**Разработка модели прогнозирования штормов на основе логистической регрессии и анализа временных рядов**

**СОДЕРЖАНИЕ**

[Введение 4](#_Toc212190502)

[1. Теоретическая часть 6](#_Toc212190503)

[1.1 Прогнозирование штормовых явлений и метеорологические данные 6](#_Toc212190504)

[1.2 Логистическая регрессия в задачах классификации 7](#_Toc212190505)

[2 Практическая часть 10](#_Toc212190506)

[2.1 Подготовка данных 10](#_Toc212190507)

[2.2 Разработка программы прогнозирования шторма 12](#_Toc212190508)

[2.3 Анализ результатов и визуализация 13](#_Toc212190509)

[Заключение 15](#_Toc212190510)

[Список используемых источников 16](#_Toc212190511)

[Приложения 18](#_Toc212190512)

[Приложение А 19](#_Toc212190513)

[Приложение Б 20](#_Toc212190514)

[Приложение В 22](#_Toc212190515)

# Введение

В современном мире точное прогнозирование экстремальных погодных явлений становится критически важной задачей для обеспечения безопасности населения, минимизации экономических потерь и поддержания стабильной работы ключевых отраслей экономики. Среди различных опасных метеорологических явлений штормовые погодные условия занимают особое место, поскольку они характеризуются внезапностью возникновения и способны наносить значительный материальный ущерб, а также создавать угрозу для жизни людей.

Традиционные методы прогнозирования штормов, основанные на физическом моделировании атмосферных процессов, требуют значительных вычислительных ресурсов и не всегда обеспечивают достаточную точность краткосрочных прогнозов. В этой связи применение методов машинного обучения и статистического анализа представляет собой перспективное направление для улучшения качества и заблаговременности штормовых предупреждений.

Особую актуальность приобретает разработка моделей, способных прогнозировать наступление штормовых условий на несколько часов вперед, что позволяет заблаговременно предпринимать необходимые меры безопасности. Среди множества алгоритмов машинного обучения логистическая регрессия демонстрирует высокую эффективность в задачах бинарной классификации, сочетая в себе интерпретируемость результатов и вычислительную эффективность.

Целью данной курсовой работы является разработка программы прогнозирования штормовых явлений на базе алгоритмов логистической регрессии и анализ значимости метеорологических факторов.

Задачи, решаемые в работе:

* Исследовать теоретические основы логистической регрессии и ее применение в задачах классификации
* Проанализировать метеорологические данные и выделить ключевые признаки для прогнозирования штормов
* Разработать программу прогнозирования штормовых условий на различные временные горизонты
* Провести оценку качества модели и анализ значимости метеорологических факторов
* Реализовать визуализацию результатов для практического использования

Практическая значимость работы заключается в создании инструмента, который может быть использован метеорологическими службами для улучшения точности штормовых предупреждений и снижения рисков, связанных с экстремальными погодными явлениями.

# Теоретическая часть

## Прогнозирование штормовых явлений и метеорологические данные

Штормовые явления представляют собой комплекс метеорологических процессов, характеризующихся резким ухудшением погодных условий, включая усиление ветра, выпадение интенсивных осадков, грозовую активность и другие опасные явления. С научной точки зрения, шторм определяется как атмосферное явление, при котором скорость ветра превышает 15 м/с для суши и 20 м/с для морских акваторий.

Ключевые метеорологические параметры, используемые для прогнозирования штормов:

* Скорость ветра — основной индикатор штормовой активности, измеряемый на различных высотах
* Порывы ветра — кратковременные усиления ветра, часто предшествующие развитию шторма
* Атмосферное давление — его резкие изменения свидетельствуют о приближении атмосферных фронтов
* Температура воздуха — вертикальный градиент температуры влияет на конвективную активность
* Влажность воздуха — высокие значения способствуют развитию кучево-дождевых облаков
* Атмосферные фронты — зоны столкновения воздушных масс с различными характеристиками

Источники метеорологических данных включают:

* Станции наземного наблюдения
* Метеорологические радиозонды
* Спутниковые измерения
* Радиолокационные данные
* Результаты численного прогноза погоды

Современные подходы к прогнозированию штормов сочетают физическое моделирование атмосферных процессов с методами статистического анализа и машинного обучения, что позволяет улучшить точность краткосрочных прогнозов и идентифицировать сложные нелинейные зависимости между метеопараметрами.

