**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**

**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**(РУТ (МИИТ)**

Институт/факультет «Академия “Высшая Инженерная Школа”, АВИШ»

Специальность/Направление подготовки Информатика и вычислительная техника

**КУРСОВАЯ РАБОТА**

**по дисциплине:** Машинное обучение и анализ данных

|  |  |
| --- | --- |
| **на тему:** | "Прогнозирование тяжести ДТП на основе погодных условий с помощью машинного обучения" |

|  |  |
| --- | --- |
| **Студент группы ШАД 21**2 | **Гаганов А. А.** |
| **Научный руководитель** | **Атамасов Н. В.** |

**Москва 2023 г.**

Оглавление

[Основные термины и сокращения 3](#_Toc155738380)

# Основные термины и сокращения

ДТП – дорожно-транспортное происшествие.

Датасет – набор данных, используемый для анализа и машинного обучения.

EDA – (exploratory data analysis) разведочный анализ данных (этап работы с датасетом).

Енкодинг – процесс, с помощью которого категориальные переменные преобразуются в подходящую ​​алгоритмам машинного обучения форму с помощью кодирования.

# Введение

Современное общество сталкивается с постоянным увеличением числа (ДТП), что приводит к серьезным последствиям для людей и имущества. Одним из важных факторов, влияющих на обстановку на дорогах, являются погодные условия.

В данной курсовой работе будет рассмотрено применение методов машинного обучения для прогнозирования тяжести ДТП на основе климатических данных.

Изучение этой темы имеет большое значение для разработки инновационных систем безопасности на дорогах и снижения риска происшествий, что делает ее актуальной и важной для научного и практического аспектов.

# Описание задачи

Основной задачей стоит цель построить модель машинного обучения, которая на основе погодных условий будет составлять прогноз такой величины, как “тяжесть ДТП”. Такая задача называется мульти классификацией. Для того, чтобы оценить качество модели, будем рассматривать такие метрики, как F1-Score и Accuracy. Также необходимо провести моделирование с помощью библиотеки автоматизированного машинного обучения и собственной, выбранной модели, и сравнить их.

# Основные шаги выполнения. EDA и подготовка данных

В данной курсовой работе были использованы данные из Excel по ДТП – “DataDTPкор.xlsx” и по погоде – “ Pogoda.xlsx” После загрузки объединил все в один датасет по признаку “join\_key”. Затем выбрал интересующие временные условия, относящиеся к сумеркам или тёмному времени суток. Начальный объем данных стал из себя представлять 14269 строк и 28 столбцов, представляющие признаки.

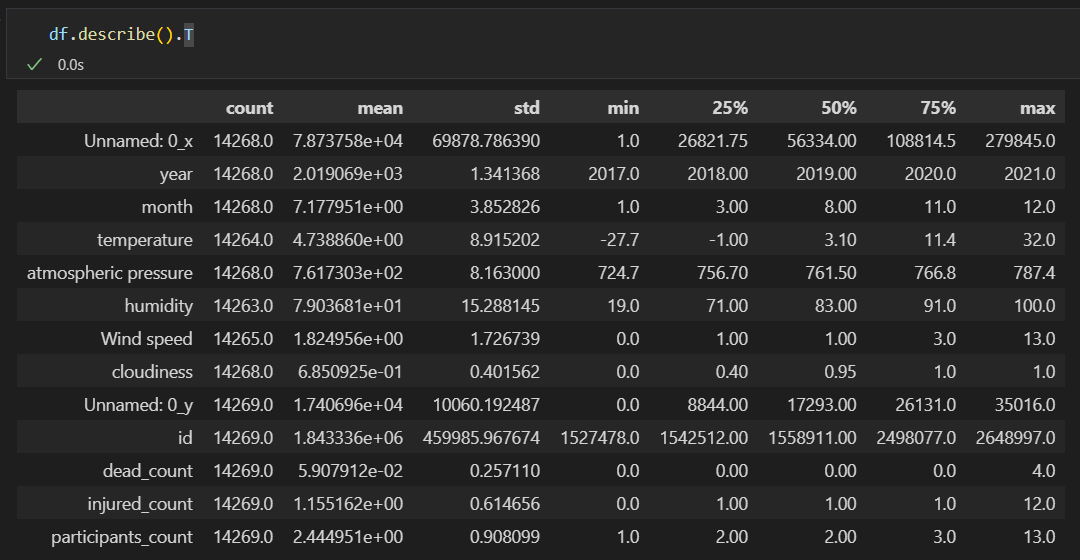
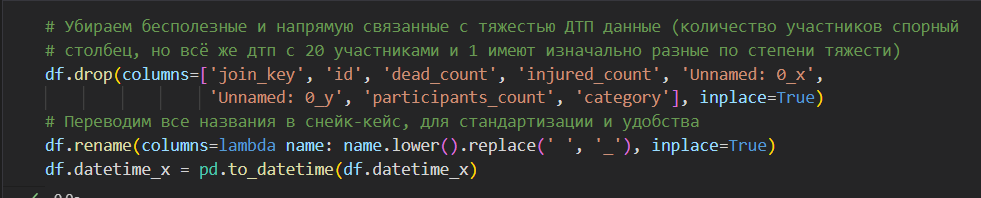


Рис. 1



Рис. 2

Основными действиями для этапа EDA стали удаление дубликатов, приведение корректного типа в некоторых столбцах, заполнение и удаление пропусков, удаление ненужных признаков.



