**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное автономное образовательное учреждение высшего образования**
**«Южный федеральный университет»**

**Институт высоких технологий и пьезотехники**


**Кафедра прикладной информатики и инноватики**

**Направление: 09.03.03 "Прикладная информатика"**

# Большие данные

# Отчёт по проекту по теме:

# «Предобработка и анализ данных систем противопожарной безопасности»

Выполнили студенты 3 курса 2_ВТ-09.03.03.01-о_ группы:

_____ Логинов А.И.

подпись

_____ Юрченко Е.В.

подпись

**Ростов-на-Дону – 2024**

Оглавление

## Введение

### Постановка задачи

В современном мире пожары представляют серьезную угрозу для жизни и здоровья людей, а также наносят огромный материальный ущерб. Мы решили использовать технологию машинного обучения для того чтобы оптимизировать процессы сигнализации о возгорании, а также снизить процент ложных срабатываний систем противопожарной безопасности.

### Актуальность темы

Прогнозирования пожароопасных ситуаций: Анализируя исторические данные о пожарах, погодные условия, состояние пожарной безопасности объектов, системы могут выявлять зоны повышенного риска и прогнозировать потенциальные возгорания.

Вот лишь некоторые факторы, подчеркивающие важность противопожарной безопасности:

**Человеческие жертвы**: Пожары являются одной из основных причин случайной смерти в мире.

**Материальный ущерб**: Пожары уничтожают здания, инфраструктуру, оборудование, запасы – нанося значительный экономический ущерб.

**Экологический ущерб**: Пожары, особенно лесные, выбрасывают в атмосферу огромное количество вредных веществ, загрязняют воздух и способствуют изменению климата.

**Психологическая травма**: Пожары – это огромный стресс для людей, переживших их. Многие сталкиваются с посттравматическим стрессовым расстройством.

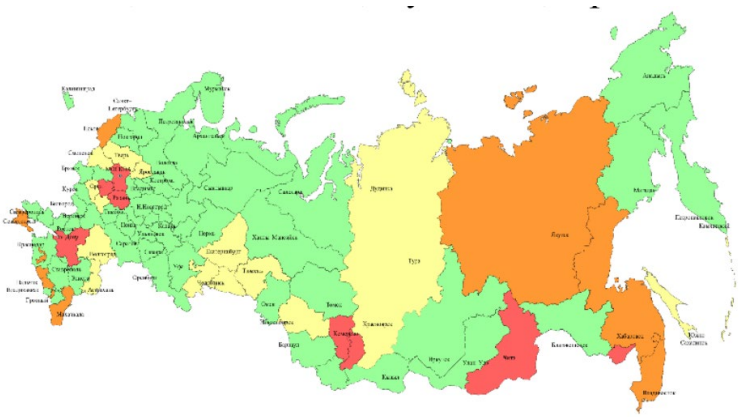
## Статистика

Статистика пожаров в России, по причине неисправности систем противопожарной безопасности

"В течение 2023 года пожарно-спасательные подразделения реагировали на более чем 350 тыс. пожаров, в которых погибло порядка 7,2 тыс. человек, более 26 тыс. - спасены", - сказали в ведомстве. По сравнению с 2022 годом, как отмечают в МЧС, число пожаров снизилось на 1,5%, а гибель в огне - на 2%.

Использование технологий машинного обучения, для улучшения систем противопожарной безопасности, поможет спасти тысячи человеческих жизней.

Риск возникновения пожаров на объектах жилого, социально-бытового и культурного назначения



## Цель работы

Целью нашего проекта является анализ и предобработка систем противопожарной безопасности, исходя из статистики, это очень актуальная проблема на сегодняшний день.

Мы решили изучить эту тему подробно, используя технологии машинного обучения, а в частности Spark ml lib, изучением которого мы занимались в течение курса.

## Описание датасета.

Мы использовали датасет с площадки Kaggle.

Датасет представляет собой реальный набор данных, в котором перечисляются различные примеси газов в воздухе, а также метки, в которых указано был ли на самом деле пожар, или это тревога ложная. На основании этих данных можно будет предсказывать, когда возможно возгорание, а также сократить количество ложных тревог.

Общая информация:

Датасет состоит из 62 тысяч строк записей, каждая из которых описывается 10 столбцами данных, в которых содержится информация о количестве примесей газов в воздухе

**Данные столбцов:**

1) id строки

2) Температура воздуха(в градусах цельсия)

3) Относительная влажность воздуха

4) Общее количество летучих органических соединений, измеряемое в частях на миллиард

5) Эквивалентная концентрация CO2

6) Необработанный молекулярный водород

7) Сырой газообразный этанол

8) Давление воздуха

9) Размер твердых частиц < 1,0 мкм (PM1.0).1,0 мкм < 2,5 мкм (PM2.5)

## Ход работы

## Гипотеза

Исходя из полученных данных, мы выдвинули гипотезу, о том что можно будет улучшить процент корректно сработанных сигналов систем противопожарной безопасности на основании данных датасета.

## Машинное обучение в Apache Spark ml lib

Вот некоторые из преимуществ этой среды машинного обучения:

1. Распределённые вычисления:

Ядро Spark: В основе ML lib лежит мощный движок Spark для распределённых вычислений. Это значит, что данные и задачи обработки распределяются по кластеру компьютеров (узлов), что позволяет значительно ускорить обучение моделей.

Параллелизм:Spark может обрабатывать множество задач одновременно (параллельно), что особенно полезно для алгоритмов машинного обучения, требующих итеративной обработки данных.

2. Оптимизация выполнения:

DAG (Directed Acyclic Graph): Spark строит ациклический граф вычислений, представляющий собой последовательность операций, которые необходимо выполнить. Это позволяет оптимизировать выполнение задач и минимизировать перемещение данных между узлами.

Ленивые вычисления: Spark не выполняет операции немедленно, а лишь строит план выполнения.  Фактическая обработка запускается только тогда, когда это действительно необходимо (например, для вывода результата), что позволяет избежать ненужных операций.

Кэширование: Spark может хранить промежуточные результаты вычислений в памяти кластера, что значительно ускоряет повторное использование данных.

3. Алгоритмическая оптимизация:

Реализации алгоритмов: ML lib предоставляет высокоэффективные реализации популярных алгоритмов машинного обучения, оптимизированные для работы на больших данных.

Итеративные алгоритмы: Многие алгоритмы машинного обучения являются итеративными.  Spark ML lib оптимизирует выполнение таких алгоритмов, минимизируя обмен данными между узлами на каждой итерации.

4. Масштабируемость:

Горизонтальное масштабирование: Spark ML lib легко масштабируется горизонтально – для обработки большего объёма данных достаточно просто добавить в кластер новые узлы.

Линейное масштабирование: В идеальных условиях Spark ML lib демонстрирует почти линейное масштабирование. Это значит, что увеличение вычислительных мощностей в 2 раза приводит к сокращению времени обучения модели примерно в 2 раза.

5. Интеграция с экосистемой Hadoop:

Spark ML lib отлично интегрируется с экосистемой Hadoop, что позволяет использовать его для обработки данных, хранящихся в Hadoop Distributed File System (HDFS).

В итоге, Spark MLlib  - это мощный инструмент для машинного обучения на больших данных, который обеспечивает высокую скорость обработки, масштабируемость и эффективность благодаря распределённым вычислениям, оптимизации выполнения, алгоритмической оптимизации и тесной интеграции с экосистемой Hadoop.

## Методы машинного обучения

Для того чтобы проанализировать, и выявить лучший метод, с наибольшей точностью, мы решили взять по методу из каждого типа машинного обучения

**Машинное обучение с учителем:**

Классификация:

Метод Дерева принятий решений

Точность составляет 99%

Метод Случайного Леса

Точность составляет 99%

Регрессия:

Метод Линейной Регрессии

Точность составляет 84,7%

Метод Опорных Векторов

Точность составляет 84%

## Визуализация данных

Для визуализации данных нашего датасета мы использовали самую популярную и удобную библиотеку на Python – matplotlib.

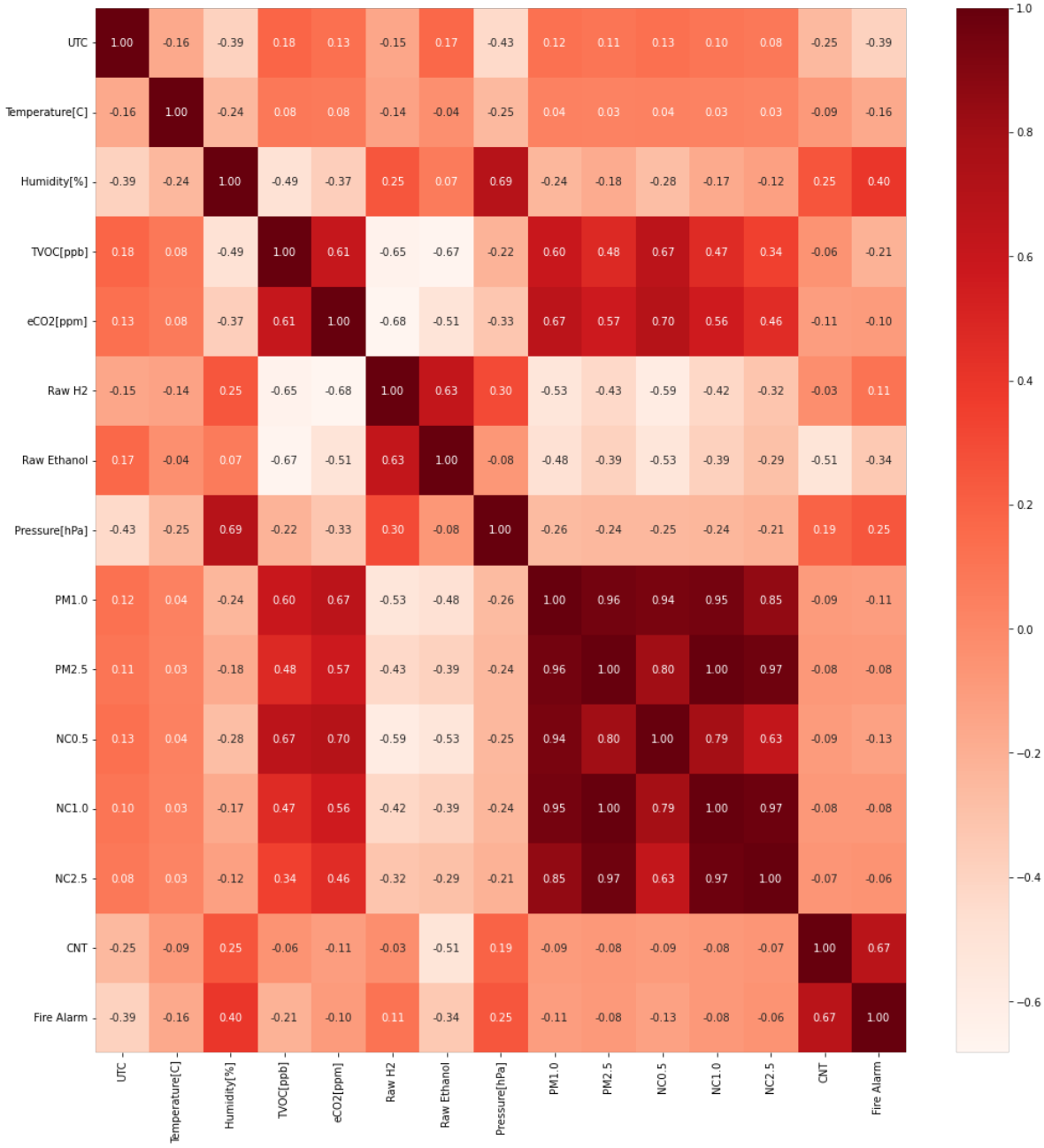
Корреляционный график (тепловая карта корреляции) визуализирует корреляцию между признаками, что позволяет:

1. Обнаружить мультиколлинеарность: Сильная корреляция (положительная или отрицательная) отображается яркими цветами, что помогает быстро выявить проблемные пары признаков.

2. Понять взаимосвязь признаков: График дает общее представление о структуре данных и взаимосвязях между переменными.

3. Принять решение о выборе признаков: Помогает решить, нужно ли удалять или комбинировать сильно коррелирующие признаки для улучшения модели.

График корреляции.




Как можно заметить, на нашем графике преимущественно светлые цвета, следовательно набор данных отлично подходит для машинного обучения.