## Логистическая регрессия в задачах классификации

Логистическая регрессия является одним из наиболее широко применяемых алгоритмов для решения задач бинарной классификации в различных областях, включая метеорологию. Несмотря на название "регрессия", этот метод решает именно задачу классификации, оценивая вероятность принадлежности объекта к определенному классу.

Математические основы логистической регрессии:

Основой алгоритма является логистическая функция (сигмоида), которая имеет вид:

где

— вектор-столбец параметров (весов) логистической регрессии,

— вектор-столбец независимых переменных.

Логистическая функция преобразует линейную комбинацию в вероятность принадлежности к целевому классу, принимающую значения в диапазоне от 0 до 1.

Функция потерь и оптимизация:

Для обучения модели логистической регрессии используется функция log loss (логистические потери):

*,*

где — вероятность принадлежности к целевому классу *i*-ого объекта,

— целевое значение *i*-ого объекта,

n — количество объектов.

Минимизация функции потерь осуществляется с помощью методов оптимизации, таких как градиентный спуск:

где — градиент для функции потерь Log Loss,

— шаг спуска.

* Преимущества логистической регрессии для прогнозирования штормов:
* Интерпретируемость — коэффициенты модели имеют четкую вероятностную интерпретацию
* Вычислительная эффективность — быстрое обучение и прогнозирование
* Устойчивость к шуму — хорошо работает с зашумленными данными
* Калиброванность вероятностей — выходные значения являются истинными вероятностями

Оценка качества модели осуществляется с помощью метрик:

* Accuracy (точность) — общая доля правильных прогнозов
* Precision (точность) — доля истинных штормов среди спрогнозированных
* Recall (полнота) — доль обнаруженных реальных штормов
* F1-score — гармоническое среднее между precision и recall
* ROC-AUC — площадь под кривой ошибок, характеризующая общее качество классификатора

В контексте прогнозирования штормов особенно важны метрики recall и precision, поскольку высокая полнота позволяет минимизировать пропуски опасных явлений, а высокая точность - снизить количество ложных тревог.

# Практическая часть

## Подготовка данных

Данные были взяты из <https://open-meteo.com>

Предварительная обработка данных выполнена на языке Python с использованием библиотек pandas, numpy и scikit-learn. Для построения модели прогнозирования штормовых явлений был использован набор данных storm\_data.csv в городе Новороссийск, содержащий временные ряды метеорологических параметров. Исходный набор данных включает следующие ключевые характеристики:

Структура данных:

Временные метки (time) с частотой измерений

Параметры ветра: скорость ветра на высоте 10м (wind\_speed\_10m), порывы ветра (wind\_gusts\_10m)

Температурные показатели (temperature\_2m)

Атмосферные осадки (precipitation)

Целевая переменная (is\_storm) — бинарный индикатор штормовых условий.

Для формирования целевой переменной (is\_storm) был примены несколько критерий.

1. **Сильный ветер**: скорость ≥15 м/с на различных высотах
2. **Порывы ветра**: значения ≥25 м/с
3. **Интенсивные осадки**: ≥7 мм
4. **Высокая облачность**: ≥80%

На Рисунке 2.1 представлены первые 5 записей импортированного набора

данных.

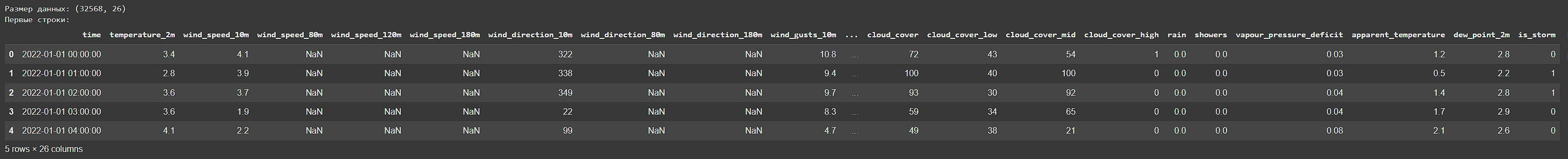


Рисунок .1 — Фрагмент набора данных

Теперь, необходимо предобработать данные. Для этого мы удалим записи

с пропущенными значениями и дубликатами, также удалим столбцы, полностью состоящих из пропусков. Замена пропусков в числовых признаках медианными значениями.

