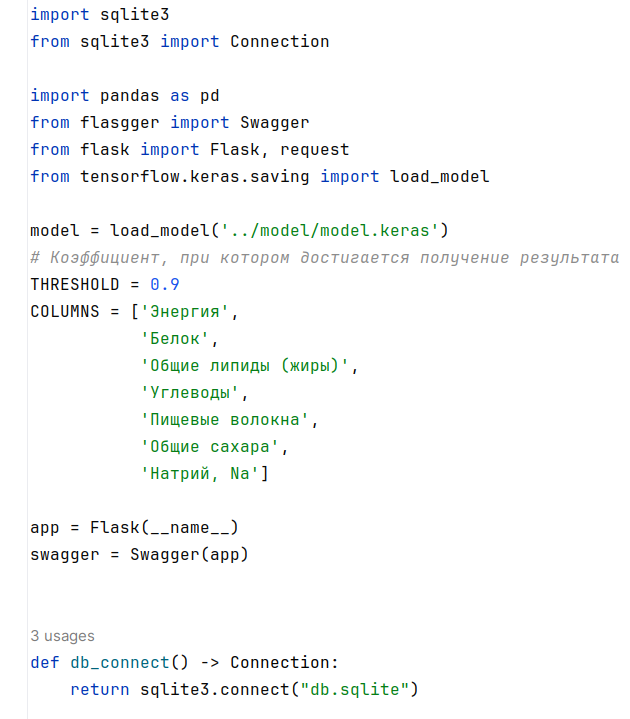


Рис. 14. Код разработки «функция подключения к базе данных».

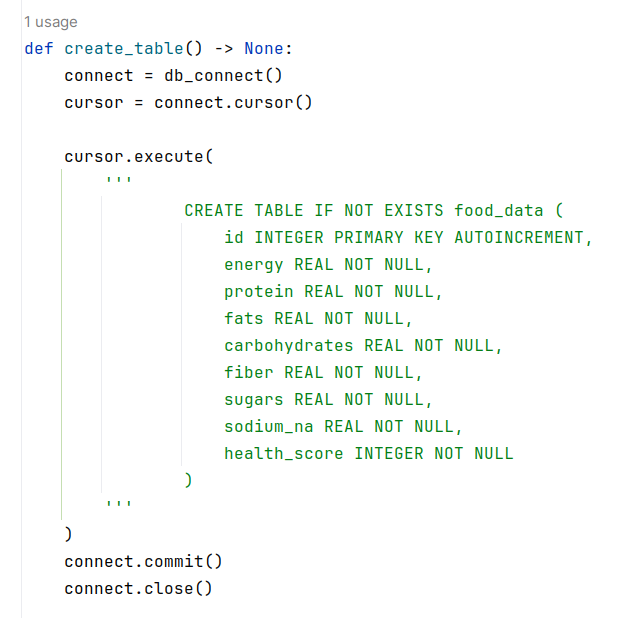


Рис. 15. Код разработки «создание таблицы в базе данных».

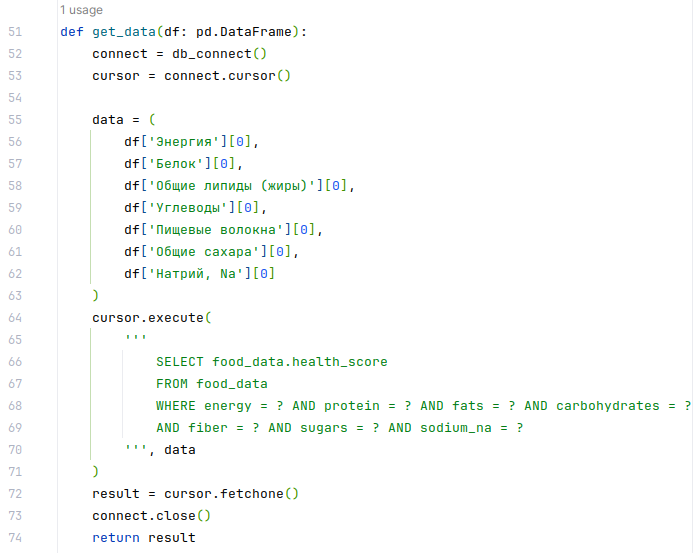


Рис. 16. Код разработки «функция для получения параметра оценки из базы данных».



