**Слайд 1**

Тема курсового проекта – Численный анализ данных о разрушающей силе торнадо

**Слайд 2 -** Постановка задачи.

Торнадо, как стихийное бедствие, представляет серьезную угрозу для жизни и имущества людей. Проблема увеличения случаев возникновения торнадо становится все более значимой ввиду изменений климата. В данной работе акцент сделан на применении численных методов статистического анализа для прогнозирования разрушающей силы торнадо по характеристикам торнадо. Цель работы - изучение и сравнение актуальных численных методов статистического анализа данных, а также создание на их основе предсказывающих моделей для прогнозирования силы торнадо. В результате работы предполагается выявить наиболее подходящую и точную модель для данной задачи и реализовать приложение, предсказывающее разрушающую силу торнадо по характеристикам торнадо.

**Слайд 3 -** Актуальность

Актуальность данной работы объясняется тенденцией к увеличению случаев возникновения торнадо. На рисунке представлены распределение торнадо по годам от 1950 до 2022 и диаграмма рассеяния с линией тренда для визуализации общей тенденции.

**Слайд 4 –** Шкала Фудзиты

Как же оценивается разрушающая сила торнадо? Она же магнитуда. Используется Расширенная шкала Фудзиты (EF-шкала), которая отвечает за классификацию торнадо по степени серьезности на основе ущерба, который они могут причинить. Она имеет шесть категорий интенсивности силы торнадо от нуля до пяти, представляющих возрастающие степени ущерба

**Слайд 5 -** Предобработка данных.

В работе был использован датасет, включающий различные параметры данного природного явления, зарегистрированного в США в различных штатах в период с 1950 по 2022 год. Датасет предоставляет информацию о точных датах, размерах и геопозиции торнадо, а также данные о нанесенном ущербе, количестве жертв и раненных вследствие этого стихийного бедствия в течение семи десятилетий. Всего в наборе данных присутствует 27 колонок, из них были выбраны колонки, которые потенциально могут быть полезны в дальнейшем анализе и иметь зависимости между собой: Год и месяц зарегистрированного торнадо, штат зарегистрированного торнадо, магнитуда (разрушающую силу) торнадо по расширенной шкале Фудзиты, количество травм вследствие торнадо, количество погибших вследствие торнадо, предполагаемая потеря имущества в долларах, длина торнадо в милях, ширина торнадо в ярдах, начальные географичесие долгота и широта торнадо в градусах.

Первым шагом стали тщательный анализ и очистка данных. Что подразумевает обработку пропусков в данных, удаление дубликатов и исправление аномальных значений, чтобы гарантировать точность анализа.

**Слайд 6**

Важным этапом был анализ корреляций для понимания взаимосвязей между различными переменными.

Изначально планировалось использовать коэффициент корреляции Пирсона, однако он обычно используется для оценки линейной взаимосвязи между двумя количественными непрерывными переменными. При работе с категориальными и количественными признаками, которые могут использоваться для анализа торнадо, коэффициент корреляции Пирсона не является подходящим методом для оценки их взаимосвязи.

Для оценки связи между категориальным признаком, таким как разрушающая сила торнадо по EF-шкале, и количественными признаками, такими как длина торнадо, ширина, географические координаты, количество смертей, может быть полезно использовать различные методы визуализации, например боксплоты, гистограммы или сводные таблицы.

И наконец, данные были разделены на обучающую и тестовую выборки в соотношении 75% к 25% с помощью библиотечной функции train\_test\_split.

**Ипсользуемые признаки для предсказания целевого:**

* длина торнадо в милях,
* ширина торнадо в ярдах,
* количество людей, погибших вследствие торнадо,
* географические широта и долгота места происшествия в градусах.

**Слайд 6 –**

В рамках данной курсовой работы по прогнозированию разрушающей силы торнадо был проведен сравнительный анализ нескольких численных методов статистического анализа данных, представляющих собой стохастические модели. Были использованы как библиотечные функции, так и реализованы собственные версии.

**Логистическая регрессия**

Логистическая регрессия - это статистический метод, используемый для предсказания вероятности наступления события. Модель преобразует линейную комбинацию входных переменных в вероятность с использованием сигмоидной функции, которая ограничивает выходное значение в диапазоне от 0 до 1.

**Деревья решений**

Метод деревьев решений в машинном обучении представляет собой структуру, где каждый узел символизирует тест на атрибут, каждая ветвь - результат теста, а каждый лист - классификационную метку. Алгоритм работает, разделяя набор данных на подмножества на основе атрибутов, выбирая атрибуты на каждом этапе с использованием такого критерия, как примесь(индекс) Джини, для достижения наиболее 'чистых' подразделений.

**kNN**

Метод k-ближайших соседей (kNN) — это алгоритм, используемый для классификации и регресии. Он работает, определяя "k" ближайших обучающих образцов в пространстве признаков к новому образцу и принимая решение на основе большинства голосов соседей. Метод основан на идее анализа близости между объектом и уже классифицированными точками в обучающем наборе данных, что и обеспечивает эффективное прогнозирование категории для новых данных

**Линейный дискриминантный анализ**

Линейный дискриминантный анализ (ЛДА) - это статистический метод, используемый для моделирования различий или разделения между двумя или более классами или категориями. ЛДА стремится найти оптимальную разделяющую гиперплоскость, которая максимизирует расстояние между средними значениями классов и минимизирует разброс внутри классов. На основе этой гиперплоскости происходит проекция изначальных данных на новое пространство, в котором классы становятся более линейно разделимыми.

**Наивный байесовский классификатор**

Наивный байесовский классификатор – это метод классификации, основанный на применении теоремы Байеса с допущением о независимости признаков. Наивный байесовский классификатор использует теорему Байеса для вычисления вероятностей принадлежности объекта к каждому классу, опираясь на условные вероятности признаков. Ключевая особенность заключается в предположении о наивной (строгой) независимости между признаками, что означает, что вероятности признаков условно независимы при условии класса. Хотя это предположение на практике редко выполняется, наивный байесовский классификатор все равно часто дает хорошие результаты.

**Слайд 8 - Сравнение**

В данном исследовании, хотя все модели показали неплохие результаты,

Сравнительный анализ точности предсказаний рассмотренных реализаций выбранных численных методов позволил сделать следующие вывода:

- лучшие результаты на тестовой выборке показало дерево решений с заданной максимальной глубиной.

- Несмотря на то, что алгоритм стандартного дерева решений показал отличные результаты на тренировочном наборе данных, обобщающая способность данной модели оказалась хуже.

- К сожалению, собственная реализация метода классификации оказалась хуже и по предсказывающей, и по обобщающей способности. Поэтому выбор пал в пользу библиотечной реализации.

**Слайд 9 - Заключение**

В результате данной курсовой работы было успешно создано приложение для прогнозирования силы торнадо по нескольким характеристикам.

В заключении, данное исследование подчеркивает важность численных методов в анализе метеорологических данных. Были рассмотрены разнообразные подходы, включая логистическую регрессию, деревья решений и k-ближайших соседей, Линейный дискриминантный анализ, Наивный байесовский классификатор, выявив их значимость в улучшении точности прогнозирования и анализа. Это исследование демонстрирует, как численные методы могут способствовать более глубокому пониманию и эффективному лечению медицинских состояний, предоставляя ценные знания для клинического применения.