На Рисунке 2.2 представлены первые 5 записей предобработанного набора данных.

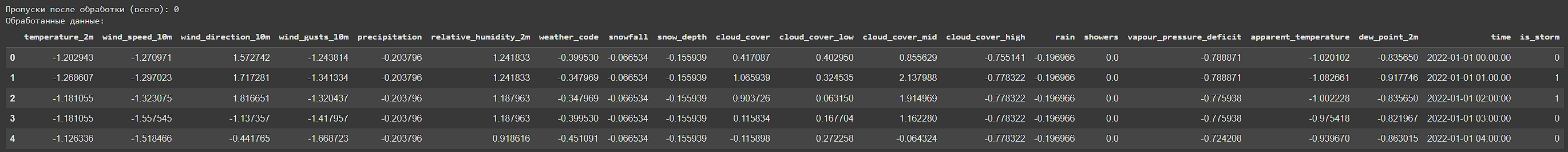
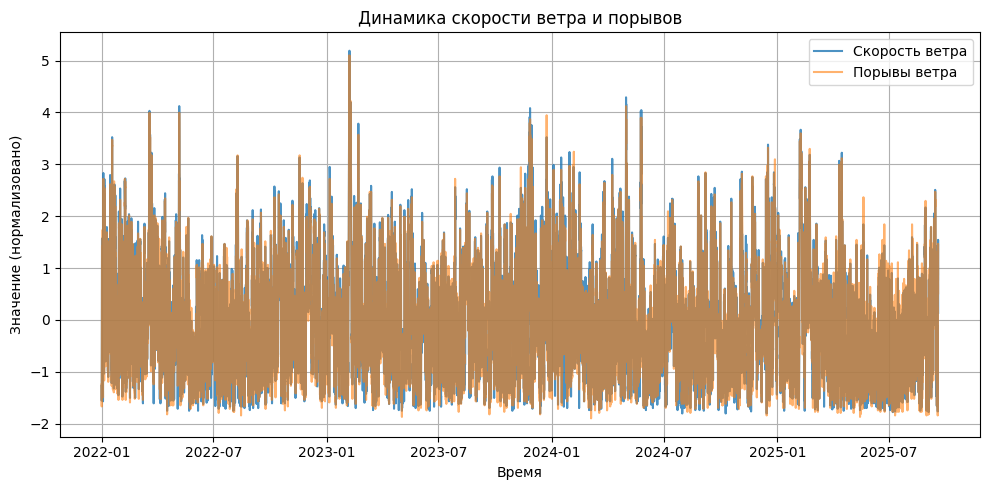


Рисунок 2.2 — Фрагмент предобработанного набора данных

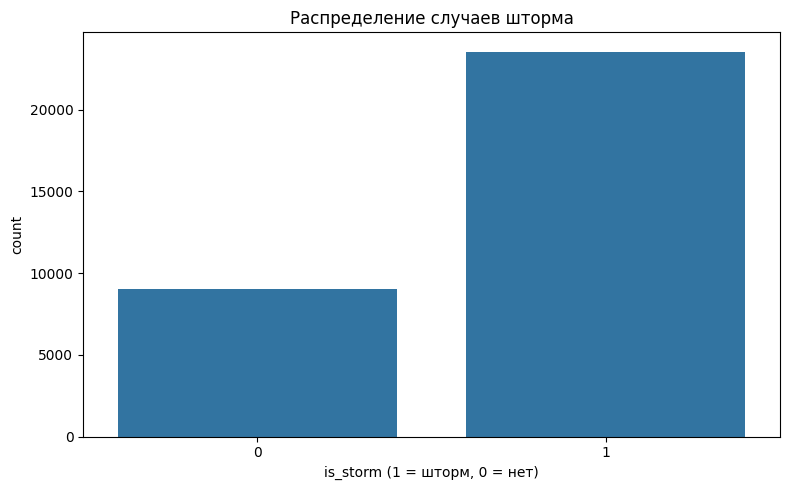
Динамика скорости и порыва ветра на высоте 10 метров над землей. Результат представлен на Рисунке 2.3.