Рис. 17. Код разработки «функция для получения параметра оценки с модели данных и сохранение результата в базе данных».

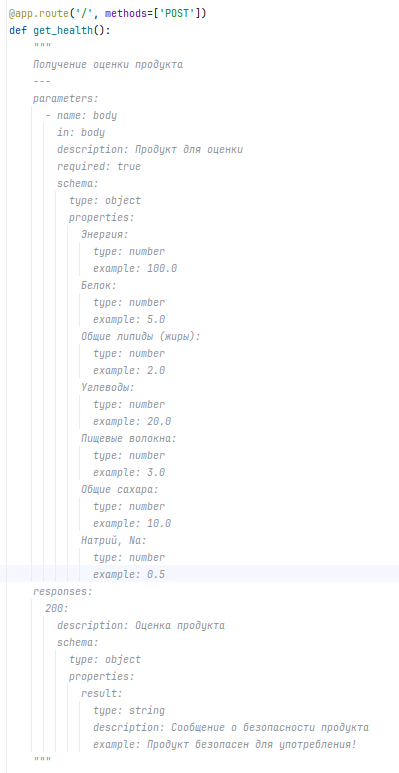


Рис. 18. Код разработки «добавление описания для Swagger-UI».

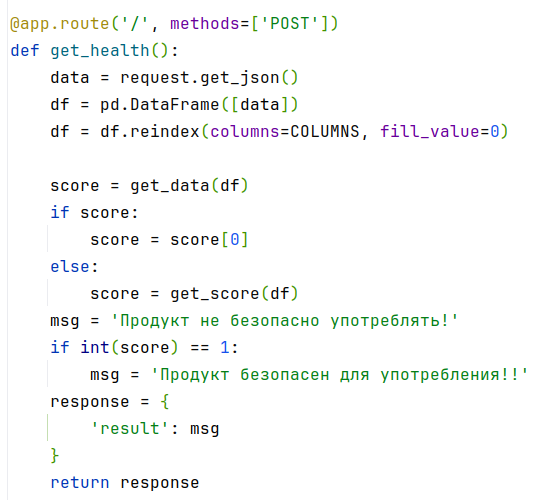


Рис. 19. Код разработки «реализация функции выдачи данных».

Приложение предоставляет API, с помощью которого можно передать данные о продукте и получить оценку его безопасности. Пользователь отправляет данные о продукте в формате JSON через POST-запрос. Если продукт уже существует в базе данных, приложение возвращает ранее рассчитанную оценку, если данные о продукте отсутствуют, модель машинного обучения анализирует его состав, вычисляет оценку и сохраняет результат в базе данных. Для удобства взаимодействия с приложением используется библиотека «Swagger-UI», которая после запуска расположена по адресу: http://127.0.0.1:5000/apidocs/.

