**МИНОБРНАУКИ РОССИИ**

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

**Институт высоких технологий и пьезотехники**

**Изображение выглядит как текст, логотип, Шрифт, символ

Автоматически созданное описание**

**Кафедра прикладной информатики и инноватики**

**Направление: 09.03.03.01 "Прикладная информатика"**

**Большие данные  
 Отчёт по проекту**

**“** **Сбор, предобработка и анализ информации, предсказание дождя на следующий день в Австралии”**

Выполнили студенты 3 курса ВТ-09.03.03.01-о3 группы

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Ли К. Э.

подпись

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Джамалбеков А. Э.

подпись

**Ростов-на-Дону – 20****24**

**Содержание**

[**1 Цели и задачи проекта** 3](#_Toc170343229)

[**2 Ход работы** 4](#_Toc170343230)

[2.1 Импортирование библиотек 4](#_Toc170343231)

[2.2 Загрузка данных 4](#_Toc170343232)

[2.3 Визуализация данных 6](#_Toc170343233)

[2.4 Предобработка данных 8](#_Toc170343234)

[2.5 Кодирование 10](#_Toc170343235)

[2.6 Разделение данных 11](#_Toc170343236)

[2.7 Обучение модели и анализ предсказаний 11](#_Toc170343237)

[**3 Заключение 17**](#_Toc170343238)

[**4 Литература 18**](#_Toc170343239)

# **1 Цели и задачи проекта**

Целью нашего проекта является обучение модели, которая будет удовлетворять следующим требованиям:

1. Модель должна предсказывать дождь на следующий день
2. Точность предсказания должна быть хорошей
3. Работа должна включать в себя полученные знания из курса

Для достижения целей проекта наша команда поставила перед собой некоторые задачи:

1. Предобработка и анализ данных
2. Изучение методов обучения и поиск наиболее подходящих для нашей модели
3. Повышение точности предсказания
4. Анализ предсказаний

Испокон веков люди пытались предсказать погодные изменения по приметам или с помощью гаданий. Современные методы прогнозирования значительно упростили жизнь. Например, вы когда-нибудь задумывались, стоит ли вам завтра брать с собой зонтик? С помощью этого проекта вы можете предсказать дождь на следующий день, используя машинное обучение.

# **2 Ход работы**

## **2.1 Импортирование библиотек**

Импортируем нужные библиотеки для дальнейшей работы.

**Изображение выглядит как текст, Шрифт, снимок экрана, алгебра

Автоматически созданное описание**

## **2.2 Загрузка данных**

При импортировании датасета замечаем, что структура данных не соответствует действительности, поэтому необходимо задать ее вручную.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Задаем схематику датафрейма и подгружаем его.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, черно-белый, Шрифт, снимок экрана

Автоматически созданное описание

Так выглядит наш датафрейм

Изображение выглядит как текст, снимок экрана, Шрифт, черно-белый

Автоматически созданное описание

Краткая сводка о столбцах

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Датафрейм состоит из 145460 строк

Изображение выглядит как текст, Шрифт, снимок экрана, белый

Автоматически созданное описание

## **2.3 Визуализация данных**

Диаграмма количества строк в каждом столбце.

Можно заметить большое количество отсутствующих данных в столбцах Sunshine, Evaporation, Cloud9am, Cloud3pm

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Заполненность данных построчно.

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Heatmap покажет зависимость одного признака от другого. Индекс корреляции варьируется от -1 до 1. Если значение близится к -1, то признаки обратно пропорциональны друг другу, к 1, то прямо пропорциональны, 0 – не зависят.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Изображение выглядит как шаблон, снимок экрана, прямоугольный, Красочность

Автоматически созданное описание

## **2.4 Предобработка данных**

Вывод количества null и NaN значений в каждом столбце.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Null значения будут заменены на средние в этом столбце. Это позволит не терять необходимые данные для предсказания, поскольку пробелы в датах не позволят нам покрывать все дни. Хотя большое количество одинаковых данных может привести к плохому обучению из-за однородности признака.

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

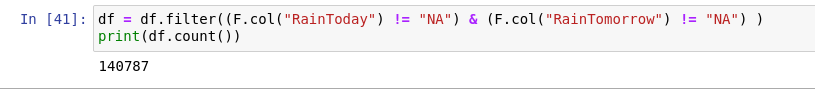
Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

RainToday и RainTomorrow должен иметь 2 категории – yes/no, однако есть и NA значения, которые указывают на отсутствие данных. Поскольку столбец String типа эти пробелы не были обнаружены в визуализации данных. Удаляем эти строки. Потеряли около пяти тысяч строк.



