**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 2](#_Toc14012944)

[Введение 3](#_Toc683750777)

[Описание задачи 4](#_Toc1941188577)

[Основные шаги выполнения. EDA и подготовка данных 5](#_Toc1619188049)

[Основные шаги выполнения. Разделение на выборки 9](#_Toc2106429262)

[Основные шаги выполнения. Построение модели 10](#_Toc1896111161)

[Основные шаги выполнения. Лучшая модель 11](#_Toc1406631840)

[Основные шаги выполнения. Интерпретация лучшей модели 13](#_Toc1427055974)

[Заключение 14](#_Toc21489756)

[Список использованных источников и программ 15](#_Toc473770367)

[Приложение 16](#_Toc1258544963)

# 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.

Фича - это переменная (столбец в таблице), которая описывает отдельную характеристику объекта.

# 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.

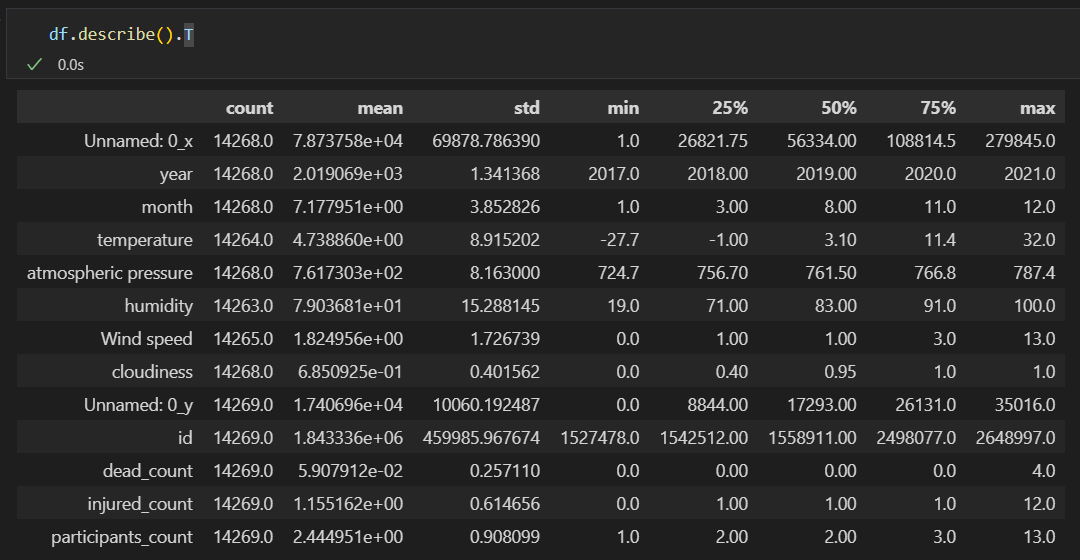
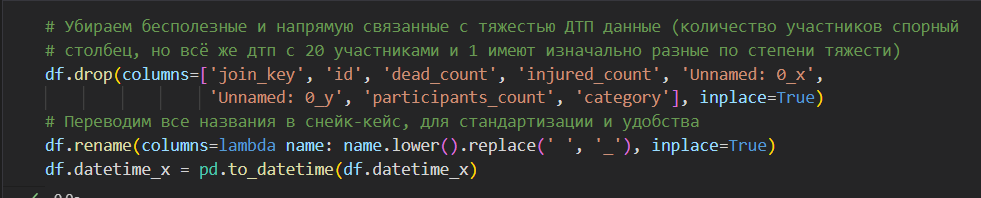


Рис. 1



Рис. 2

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



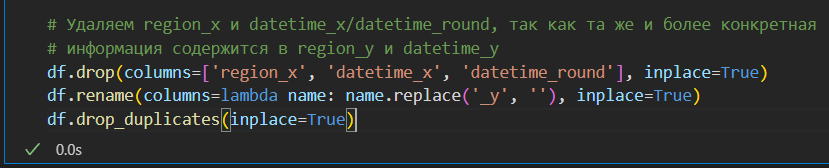
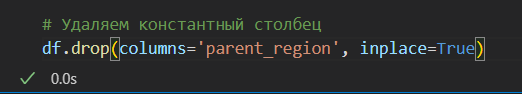


