МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»**

*ИНСТИТУТ* ИТАСУ

*КАФЕДРА АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ*

*НАПРАВЛЕНИЕ* 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

**Практическое занятие №1**

**По дисциплине**: Технологии интеллектуального анализа данных

**На тему:** Метод деревьев решений для задач бинарной и мультикласссовой классификации

**Студент: Костромин Д.А.**

**Группа: МИВТ 18-5-7**

**Преподаватель: Фомичева О.Е.**

**Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Дата защиты: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Москва, 2019

**Оглавление**

[**Введение** 3](#_Toc24078601)

[**1. Простая мультиклассовая классификация на примере классификации сортов ириса** 4](#_Toc24078602)

[**2. Классификация и прогнозирование погодных явлений на примере исторических данных о погоде в Остине.** 8](#_Toc24078603)

[**2.1 Предварительная обработка данных** 8](#_Toc24078604)

[**2.2 бинарная классификация погодных явлений** 11](#_Toc24078605)

[**2.3 Мультиклассовая классификация погодных явлений** 13](#_Toc24078606)

[**Выводы** 16](#_Toc24078607)

[**Список источников** 17](#_Toc24078608)

# **Введение**

Цель работы изучение и применение простых моделей машинного обучения (ML) с учителем, основанных на алгоритме дерева решений для задач многоклассовой классификации на примере классификации сортов ириса и определения типа осадков по историческим данным погоды в Остине.

Деревья решений – модель машинного обучения с учителем, широко применяемая для задач как классификации, так и регрессии. Данный метод основывается на построении иерархии правил “ЕСЛИ… ТО”, по данным из обучающей выборки. Такой подход позволяет легко визуализировать простые модели на основе деревьев решений, что делает их понятными даже для неспециалистов. Также к однозначным достоинствам деревьев решений относят высокую скорость обучения и отсутствие необходимости в масштабировании данных. С другой стороны модели данного типа склонны к переобучению и показывают небольшую обобщающую способность при работе со сложными наборами данных.

При этом для обоих недостатков существует решение. Основным способом борьбы с переобучением деревьев решений является “обрезание” дерева, когда задается фиксированный параметр максимальной глубины иерархии “ЕСЛИ ТО” как гиперпараметр[[1]](#footnote-1) модели (на этапе до обучения) или после обучения осуществляется “пост-обрезка” дерева, когда из иерархии “ЕСЛИ ТО” удаляются лишние, мало информационные узлы.

В свою очередь для повышения обобщающей способности модели на сложных наборах данных вместо одиночных деревьев решений нередко применяются ансамбли деревьев – методы машинного обучения, позволяющие объединить несколько моделей машинного обучения для получения обобщенного результата. В качестве примера ансамбля можно привести модель случайный лес – набор деревьев решений, в котором каждое дерево отличается от других [1].

При выполнении работы будут использованы модели машинного обучения из библиотеки scikit-learn для языка Python.

# **1. Простая мультиклассовая классификация на примере классификации сортов ириса**

Простейший пример применение метода деревьев решений для задач классификации будет рассмотрен на примере классификации сортов ириса. В качестве исходных данных будет использован “классический” набор данных, содержащий сведений о 150 растениях, относящихся к одному из трех сортов ириса: Iris-setosa, Iris-versicolor или Iris-virginica. При условии, что каждый сорт представляют ровно 50 растений [2].

Исходные данные имеют шесть атрибутов:

* Id – уникальный идентификатор конкретного растения
* SepalLengthCm – измеряемая в сантиметрах длинна чашелистика
* SepalWidthCm - измеряемая в сантиметрах ширина чашелистика
* PetalLengthCm – измеряемая в сантиметрах длинна лепестка
* PetalWidthCm – измеряемая в сантиметрах ширина лепестка
* Species – сорт ириса

Загрузим исходные данные и для проверки отобразим первые 5 строк (Рисунок 1)

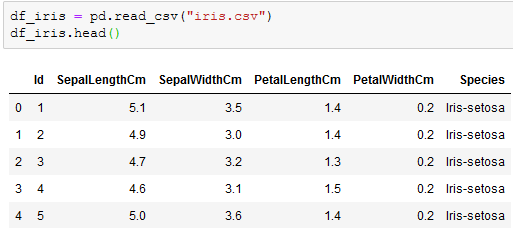


Рисунок 1 – Загрузка исходных данных

Данные успешно загружены. На следующем этапе можно перейти к разделению исходных на тестовую и обучающую выборку и обучение модели. Разбиение на обучающую и тестовую выборку осуществляется с помощью функции train\_test\_split библиотеки sklearn, в результате данные будут случайным образом распределены и 75% исходных данных будут использованы для обучения модели, остальные 25% для тестирования.