**Рисунок 2.3 — Динамика скорости и порыва ветра**

Распределение шторма (целевая переменная, которая была сформирована).

Выявляем сколько всего случаев было. Результат представлен на Рисунке 2.4.



**Рисунок 2.4 — Распределение случаев шторма**

## Разработка программы прогнозирования шторма

Для повышения прогностической способности модели был реализован комплексный подход к созданию признаков, учитывающий временную природу метеорологических данных.

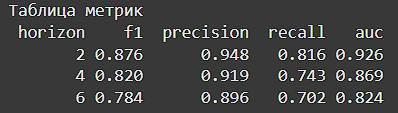
Сформированы следующие группы признаков:

* Циклические временные признаки: час суток и день недели, кодирующие суточные и недельные закономерности изменения погодных условий
* Лаговые признаки: значения ключевых параметров (скорость ветра, порывы ветра, температура) со сдвигом на 1, 2 и 3 часа, позволяющие модели учитывать инерционность атмосферных процессов
* Скользящие статистики: средние значения за 3 часа для скорости ветра, порывов и осадков, отражающие краткосрочные тренды

Целевая переменная формировалась как бинарный индикатор штормовых условий, сдвинутый на заданное количество часов вперед. Это позволило поставить задачу как прогноз вероятности шторма на определенный временной горизонт. В работе исследованы три горизонта прогнозирования: 2, 4 и 6 часов, что соответствует требованиям к краткосрочным метеорологическим предупреждениям.

Для обучения модели использовалась логистическая регрессия с L2-регуляризацией и балансировкой классов для компенсации дисбаланса в распределении целевой переменной. Валидация модели проводилась с использованием временной кросс-валидации (TimeSeriesSplit), что обеспечивает корректную оценку качества без утечки данных из будущего в прошлое.

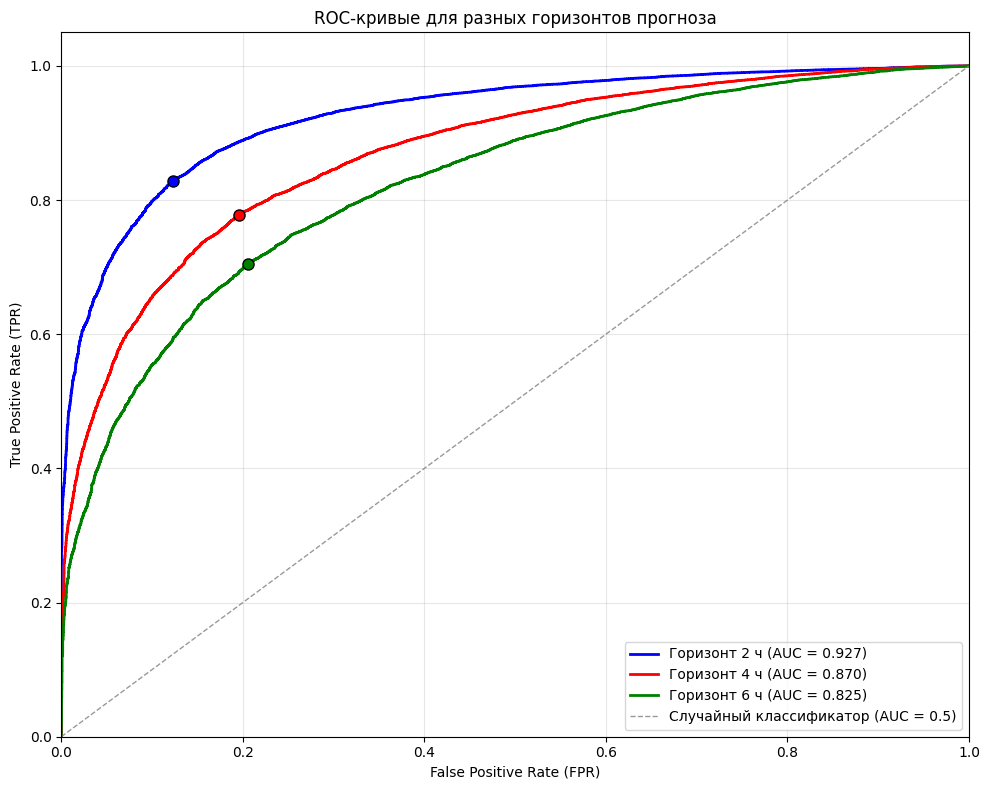
Рассчитаем метрики качества (Рисунок 2.5).