Рис. 3-4

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

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



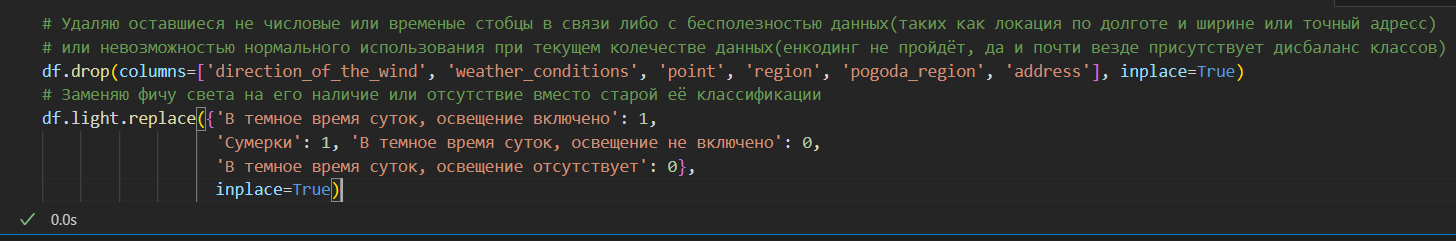


Рис. 5-6

Добавляю новые фичи, которые требуются по заданию.

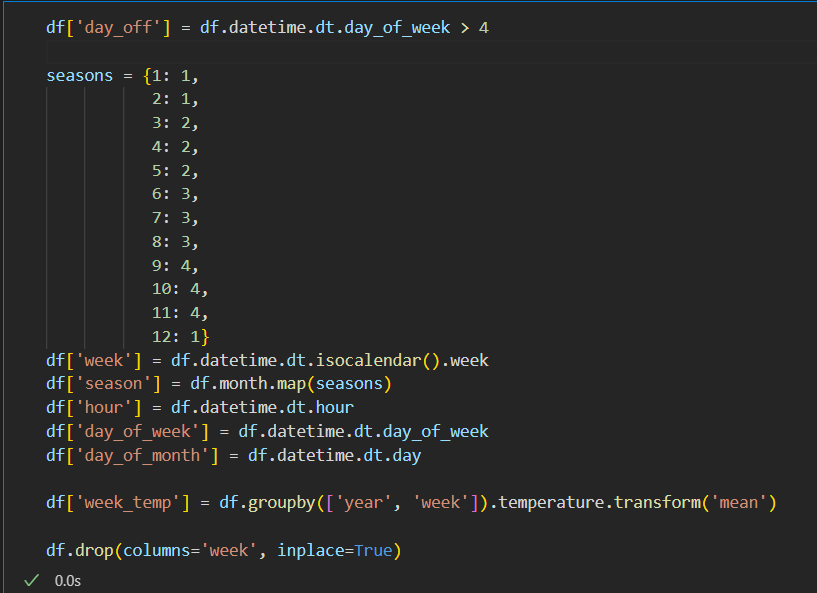


Рис. 7

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

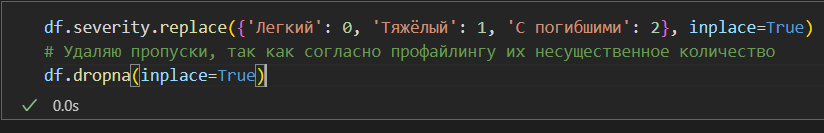


Рис. 8

# Основные шаги выполнения. Разделение на выборки

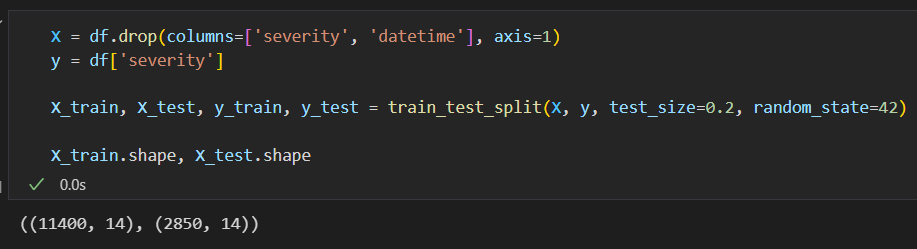


Рис. 9

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.

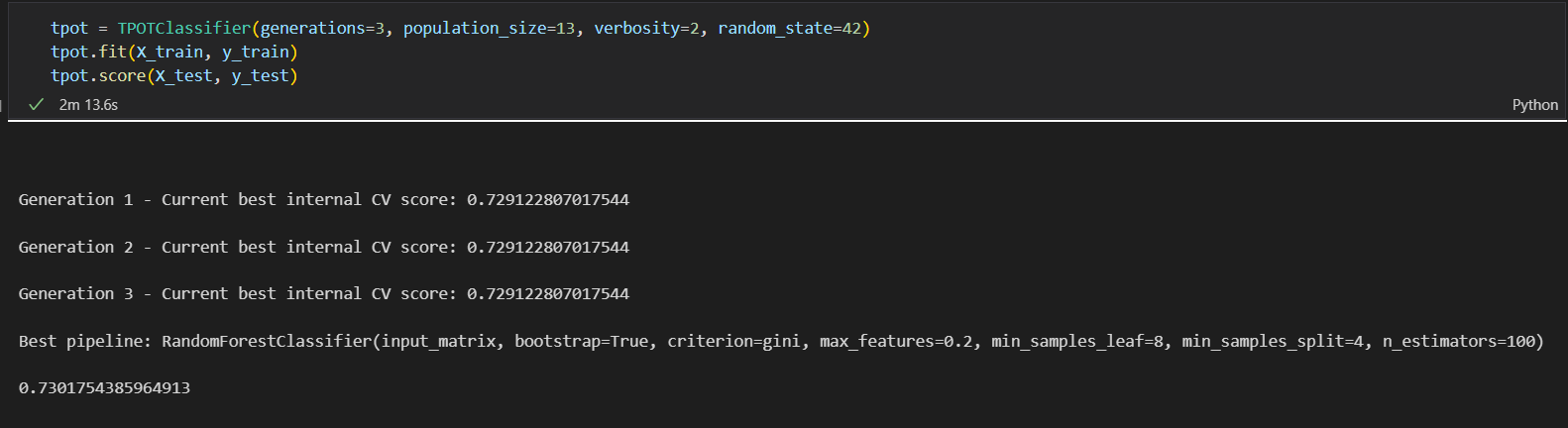


Рис. 10

В результате TPOT выбрал RandomForestClassifier, с итоговым результатом:

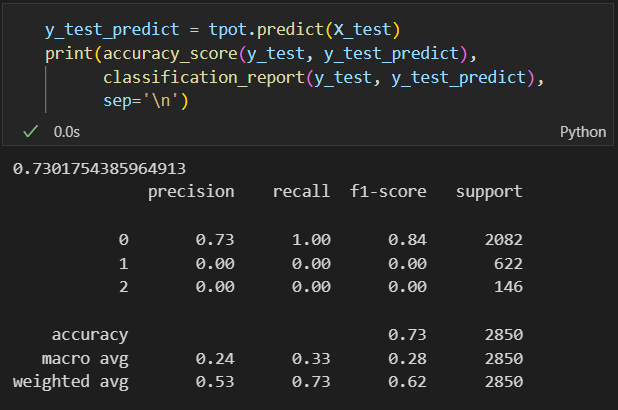


Рис. 11

# 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.

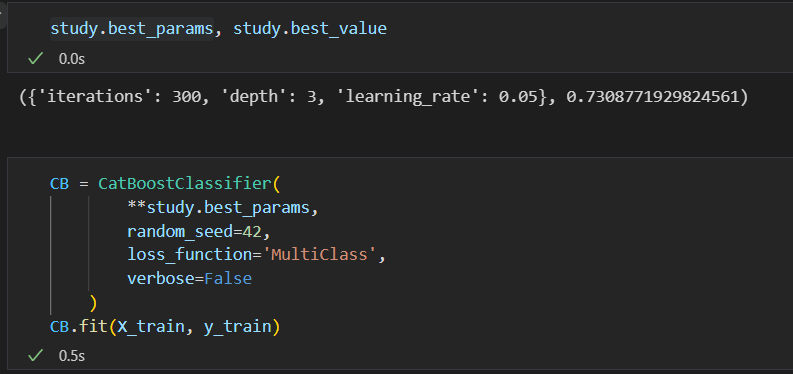


Рис. 12

Финальная модель категориального бустинга имеет параметры: количество итераций – 300, максимальная глубина дерева – 3, скорость обучения – 0.05 и имеет следующие результаты по метрикам:

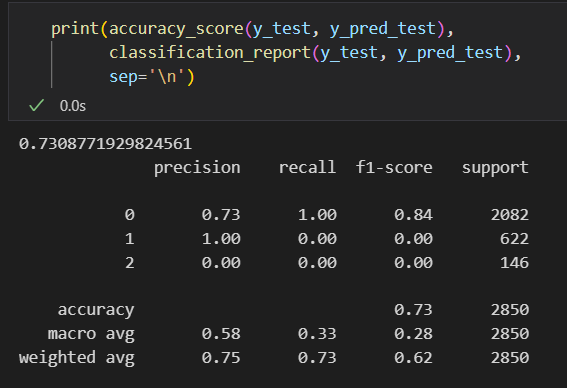


Рис.13

Итоговый результат лучшей модели по метрикам лишь немного лучше базовой модели, что вызвано в первую очередь скудностью данных, а также их качеством: огромный дисбаланс классов целевой метки, который, при попытке устранения, лишь ухудшает результаты моделей. Так же сыграло роль то, что фактор погодные условия, который мог бы сыграть ключевую роль, было практически невозможно использовать, так как он категориальный и не поддаётся нормальному кодированию, а использование one-hot енкодинга было нецелесообразно, так как при таком количестве строк данных, количество фичей стало бы слишком большим, а самое главное — это наличие огромного дисбаланса классов по этой фиче.

# Основные шаги выполнения. Интерпретация лучшей модели

Для интерпретации модели было решено использовать библиотеки shap и lime.

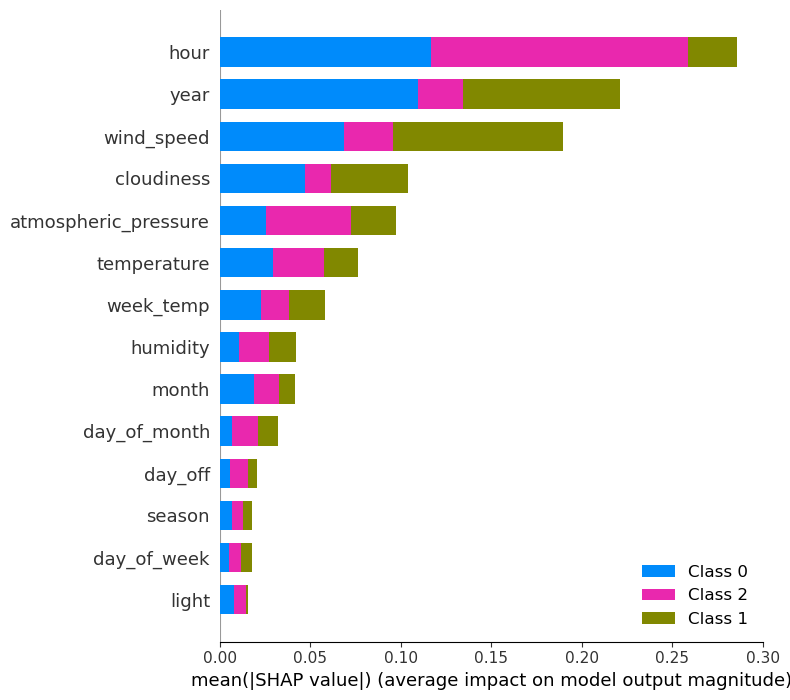


Рис. 14

Важнейшими факторами для модели оказались время, год и скорость ветра. Если первое и последнее ясны, чем позже и сильнее ветер - тем меньше света и сложнее управлять машиной, особенно при выезде с закрытых участков дороги на обветриваемые. А вот влияние параметра года достаточно странно, однако может просто указывать на улучшение условий для вождения (так как он больше способствует выбору более лёгких последствий ДТП, нежели летальных исходов). Так же можно заметить влияние облачности на тяжесть ДТП, что может быть так же связано с уменьшением количества света, так же, как и у параметра часов - а следовательно и видимости. Однако сам параметр наличия света не имеет практически никакого влияния, что может быть связанно с тем, как он считается.

Такие погодные данные как атмосферное давление, влажность и температура имеют среднее влияние, что означает, что погода может влиять на тяжесть ДТП, однако она не является ключевым фактором.

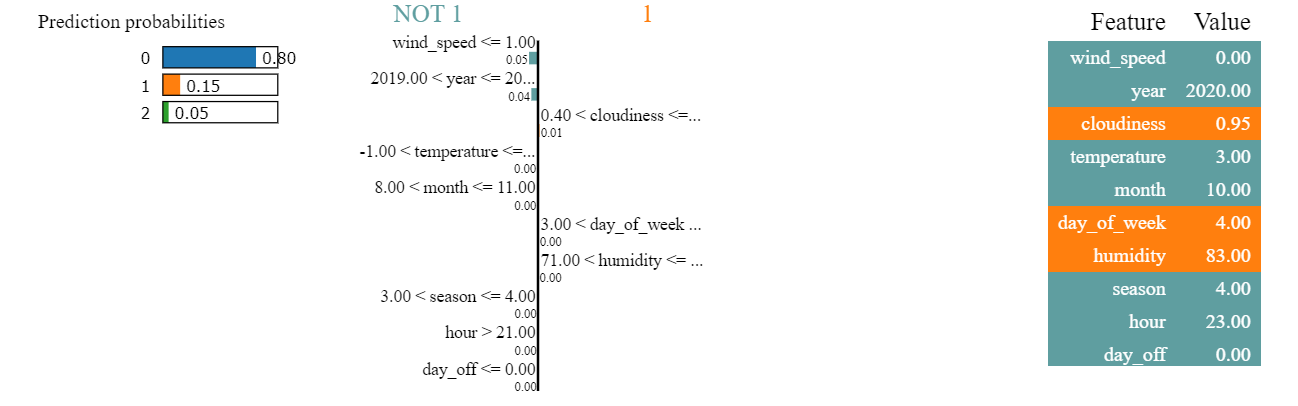


Рис. 15

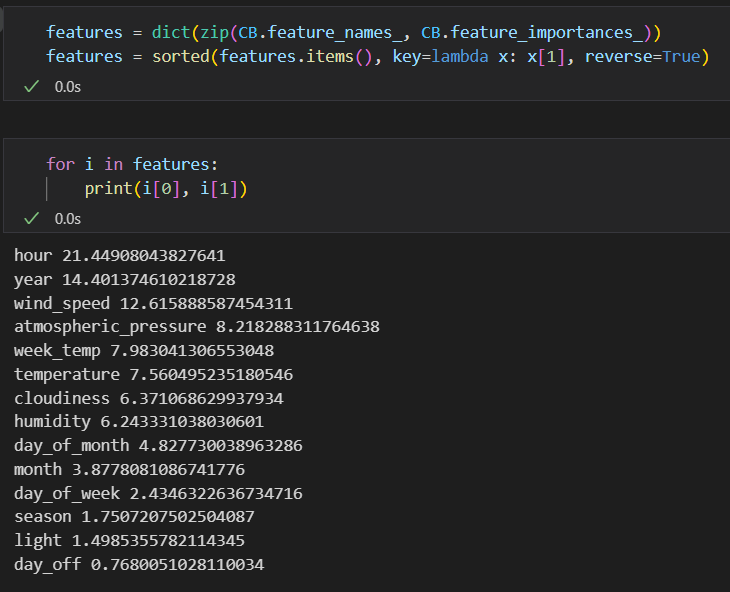


Рис. 16

# Заключение

В заключении можно сказать, что поставленные цели были достигнуты условно, ввиду скудности выборки и фичей, малосвязанных с целевой меткой. Наилучший полученный результат от простого гадания с ответом, что ДТП всегда имеет лёгкие последствия, отличается лишь на десятые процентов и всё так же равно примерно 70%, что означает, что данные изначально не были приспособлены к обучению по ним.

Так же хотелось бы добавить, что смысла прогнозировать тяжесть ДТП с условием, что оно уже произошло, фактически бессмысленно, так как тяжесть становится ясна в тот же момент, что оно и произошло.

Однако данное и подобные исследования могут помочь понять при каких условиях ДТП имеет наихудшие последствия.

Это может пригодиться для регулирования ПДД в определённые погодные условия, ради снижения как рисков самого ДТП, так и его последствий.

# Список использованных источников и программ

1. Python (Anaconda3) с его библиотеками для анализа данных и машинного обучения: pandas, ydata-profiling, scikit-learn, tpot, optuna, shap, lime
2. Редактор кода Visual Studio Code
3. Платформа итеративного программирования Jupiter Notebook

# Приложение

1. Ссылка на отчёт ydata-profiling: [Pandas Profiling Report (althgamer.ru)](https://althgamer.ru/report.html)
2. Репозиторий github с кодом и копией этого документа: [GaganovAlexander/ML\_course\_paper (github.com)](https://github.com/GaganovAlexander/ML_course_paper)
3. Альтернатива для скачки файла с кодом: [ml\_course\_paper.ipynb (althgamer.ru)](https://althgamer.ru/storage/ml_course_paper.ipynb)