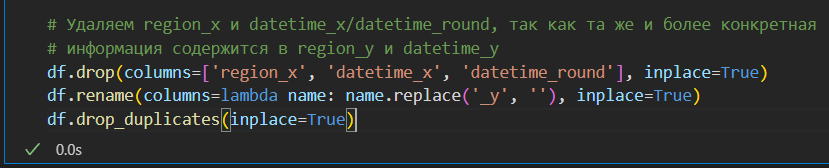
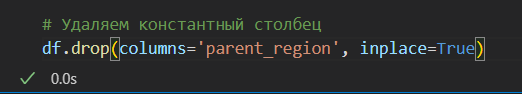


Рис. 3-4

Так же использовался ydata-profiling, посмотреть отчёт можно тут:

[Pandas Profiling Report (althgamer.ru)](https://althgamer.ru/report.html)



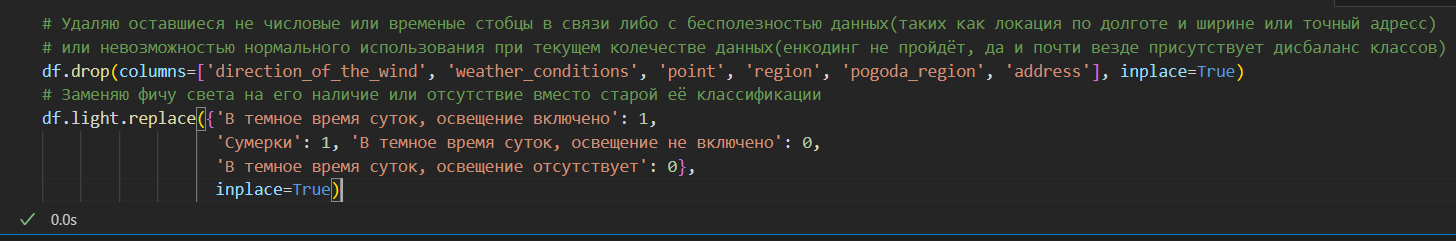


Рис. 5-6

Добавляю новые фичи, которые требуются по заданию.

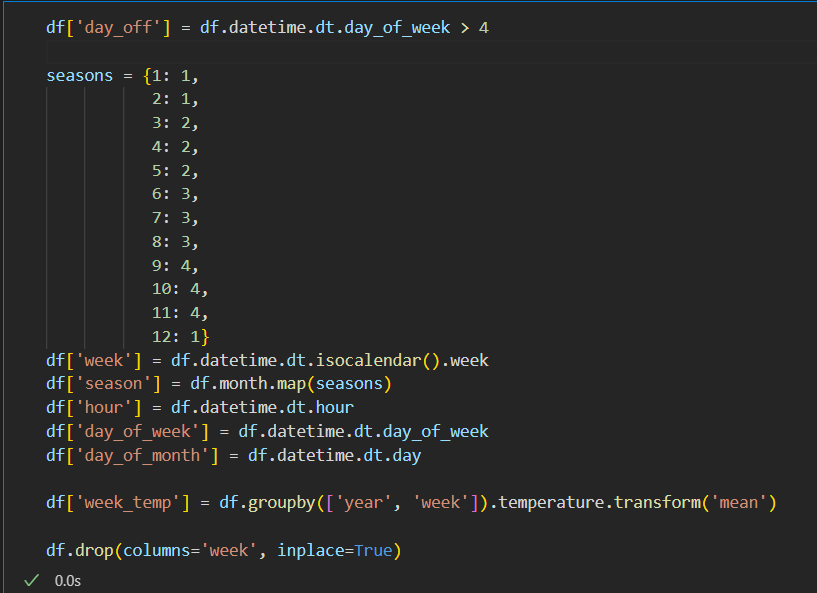


Рис. 7

Перекодирую целевую фичу по порядковым енкодингом согласно логике и удаляю все строки с пропущенными значениями.