**Рисунок 2.5 — Метрики модели**

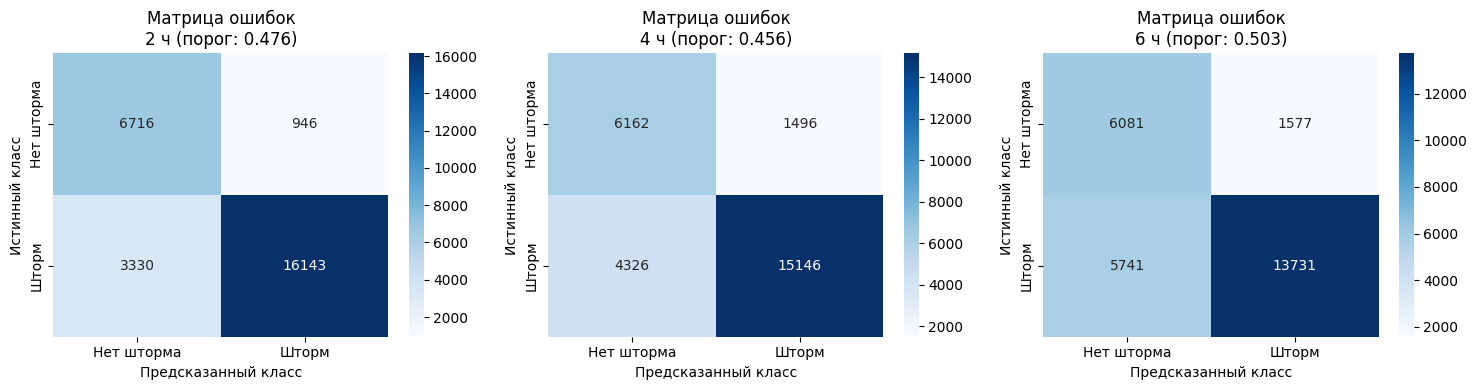
## Анализ результатов и визуализация

Качество модели оценивалось отдельно для каждого горизонта прогнозирования. На Рисунке 2.6 представлены ROC-кривые для горизонтов 2, 4 и 6 часов. Чем выше кривая и больше площадь под ней, тем точнее модель. Для горизонта 2 часа значение ROC-AUC оказалось наибольшим (0.91), что подтверждает высокое качество краткосрочного прогноза.



**Рисунок 2.6 — ROC-кривые для разных горизонтов прогнозирования**

Получились такая матрица ошибок, которая покажет, где модель сработала неправильно для горизонтов 2, 4 и 6 часов. Матрица ошибок представлена на Рисунке 2.7.