## **2.5 Кодирование**

Датафрейм имеет как числовые, так и категориальные признаки. Необходимо перевести все в числовой тип для дальнейшей работы с моделью.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Просмотр данных о дожде. Представление его количества на диаграмме.

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

## **2.6 Разделение данных**

Для обучения модели создан вектор признаков, который включает себя все столбцы, кроме RainTomorrowBin, поскольку он используется для предсказания.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Разделение данных на тренировочные и тестовые будет происходить случайным образом. Отношение 70% на 30%.

## **2.7 Обучение модели и анализ предсказаний**

В качестве методов обучения выбраны DecisionTree, RandomForest, LinearSVC.

Создание метода для отрисовки ROC кривой

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

**Метод решающего дерева** основан на определении, в какую из дочерних вершин нужно поместить рассматриваемый объект. Строится дерево из признаков, и, в зависимости от тех или иных значений, получаем результат будет дождь или нет

Изображение выглядит как текст, Шрифт, веб-страница, Веб-сайт

Автоматически созданное описание

Таблица и точность предсказания.

Изображение выглядит как текст, снимок экрана, черно-белый, число

Автоматически созданное описание

ROC кривая и F1 score.

Изображение выглядит как текст, линия, снимок экрана, График

Автоматически созданное описание

Метод случайного леса строит множество деревьев решений во время обучения и выводит класс в режиме классификации или среднее предсказание в режиме регрессий. Каждое дерево в случайном лесу обучается на различном подмножестве данных, а случайность, добавленная во время обучения, помогает улучшить общие показатели модели и ее обощающие способности.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Таблица и точность предсказания.

Изображение выглядит как текст, снимок экрана, документ, число

Автоматически созданное описание

ROC кривая и F1 score

Изображение выглядит как текст, линия, снимок экрана, График

Автоматически созданное описание

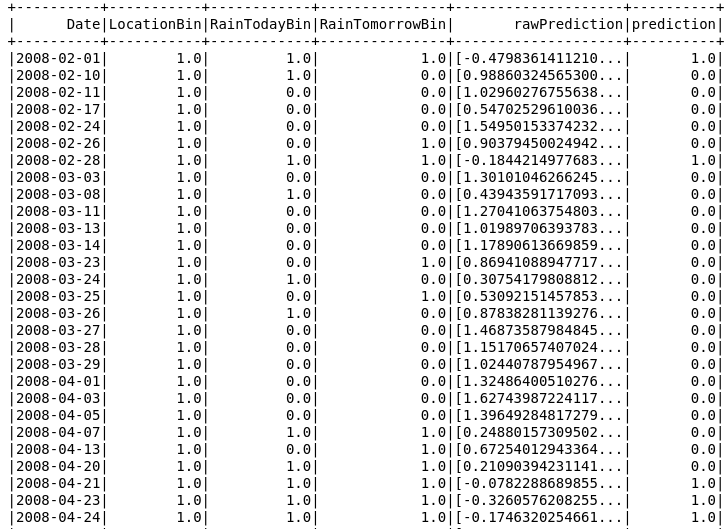
LinearSVC - алгоритм классификации, в основе которого лежит определяемая разделяющая гиперплоскость (линия, прямая, многомерные плоскости).

При заданных тренировочных данных алгоритм находит такую гиперплоскость, которая разделяет данные, принадлежащие разным классам, самым оптимальным способом. В двухмерном пространстве гиперплоскостью служит прямая линия.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Таблица предсказаний



Хотя и точность неплохая, ROC кривая и F1 score дают понять, что большое количество данных предсказаны ошибочно.

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

# **3 Заключение**

В ходе работы построена модель предсказания данных о дожде на следующий день. Было использовано 3 метода обучения: решающее дерево, случайный лес и метод опорных векторов. Лучшие предсказания точности были с использованием решающего дерева. Точность составила 83.6 %, что является хорошим результатом в сочетании с результатами F1 score и ROC score. Наихудшее обучение было с методом опорных векторов, хотя точность и составляла 80%, F1 score был очень низким, что указывает на плохую точность и классификацию данных.

# **4 Литература**

1. Датасет [Электронный ресурс]

URL: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package/data

2. Пример анализа и обработки данных [Электронный ресурс]

URL: [https://www.kaggle.com/code/fahadmehfoooz/rain-prediction-with-90-65-accuracy#3.-Training-The-Models](https://www.kaggle.com/code/fahadmehfoooz/rain-prediction-with-90-65-accuracy#3.-Training-The-Models(Дата)

3.Обучение модели [Электронный ресурс]

URL: <https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial#8.-Declare-feature-vector-and-target-variable->

4. Методы обучения [Электронный ресурс]

URL: <https://habr.com/ru/companies/ods/articles/322534/>