Представляю ВКР «Прогнозирование конечных свойств новых композиционных материалов».

В первой части работы проведен разведочный анализ данных для получения первоначальных представлений о **характерах** распределений переменных **исходного** набора данных, формирования оценки качества исходных данных (наличие пропусков, дубликатов, выбросов), выявления характера взаимосвязи между переменными. Это нужно для подбора п**одходящих** для решения задачи **моделей** машинного обучения. В качестве инструментов использовались: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной; диаграммы ящика с усами; попарные графики рассеяния точек; график «квантиль-квантиль»; тепловая карта; описательная статистика для каждой переменной; анализ выбросов; проверка наличия пропусков и дубликатов; ранговая корреляция Спирмена, Кендалла и Пирсона .

Во второй части работы я создала модель линейной регрессии после предобработки данных. После объединения датасета по **индексу** тип объединения INNER, я удалила первый неинформативный столбец с индексами. Затем провела с помощью MinMaxScaler нормализацию всех данных. Потом методом межквартильного расстояния удалила выбросы. Сначала создала 1norm-модель линейной регрессии для признака «Модуль упругости при растяжении, Гпа», используя все остальные признаки. Получившаяся модель плохо выдавала прогнозные значения на тестовой выборке. Поэтому построила новую модель model\_norm2, используя только 2 признака 'Содержание эпоксидных групп,%\_2' и 'Потребление смолы, г/м2'. Получившийся коэффициент детерминации стал положительным, хотя и оказался слишком мал (0.01). Затем попробовала построить такие же модели линейной регрессии на StandardScaler-стандартизированных предварительно очищенных данных, но результат не оправдал ожиданий. Поэтому при создании модели линейной регрессии для определения значения "**Прочность** при растяжении, МПа", работала на MinMaxScaler-нормализованных данных по одному признаку 'модуль упругости, ГПа'.

В третьей части работы применила следующие алгоритмы машинного обучения: метод опорных векторов; случайный лес; градиентный бустинг; К-ближайших соседей; дерево решений; стохастический градиентный спуск. При создании моделей К-ближайших соседей использовала 5 и 20 соседей. Лучше результаты оказались при меньшем количестве соседей. Во всех созданных моделях коэффициент детерминации получился отрицательным. Таким образом, ни одна из рассмотренных моделей не годится для прогноза ‘Модуль упругости при растяжении, ГПа’ и 'Прочность при растяжении, МПа' по предоставленному набору данных.

В четвертой части создала модели нейронной сети. Использовала Sequential. Функции активации слоев relu, выходного слоя sigmoid. Экспериментировала с количеством слоев, количеством нейронов, эпох. Коэффициент детерминации везде оказывался отрицательным. И визуализация показывает несостоятельность моделей.

Таким образом, использованные при разработке моделей подходы не позволили получить сколько-нибудь достоверных прогнозов.

Добавила в репозиторий Пояснительную записку, README, презентацию ВКР, модель lr\_model.pkl.