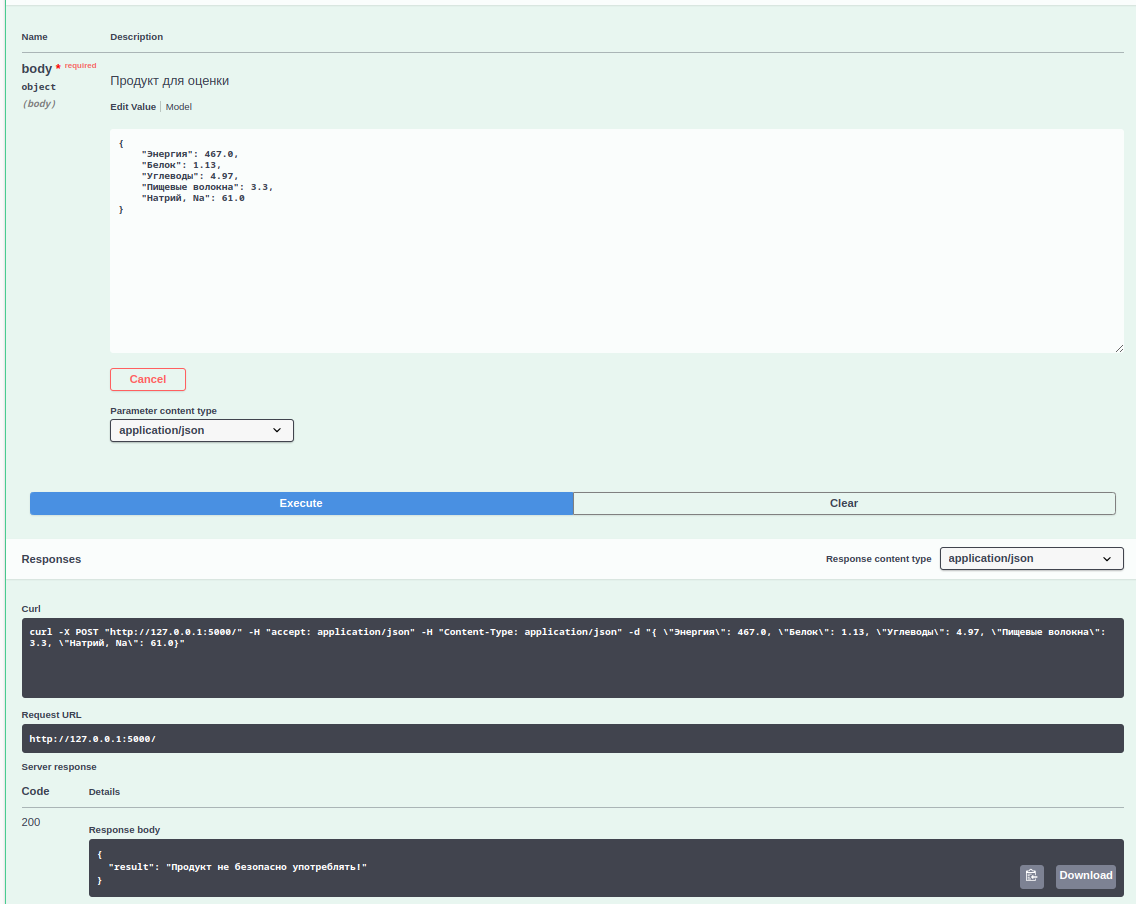


Рис. 20. Итоговый интерфейс для взаимодействия с приложением.

1. Создание удаленного репозитория и загрузка результатов

Код исследования размещен в созданном на GitHub репозитории:

https://github.com/Mikhail8881/13377\_Data\_Architect-\_Pro

# **Заключение**

В ходе выполнения работы была построена архитектура данных здорового питания и протестированы модели машинного обучения. Разработанное приложение позволяет автоматизировать процесс оценки и классификации продуктов питания.

Здоровое питание имеет важнейшее значение для обеспечения выполнения «Доктрины продовольственной безопасности Российской Федерации», согласно Указа Президента Российской Федерации В.В. Путина от 21.01.2020 г. № 20.

Настоящая Доктрина является документом стратегического планирования, в котором отражены официальные взгляды на цели, задачи и основные направления государственной социально-экономической политики в области обеспечения продовольственной безопасности Российской Федерации.

# **Список использованной литературы**

1. Библиотека pandas [Электронный ресурс] : – Режим доступа: https://pandas.pydata.org. (дата обращения: 26.08.2024).
2. Библиотека scikit-learn [Электронный ресурс] : – Режим доступа: https://scikit-learn.org. (дата обращения: 26.08.2024).
3. Industrial-Strength Natural Language Processing. SpaCy [Электронный ресурс] : – Режим доступа: https://spacy.io. (дата обращения: 25.08.2024).
4. Saishruthi Swaminathan Logistic Regression – Detailed Overview [Электронный ресурс] : – Режим доступа: https://medium.com/towards-data-science/logistic-regression-detailed-overview-46c4da4303bc. (дата обращения: 25.08.2024).
5. Пользователь @wunder\_editor Глубокое обучение для новичков: распознаем рукописные цифры [Электронный ресурс] : – Режим доступа: https://habr.com/ru/companies/wunderfund/articles/314242/. (дата обращения: 27.08.2024).
6. Пользователь Александр Гончар @Rachnog Прогнозирование финансовых временных рядов с MLP в Keras [Электронный ресурс] : – Режим доступа: https://habr.com/ru/articles/327022/. (дата обращения: 27.08.2024).