**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)»**
**(МГТУ им. Н.Э. Баумана)**

ФАКУЛЬТЕТ РТ

КАФЕДРА ИУ5

# РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

## *К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ*

## *НА ТЕМУ:*

***Задача регрессии на основе датасета "Air Quality Dataset"***

Студент РТ5-61Б (Группа) _____ (Подпись, дата) **Мамаев Т.Э.** (И.О.Фамилия)

Руководитель _____ (Подпись, дата) **Гапанюк Ю.Е.** (И.О.Фамилия)

Консультант _____ (Подпись, дата) **Гапанюк Ю.Е.** (И.О.Фамилия)

*2023 г.*

**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)»**
**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой _____
(Индекс)

_____ _____
(И.О.Фамилия)

« _____ » _____ 20 ____ г.

# З А Д А Н И Е
## на выполнение научно-исследовательской работы

по теме _____Задача регрессии на основе датасета "Air Quality Dataset"_____

_____

_____

Студент группы ____РТ5-61Б__

_____Мамаев Темирхан Эльмирханович_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

_____Исследовательская_____

Источник тематики (кафедра, предприятие, НИР) _____Кафедра_____

График выполнения НИР: 25% к _4_ нед., 50% к _8_ нед., 75% к 12_ нед., 100% к _16_ нед.

***Техническое задание*** _____

_____

_____

_____

_____

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на __5_ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
__Исходный код с графиками в формате .ipynb_____

_____

_____

_____

Дата выдачи задания « ___ » _____ 20__ г.

**Руководитель НИР** _____ ____Гапанюк Ю.Е.____
(Подпись, дата) (И.О.Фамилия)

**Студент** _____ **Мамаев Т.Э.**
(Подпись, дата) (И.О.Фамилия)

Содержание:

# Введение

В данной работе мы занимаемся решением задачи регрессии с использованием выбранного набора данных. Задача состоит в построении моделей машинного обучения, способных предсказывать целевую переменную на основе доступных признаков. В работе используются различные методы и подходы для оптимизации моделей и достижения наилучшего качества предсказаний.

# Постановка задачи

Мы выбирали набор данных "Air Quality Dataset" и строим модели регрессии для предсказания целевой переменной. Целью является достижение высокой точности предсказаний и улучшение производительности моделей путем оптимизации гиперпараметров.

# Последовательность действий по решению задачи

Мы выполнили следующие шаги для решения задачи регрессии:

1. Поиск и выбор набора данных "Air Quality Dataset".
2. Проведение разведочного анализа данных для понимания структуры данных.
3. Заполнение пропусков в данных, если таковые имеются.
4. Выбор подходящих признаков для построения моделей и кодирование категориальных признаков.
5. Масштабирование данных для обеспечения одинакового масштаба признаков.
6. Формирование вспомогательных признаков, которые могут улучшить качество моделей.
7. Проведение корреляционного анализа данных для выявления зависимостей между признаками.
8. Выбор метрик для оценки качества моделей, таких как среднеквадратичная ошибка (MSE), средняя абсолютная ошибка (MAE) и коэффициент детерминации (R^2).
9. Выбор моделей для задачи регрессии, включая Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor и Bagging Regressor.
10. Формирование обучающей и тестовой выборок на основе исходного набора данных.
11. Построение базового решения (baseline) для моделей без подбора гиперпараметров.
12. Подбор оптимальных гиперпараметров для выбранных моделей с использованием методов кросс-валидации, таких как GridSearchCV.

13. Повторное обучение моделей с оптимальными гиперпараметрами и сравнение с базовыми моделями.
14. Формирование выводов о качестве моделей на основе выбранных метрик и визуализация результатов.

## Заключение

В результате выполнения работы мы получили опыт в выборе, предварительной обработке и построении моделей машинного обучения для задачи регрессии. Оптимизация гиперпараметров и сравнение с базовыми моделями позволили получить значительное улучшение качества предсказаний. Работа показала, что правильный выбор признаков, подходящих моделей и оптимальных гиперпараметров являются важными факторами для достижения высокой точности моделей регрессии.

Список использованных источников информации:


1. Air Quality Dataset. [Online].
2. Documentation of Scikit-learn - Machine Learning in Python. [Online].
3. Kaggle. [Online].
4. Python Data Science Handbook by Jake VanderPlas. O'Reilly Media, 2016.
5. Python for Data Analysis by Wes McKinney. O'Reilly Media, 2017.