В sklearn присутствует реализация деревьев решений – DecisionTreeClassifier, который и будет использован в дальнейшем. В результате тестирования модели получена стопроцентная точность как на тестовой, так и на обучающей выборке, что может быть симптомом переобучения модели (Рисунок 2).

Для получения большей информации о обученной модели можно осуществить ее визуализацию. Как и предполагалось, визуализация демонстрирует наличие избыточно-сложных взаимосвязей между узлами модели, что свидетельствует о переобучении (Рисунок 3).

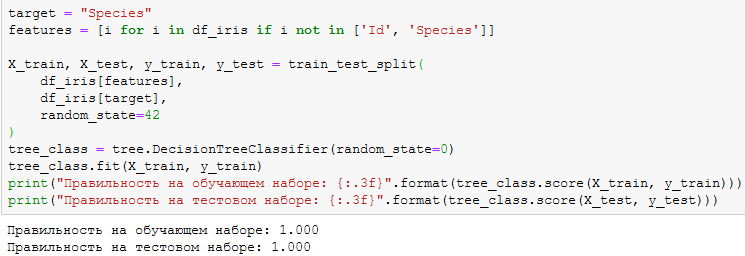


Рисунок 2 – DecisionTreeClassifier с параметрами по умолчанию

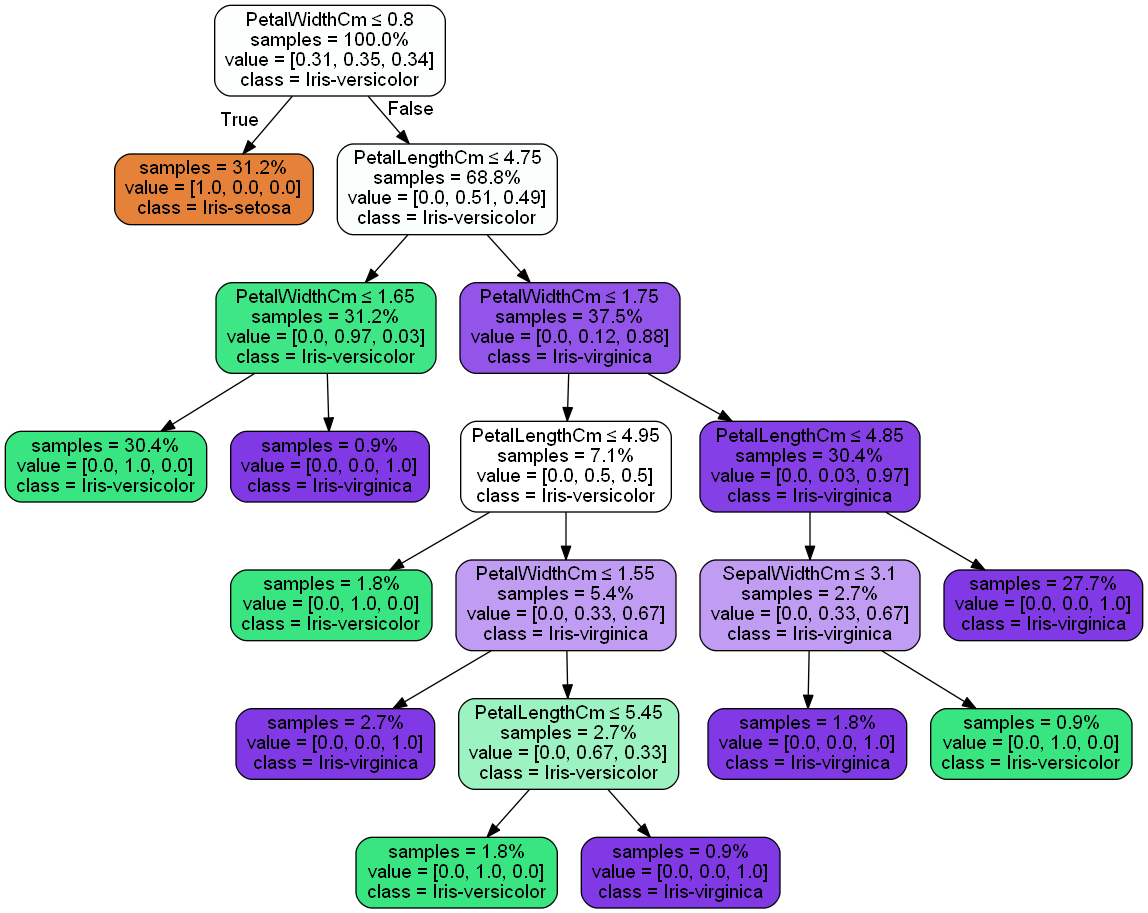


Рисунок 3 Визуальное представление переобученной модели

С целью повышения качества модели можно остановить неконтролируемый “рост” дерева, строго определив максимальную глубину, явно указав параметр max\_depth в конструкторе DecisionTreeClassifier. В данном случае установлено max\_depth=3 (Рисунок 3).

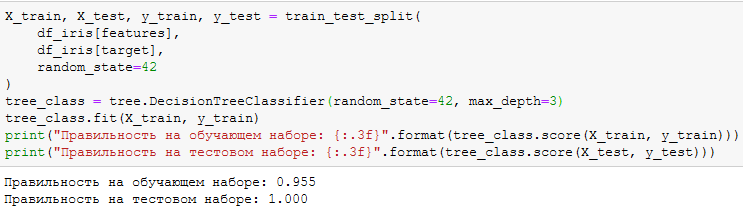


Рисунок 4 – DecisionTreeClassifier с ограничением роста