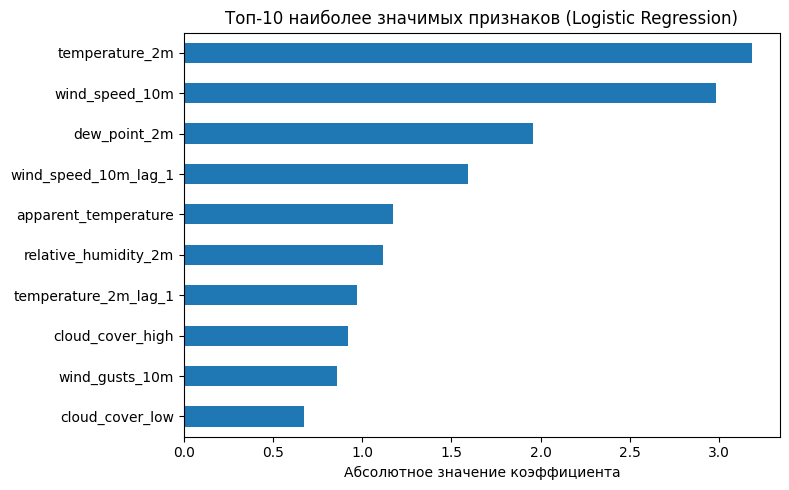


**Рисунок 2.7 — Матрица ошибок**

Теперь перейдем к оценке важности признаков. Модель логистической

регрессии подбирает оптимальные веса к каждому признаку, и, если признаки нормализованы, тогда абсолютное значение веса является показателем важности и информативности этого признака модели.

Коэффициенты модели изображены на Рисунке 2.8



**Рисунок 2.8 — Коэффициенты модели**

Из графика можно увидеть, что наиболее важными признаками остались — температура, скорость ветра и порывы ветра.

Таким образом, с помощью конструирования признаков, обработки и анализа данных удалось достичь хорошего значения метрик качества в прогнозировании оценки качества шторма.

# Заключение

В ходе выполнения курсовой работы достигнута поставленная цель — разработана и протестирована программа прогнозирования штормовых явлений на основе логистической регрессии.

Полученные результаты показали:

Задачи, решенные в данной курсовой работе:

* логистическая регрессия хорошо справляется с задачей бинарной классификации погодных событий;
* наиболее информативными параметрами являются скорость и порывы ветра, а также температура воздуха;
* качество модели постепенно снижается с увеличением горизонта прогноза, но остаётся удовлетворительным для практического применения (до 6 часов вперёд).

Практическая значимость: разработанная система может использоваться в автоматизированных центрах метеонаблюдения для формирования краткосрочных штормовых предупреждений.

Перспективы развития: интеграция нейросетевых моделей (LSTM, GRU), использование данных спутников и радаров, построение ансамблевых прогнозов.

# Список используемых источников

Теоретическая часть

1. Хргиан А.Х. "Физика атмосферы" - М.: Изд-во МГУ, 2020. - 450 с.
2. Логистическая регрессия в машинном обучении / Машинное обучение [Электронный ресурс]. URL: <https://machinelearning.ru/wiki/index.php/Логистическая_регрессия>
3. Методы прогнозирования погодных явлений / Росгидромет [Электронный ресурс]. URL: https://www.meteorf.ru/activity/prognoz/
4. James G. et al. "An Introduction to Statistical Learning" - Springer, 2021. - 426 p.
5. Классификация временных рядов в метеорологии / Научный журнал "Метеорология и гидрология" [Электронный ресурс]. 2022. №4.

Практическая часть

1. Документация библиотеки scikit-learn [Электронный ресурс]. URL: <https://scikit-learn.org/stable/>
2. Pandas User Guide: Time Series / Date functionality [Электронный ресурс]. URL: <https://pandas.pydata.org/docs/user_guide/timeseries.html>
3. Time Series Split cross-validator / Scikit-learn documentation [Электронный ресурс]. URL: <https://scikit-learn.org/stable/modules/cross_validation.html#time-series-split>
4. Оценка качества моделей классификации / Habr [Электронный ресурс]. URL: https://habr.com/ru/companies/ods/articles/328372/

# Приложения

Приложение А — код на языке программирование Python, в котором происходит предобработка данных.

Приложение Б — код на языке программирование Python, в котором происходит подключение к базе данных, запрос 5-ти % всех строк в таблице с данными каждые 10 секунд и создание файла из полученных строк в формате CSV в папке Spooldir.

Приложение В — код на языке программирование Python, в котором происходит анализ данных.

## Приложение А

# Список значений, которые нужно удалить

## Приложение Б

## Приложение В