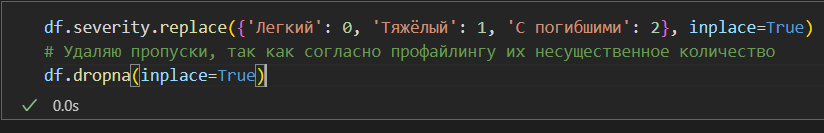


Рис. 8

# Основные шаги выполнения. Разделение на выборки

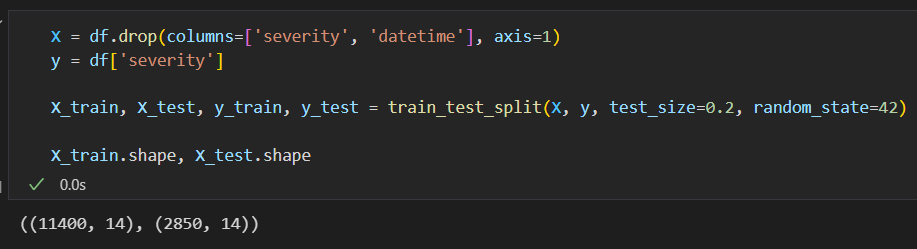


Рис. 9

Разделяем выборку на обучающую и тестовую в соотношение 80 на 20 с параметром random\_state равным 42, с random state равным 42, для устранения полной случайности.

После проверки, стало ясно, что попытка баланса классов приводит лишь к ухудшению результатов обучения, поэтому в итоговой версии его нет.

# Основные шаги выполнения. Построение модели

Выбор библиотеки для построения базовой модели выпал на TPOT. У TPOT есть то, что его разработчики называют алгоритмом генетического поиска для поиска лучших параметров и ансамблей моделей. Его также можно рассматривать как естественный отбор или эволюционный алгоритм. TPOT пробует конвейер, оценивает его производительность и случайным образом меняет части конвейера в поисках более эффективных алгоритмов.

Generations отвечает за количество поколений эволюционного алгоритма, где каждое поколение – новая финальная модель, а population\_size определяет количество моделей-кандидатов в одном поколении.

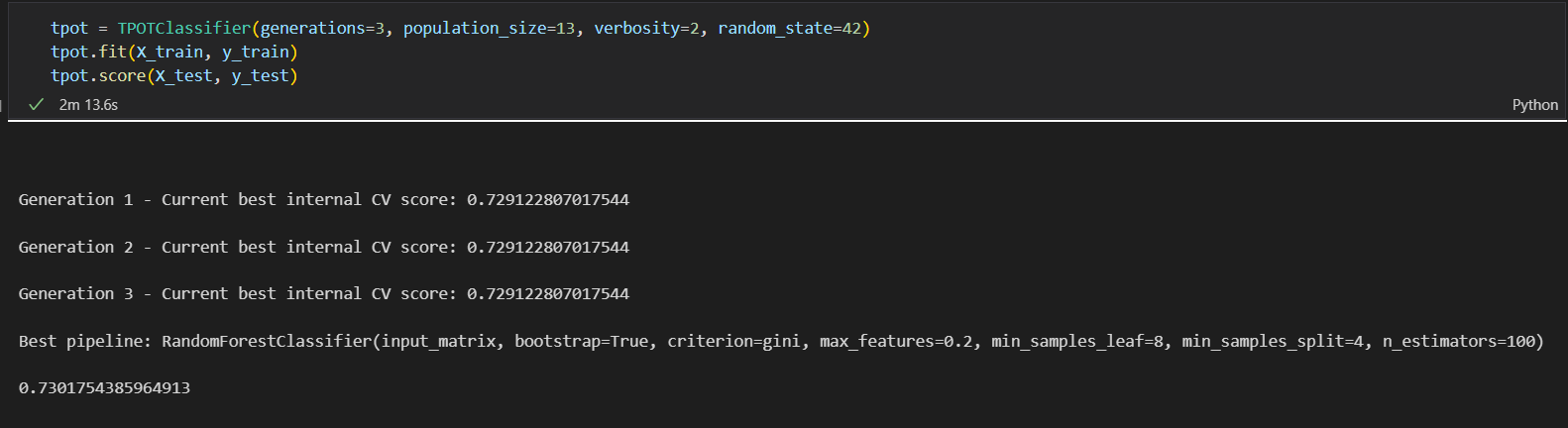


Рис. 10

В результате TPOT выбрал RandomForestClassifier, с итоговым результатом:

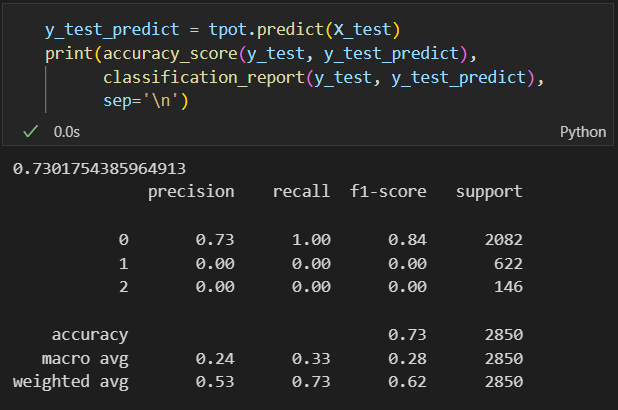


Рис. 11

# Основные шаги выполнения. Лучшая модель

На этом этапе я построил собственные модели машинного обучения с подбором гиперпараметров с помощью Optuna. Поставлена цель превзойти результат базовой автоматизированной модели по основным метрикам, например, как F1-Score. LogisticRegression и GradientBoostingClassifier выдали неплохие результаты, однако окончательный выбор пал на CatBoostClassifier, который с подбором гиперпараметров оказался лучше среди всех моделей, включая TPOT.

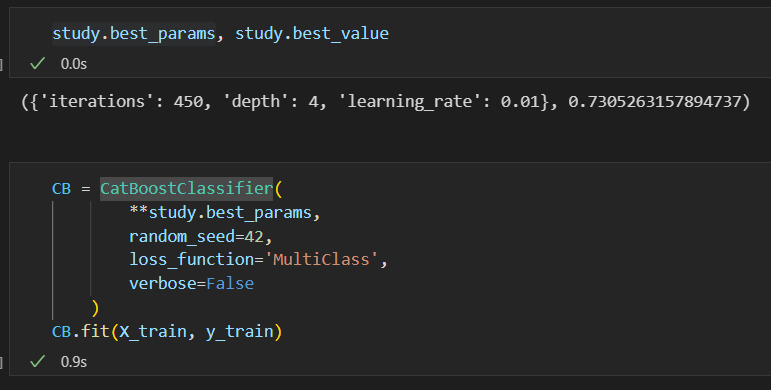


Рис. 12

Финальная модель категориального бустинга имеет параметры: количество итераций – 450, максимальная глубина дерева – 4, скорость обучения – 0.01 и имеет следующие результаты по метрикам:

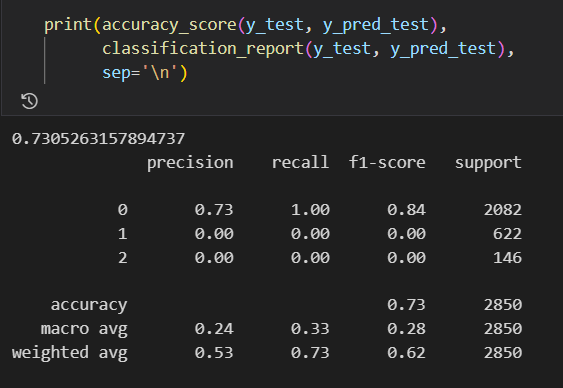


Рис.13

# Основные шаги выполнения. Интерпретация лучшей модели

Для интерпретации модели было

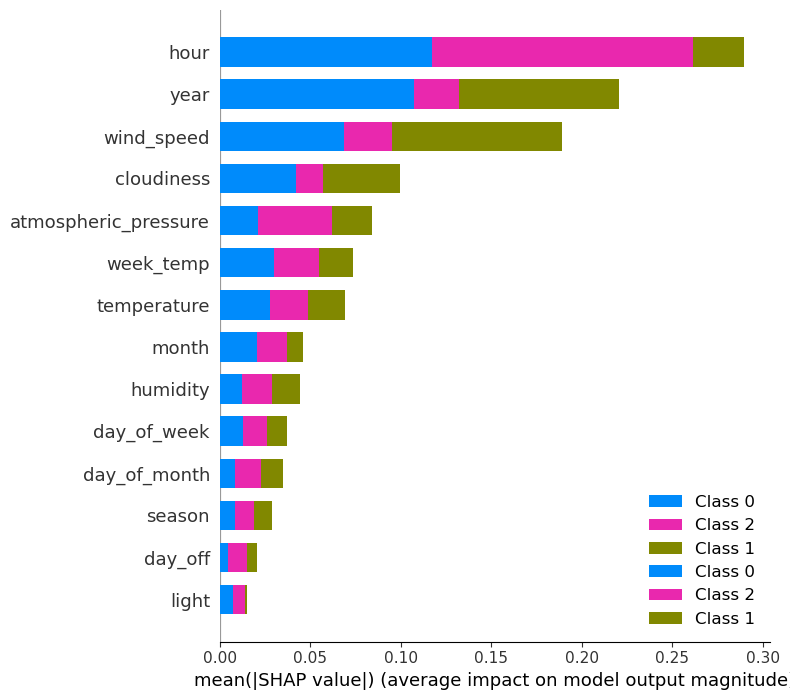


Рис. 14