В результате будет получена меньшая точность на обучающей выборке, но сохранится абсолютная[[2]](#footnote-2) точность на тестовой выборке, при этом будут получены более логичные взаимосвязи между данными (Рисунок 5).

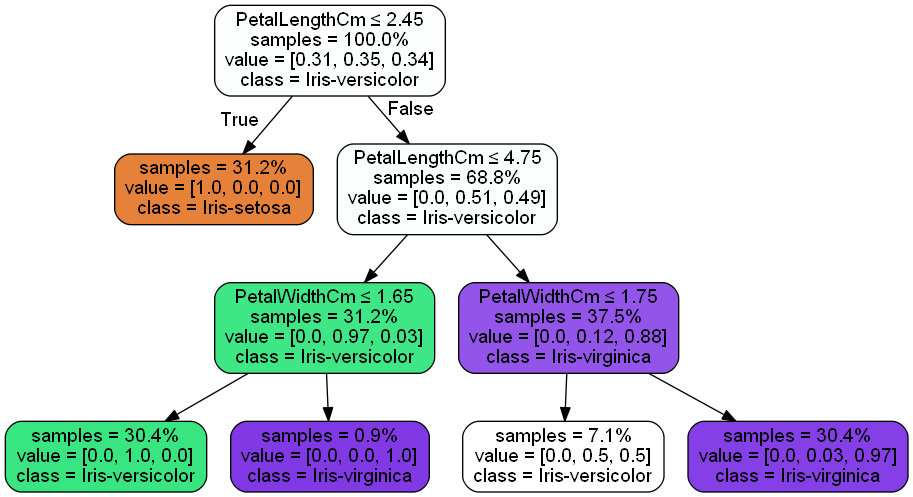


Рисунок 5 Визуальное представление модели с ограниченной глубиной

Классификация ирисов была простым примером и отправной точкой для работы с деревьями решений, теперь можно перейти к большему объему исходных данных и более сложным моделям, требующим большей настройки.

# **2. Классификация и прогнозирование погодных явлений на примере исторических данных о погоде в Остине.**

## **2.1 Предварительная обработка данных**

В этом разделе будет рассматриваться классификация погодных явлений в Остине за 2015-2017 годы. В качестве исходных данных используются исторические сведения о температуре, осадках, скорости ветра и влажности за данный период. Измерения осуществлялись с ежедневной периодичностью.

Исходные данные содержат следующие признаки:

* Date – дата в формате ГГГГ-ММ-ДД.
* TempHighF – Максимальная температура (В Фаренгейтах)
* TempAvgF - Средняя температура (В Фаренгейтах)
* TempLowF - