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**

«Data Science»

Слушатель Морозов Дмитрий Викторович

Москва, 2024

# Содержание

[Содержание 2](#_Toc106375636)

[Введение 3](#_Toc106375637)

[1 Аналитическая часть 4](#_Toc106375638)

[1.1 Постановка задачи 4](#_Toc106375639)

[1.2 Описание используемых методов 6](#_Toc106375640)

[1.3 Разведочный анализ данных 6](#_Toc106375641)

[2 Практическая часть 17](#_Toc106375642)

[2.1 Предобработка данных 17](#_Toc106375643)

[2.2 Разработка и обучение модели 18](#_Toc106375644)

[2.3 Тестирование модели 18](#_Toc106375645)

[2.4 Нейронная сеть 19](#_Toc106375646)

[2.5 Разработка приложения 20](#_Toc106375647)

[3 Создание удаленного репозитория 21](#_Toc106375648)

[Заключение 22](#_Toc106375649)

[Библиографический список 23](#_Toc106375650)

# Введение

Данная работа выполнена в рамках курса Data Science.

В качестве анализируемой задачи принята тема «Прогнозирование безопасной навигации судоходных рек Красноярского края при осуществлении "северного завоза"»

# 1 Аналитическая часть

## 1.1 Постановка задачи

Тема: Прогнозирование безопасной навигации судоходных рек Красноярского края при осуществлении "северного завоза"

По территории Красноярского края протекает огромное количество рек. Однако многие из них являются судоходными лишь в короткий период половодья. Точно предсказать срок навигации на таких реках – важнейшая задача.

Ежегодно в Красноярском крае происходит «северный завоз» - это комплекс мероприятий по доставке речным транспортом необходимых запасов, оборудования и материалов в населенные пункты, до которых можно добраться только по рекам. К таким поселениям относятся, например, поселок городского типа Тура на реке Нижняя Тунгуска или поселок Ванавара на реке Подкаменная Тунгуска. Помимо обеспечения населения, также доставляются грузы в места разработки полезных ископаемых, например, на Ванкорское нефтяное месторождение в бассейне реки Большая Хета.

Каждый год время начала и окончания навигации смещается в зависимости от фактических метеоусловий (температура и осадки в период таянья снега), запасов снега в бассейне рек. В ожидании достаточного уровня воды в устьях рек собираются караваны судов. Судам необходимо не только подняться вверх по течению до пункта назначения, но и вернуться обратно до того, как уровень воды упадет до критической отметки. Нередки случаи, когда суда оказывались на мели до следующего сезона «большой воды» или получали повреждения из-за низкого уровня воды. При этом уровень воды на некоторых реках может меняться на десятки метров всего за несколько дней.

Навигация затрудняется не только меняющимся уровнем воды, но и сложным рельефом русел. Так, например, на реке Нижняя Тунгуска при подъеме воды в Большом пороге выше отметки 30 метров, порог считается непреодолимым. И суда стоят в ожидании падения уровня воды, а затем буксируются вверх по порогу по очереди. Это сильно сказывается на сроках пути.

Для прогнозирования уровня рек предлагается использовать собранные за 2008-2017 года ежедневные наблюдения по постам гидрологического контроля рек Подкаменная Тунгуска (12 гидрологических постов). В файлах с данными представлена информация и легенда с каждого из гидрологических постов. Информация за каждый год находится в отдельном файле. Также можно получать фактические данные с <https://gmvo.skniivh.ru/> после регистрации.

Для получения суточных метеоданных (температура воздуха, осадки и др.) за необходимые даты и с нужных (по мнению участников конкурса) метеопостов предлагается использовать <http://aisori-m.meteo.ru/waisori/index.xhtml?idata=5> [Булыгина О.Н., Разуваев В.Н., Александрова Т.М.«ОПИСАНИЕ МАССИВА ДАННЫХ СУТОЧНОЙ ТЕМПЕРАТУРЫ ВОЗДУХА И КОЛИЧЕСТВА ОСАДКОВ НА МЕТЕОРОЛОГИЧЕСКИХ СТАНЦИЯХ РОССИИ И БЫВШЕГО СССР (TTTR)»].

Для конвертации данных из формата exlx в формат cvs была разработана утилита на языке python (src\utils). Утилита способна конвертировать размеченные таблицы с данными из источников <https://gmvo.skniivh.ru/> и <http://aisori-m.meteo.ru> в форматы cvs , а так же в posgresql. Сложность вызывает периодическое изменение формата таблиц из источника <https://gmvo.skniivh.ru/>

В качестве входных данных после обработки приняты следующие:

* дата;
* Высота снежного покрова;
* Степень покрытия окрестности станции снегом;
* Температура за 14 предыдущих дней;
* Текущая температура;
* Прогноз на 1 день вперед;
* Уровень воды за предыдущие 7 дней;
* Текущий уровень воды на посте 9387 (Ванавара);

Общее количество параметров для анализа – 27.

Датасеты были объединены по дате.

В итоговом датасете, пропуски отсутствуют. Элементы массива соответствуют типу float64.

**Критерии решения:**

1. Разработать алгоритм и программное решение для прогноза календарных дат начала и окончания навигации для реки Подкаменная Тунгуска. Достаточным для навигации уровнем считать уровень реки в 220 сантиметров на гидрологическом посту в поселке Ванавара.
2. Прогноз начала срока навигации должен быть составлен с использованием данных за 2 недели до самого события.
3. Прогноз окончания срока навигации должен быть составлен с использованием данных за неделю до данного события.
4. Ошибка прогноза не должна превышать одних суток.

**Гипотезы задачи:**

Проверить гипотезу о достаточности данных об уровнях рек с постов гидрологического контроля, а также данных метеосводок для решения задачи прогнозирования периода навигации на сезон.

## 1.2 Описание используемых методов

Для решения поставленной задачи выбраны методы:

* решение задачи регрессии для прогнозирования уровня реки Подкаменная Тунгуска на посту Ванавара. Для решения задачи регрессии использовались: линейная регрессия , деревья принятия решений, catboost.

Оценка качества моделей указана на рисунке 1,

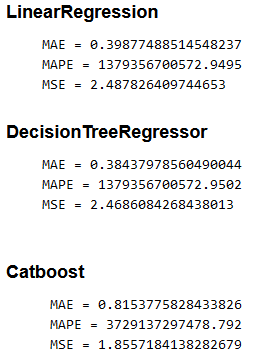


Рисунок 1 - Оценка качества моделей

Средняя квадратичная ошибка (MSE) для всех моделей находится около 2.

Среднее абсолютное отклонение в процентах (MAPE).

Средняя абсолютная ошибка (MAE)

## 1.3 Разведочный анализ данных

Для разведочного анализа данных использованы методы описательной статистики.

Датасет был проверен на наличие пропусков в значениях (команда df.isna().sum()). Пропусков обнаружено не было.

Команда Describe () позволила выявить, основные значения для всех параметров.

С помощью построения гистограмм было выявлено распределение величин, близкое к нормальному, для большей части параметров (исключение составили: поверхностная плотность – распределение со смещением вправо; угол нашивки – дискретная величина, график оказался не показателен).

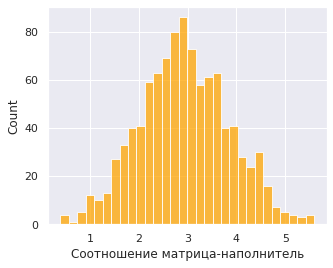


Рисунок 2 – Гистограмма «Соотношение матрица-наполнитель»

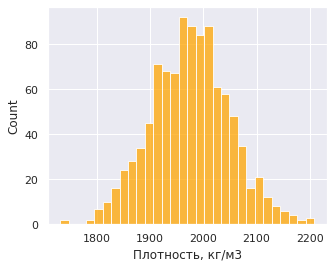


Рисунок 3 – Гистограмма «Плотность»

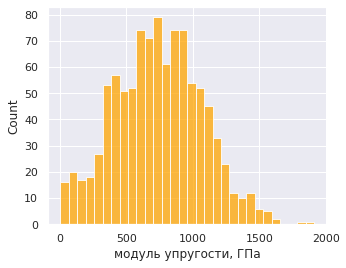


Рисунок 4 – Гистограмма «Модуль упругости»



Рисунок 5 – Гистограмма «Количество отвердителя»

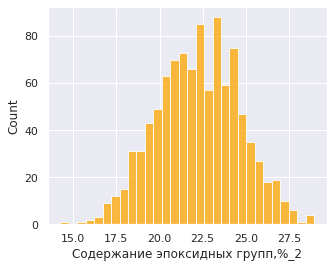


Рисунок 6 – Гистограмма «Содержание эпоксидных групп»

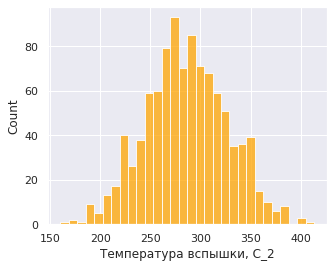


Рисунок 7 – Гистограмма «Температура вспышки»

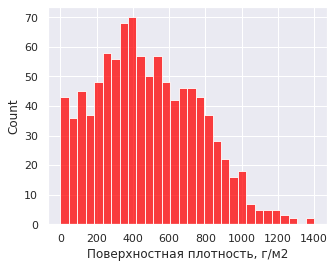


Рисунок 8 – Гистограмма «Поверхностная плотность»

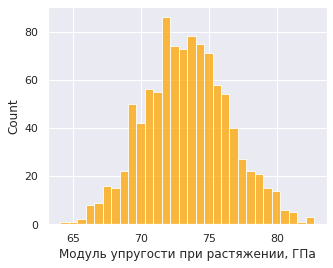


Рисунок 9 – Гистограмма «Модуль упругости при растяжении»

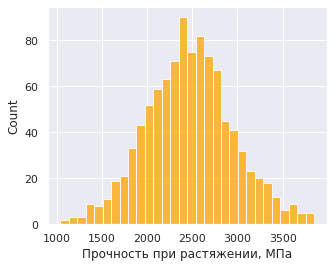


Рисунок 10 – Гистограмма «Прочность при растяжении»

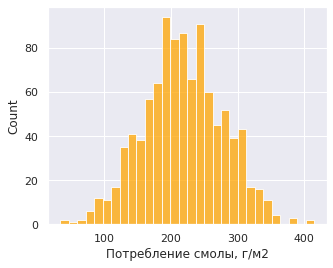


Рисунок 11 – Гистограмма «Потребление смолы»

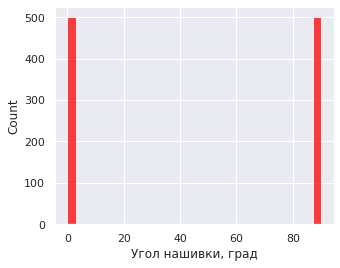


Рисунок 12 – Гистограмма «Угол нашивки»



Рисунок 13 – Гистограмма «Шаг нашивки»



Рисунок 14 – Гистограмма «Плотность нашивки»

С помощью диаграммы «ящик с усами» для всех параметров были выявлены выбросы.

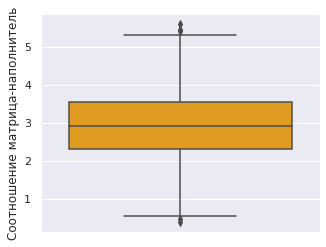


Рисунок 15 – Диаграмма «Соотношение матрица-наполнитель»

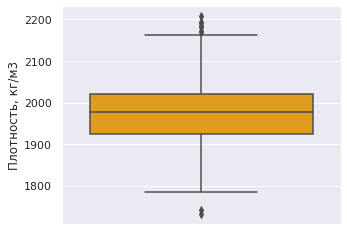


Рисунок 16 – Диаграмма «Плотность»

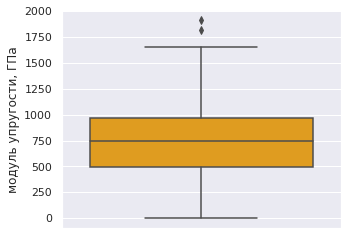


Рисунок 17 – Диаграмма «Модуль упругости»

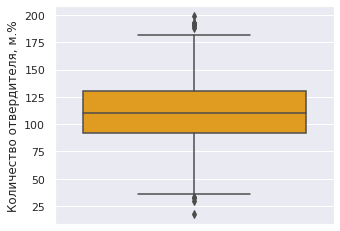


Рисунок 18 – Диаграмма «Количество отвердителя»

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 19 – Диаграмма «Содержание эпоксидных групп»

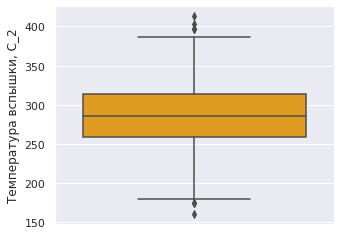


Рисунок 20 – Диаграмма «Температура вспышки»

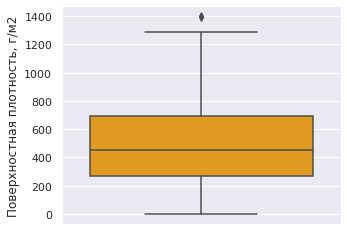


Рисунок 21 – Диаграмма «Поверхностная плотность»

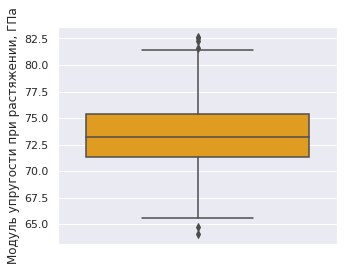


Рисунок 22 – Диаграмма «Поверхностная плотность»

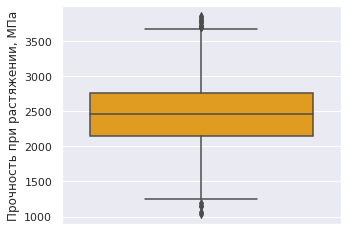


Рисунок 23 – Диаграмма «Прочность при растяжении»

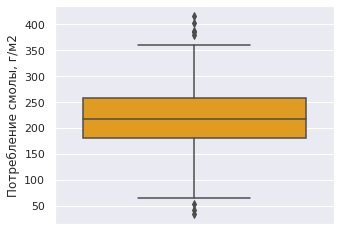


Рисунок 24 – Диаграмма «Потребление смолы»



Рисунок 25 – Диаграмма «Угол нашивки»

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 26 – Диаграмма «Шаг нашивки»

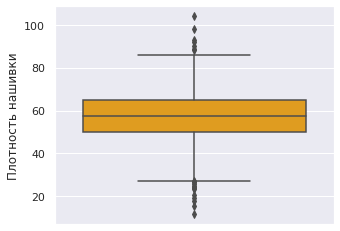


Рисунок 27 – Диаграмма «Плотность нашивки»

Выбросы наблюдаются по всем параметрам, кроме угла нашивки, т.к. данный параметр принимает дискретные значения и диаграмма «ящик с усами» для него не показательна.

Также следует отметить наличие выбросов с двух сторон (например, прочность при растяжении), наличие выбросов со стороны наименьших значения (например, содержание эпоксидных групп), наличие выбросов со стороны наибольших значений (например, шаг нашивки).

Для разведочного анализа было использовано также построение попарных графиков рассеяния точек. Графики для всех параметров показали отсутствие зависимости между переменными датасета.

На рисунке 28 приведена тепловая карта коэффициентов корреляции, значения которой показывают, что все полученные коэффициенты корреляции находятся в промежутке значений от - 0,25 до 0,25. Исходя из этого можно сделать вывод, что исходный датасет был предварительно обработан (либо сгенерирован) и переменные являются независимыми.

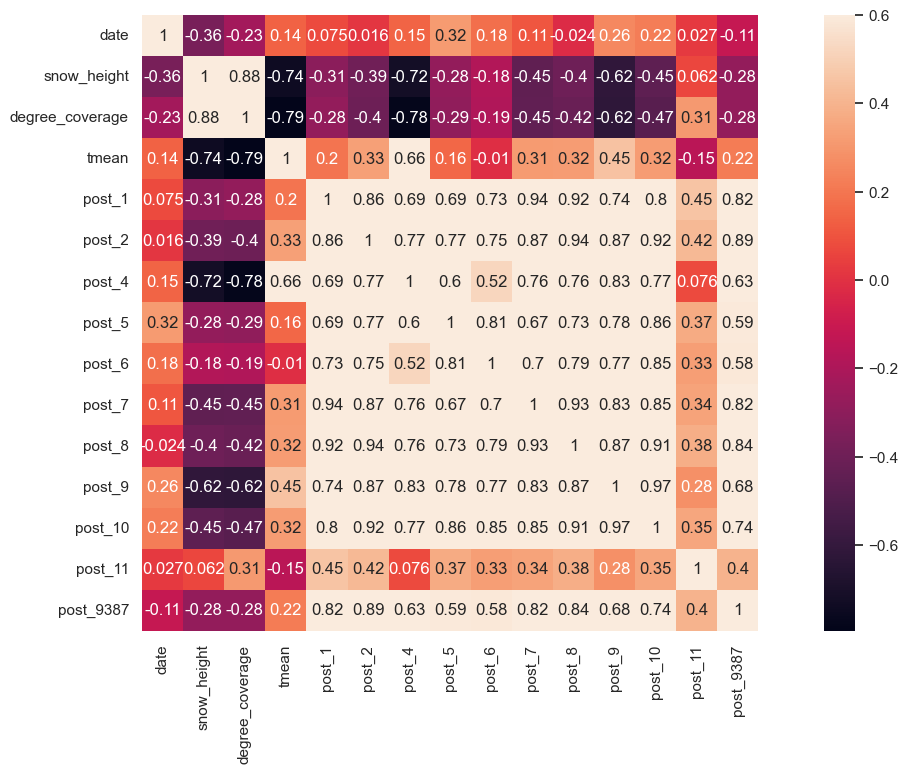


Рисунок 28 – Тепловая карта коэффициентов корреляции

# 2 Практическая часть

## 2.1 Предобработка данных

Предобработка данных осуществлялась на основании разведочного анализа данных, который показал наличие выбросов. Для удаления выбросов был произведен расчет количества выбросов для каждого параметра, данные приведены на рисунке 29.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 29 – Расчет количества выбросов

Количество выбросов говорит о их незначительности, соответственно, данные значения можно удалить из датасета.

Размер датасета после удаления выбросов и проверка наличия пропусков:

* Количество строк в очищенном датасете: 918;
* Количество столбцов (переменных) в очищенном датасете: 13;
* Количество пропусков в данных очищенного датасета: 0.

Вызов команды df\_used.describe() для датасета показывает, что количество строк датасета уменьшилось. Датасет очищен от выбросов.

Для выделения наиболее весомых признаков датасета использован факторный анализ.

С помощью метода главных компонент были получены значения влияния 90% факторов (данные приведены на рисунке 30). Указанные значения позволяют сделать вывод о том, что влияние очень слабое.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 30 – Вклад влияния 90% факторов

С целью определения весов характеристик датасета для первых пяти факторов влияния был проведен анализ, последовательно добавляя по одному фактору. Полученные результаты

* Для двух факторов: выделить наименование и смысл нового фактора сложно. но можно увидеть высокие доли характеристик внутри факторов: модуль упругости, ГПа, Температура вспышки, С\_2 Плотность нашивки;
* Для трех факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки;
* Для четырех факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки Угол нашивки, град;
* Для пяти факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки Угол нашивки, град Потребление смолы, г/м2.

Анализ полученных результатов показывает, что возможно характеристики: Количество отвердителя, Плотность нашивки, Плотность, Угол нашивки, Потребление смолы, являются наиболее существенными для построения будущих моделей.

Для дальнейшей разработки и обучения модели была выполнена нормализация данных с помощью MinMaxScaler.

## 2.2 Разработка и обучение модели

В качестве модели выбрана линейная регрессия и случайный лес. Обучение и тестирование производилось для двух параметров: модуль упругости и прочность при растяжении (в соответствии с условиями задачи).

При построении моделей «Случайный лес» был осуществлен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой На рисунке 1 приведен анализ качества разработанных моделей.

## 2.3 Тестирование модели

Из датасета была выделена прогнозируемая (зависимая) переменная. Далее, выборка была разделена на обучающую и тестовую выборки, в соответствии с условием задачи 70% (на обучение) /30% (на тестирование).

Размер обучающей выборки: 642.

Размер тестовой выборки: 276.

## 2.4 Нейронная сеть

В качестве нейронной сети был принят многослойный персептрон.

Гиперпараметры модели:

- количество скрытых слоев = 3;

- количество нейронов на слое = 256 и 64;

- активационная функция «relu»;

- количество нейронов на выходном слое = 1;

- оптимизатор «Adam»;

Обучение модели происходило за 20 эпох (количество задано после поиска опытным путем наиболее приемлемого распределения MSE и общих результатов обучения модели). Результат изменения MSE модели указан на рисунке 31. MSE уменьшается со временем по мере выполнения алгоритма. Это означает, что модель приближаемся к оптимальному решению.

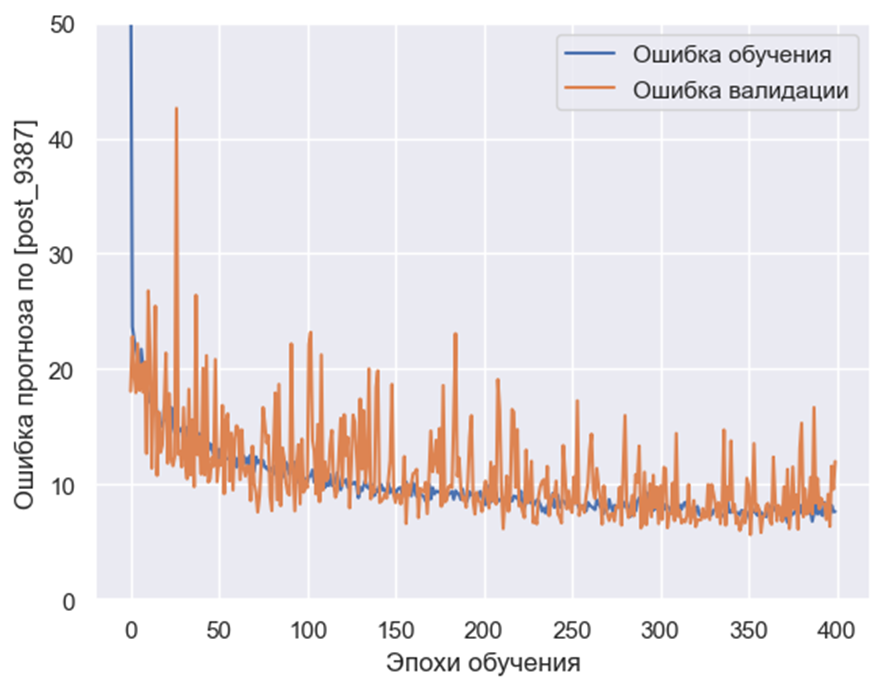


Рисунок 31 – Изменение MSE за время обучения модели

Ошибка в основном распределяется между -1 и 1, где 0 не является наибольшим. Это показывает, что модель обучения не совсем подходящая, ее можно улучшить, приведя значение ошибки к 0. Гистограмма распределения ошибки приведена на рисунке 32.

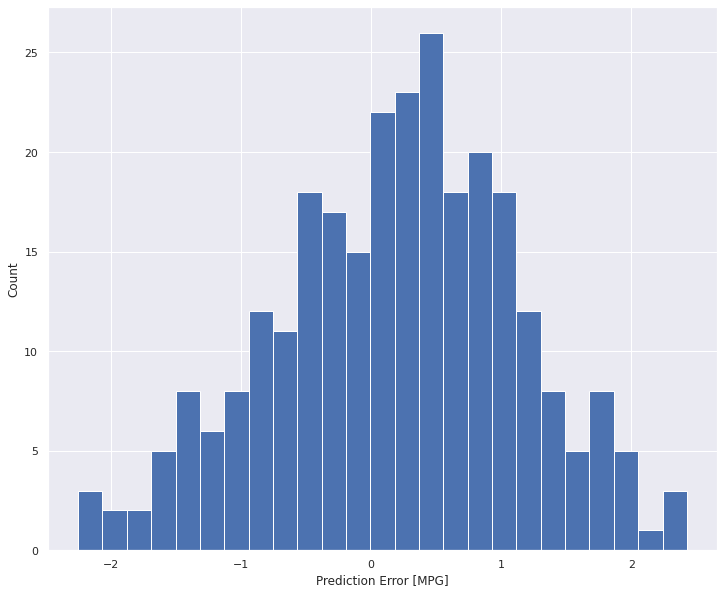


Рисунок 32 – Распределение ошибки

## 2.5 Разработка приложения

Приложение было разработано с помощью Flask.

В интерфейсе использованы наименования параметров из датасета (скриншот приведен на рисунке 33).

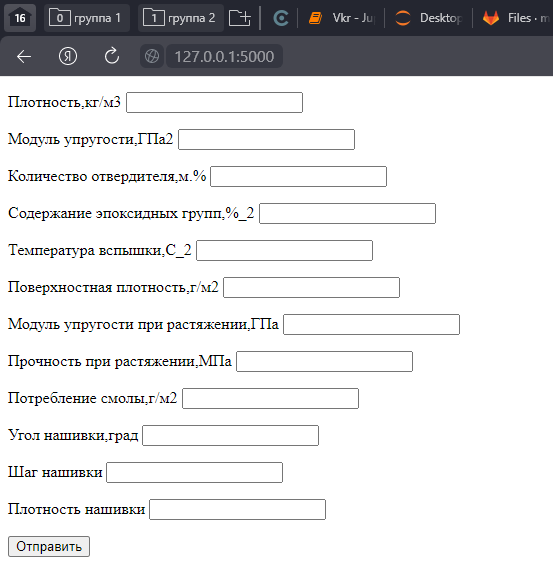


Рисунок 33 – Интерфейс приложения

При введении интересуемых данных и после нажатия кнопки «отправить» приложение, на основании модели, будет выводить рекомендуемое значение соотношение матрица – наполнитель.

# 3 Создание удаленного репозитория

Страница слушателя на GitHab

Созданный репозиторий:

https://github.com/00Apollo00/Rivers

Страница слушателя, созданный репозиторий, коммиты в репозитории.

https://github.com/00Apollo00/Rivers

# Заключение

В ходе выполнения ВКР были изучены способы анализа и предобработки данных. Построенные модели показали, что исходный датасет является предобработанным и не содержит реальных значений для отработки обучения и тренировки моделей.

Полученная модель нейронной сети не идеальна, но позволяет предсказывать значения, близкие к средним значениям параметров.

# Библиографический список

1. Л.И. Бондалетова, В.Г. Бондалетов Полимерные композиционные материалы: - Режим доступа - https://portal.tpu.ru/SHARED/b/BONDLI/stud\_work/p\_k\_m\_m/Tab1/Posobie\_PCM.pdf. (дата обращения - 10.06.2022).