График соотношения меток, где 1 – срабатывание сигнализации

0 – несрабатывание сигнализации

Метки соотносятся примерно 50/50

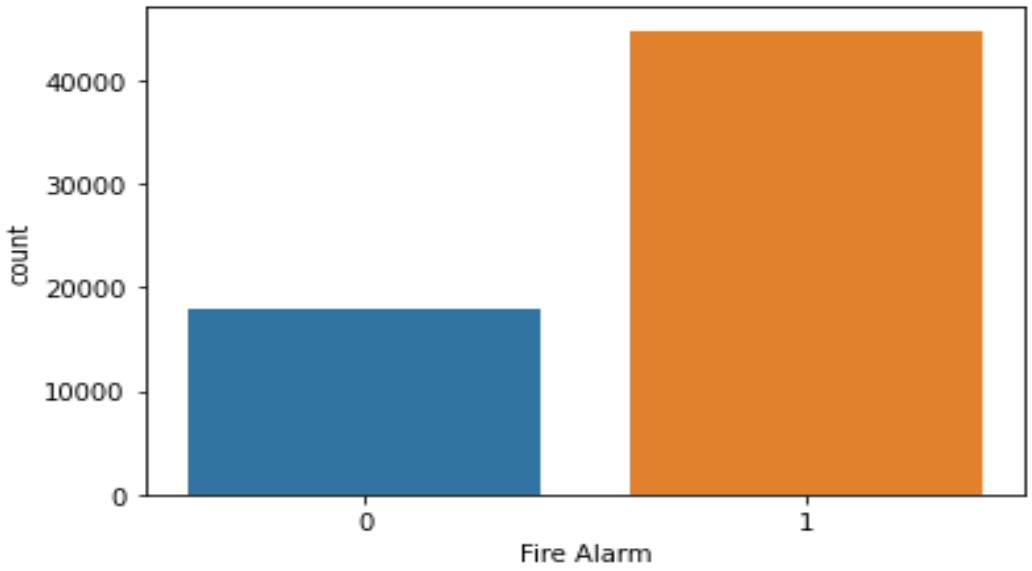



График влажности воздуха, в %

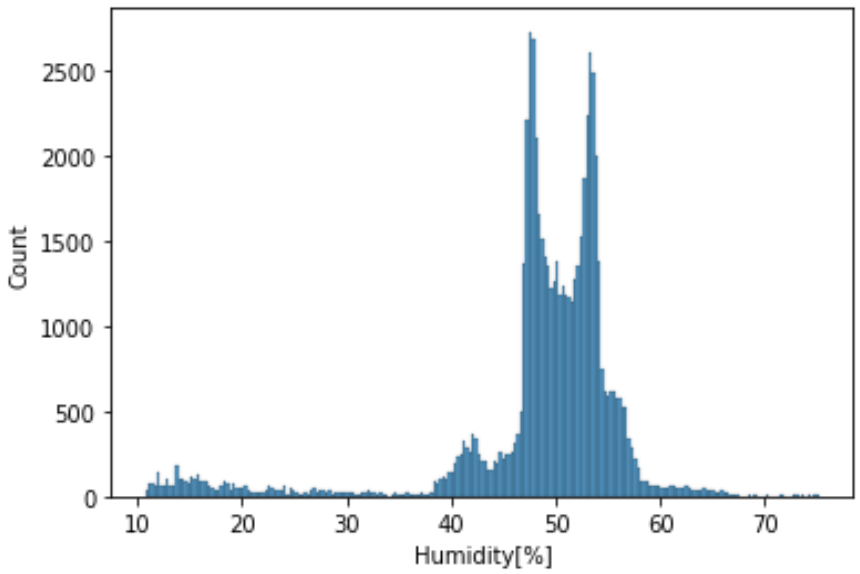
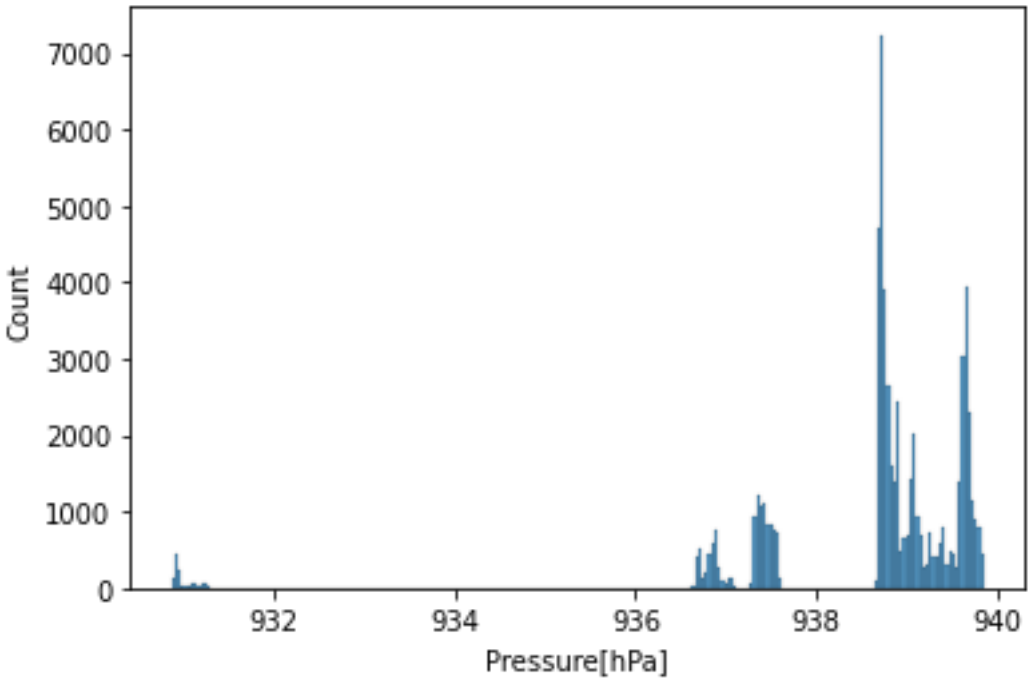


График давления, в кПа




## Вывод

В ходе решения данного кейса, было сделано:

- Проанализирован датасет, отобраны наиболее информативные данные
- Выбраны подходящие методы для модели работы с нелинейными зависимостями.
- Построены графики в библиотеке matplotlib
- Проведено обучение и тестирование выбранных моделей для достижения наилучших результатов в прогнозировании средней температуры.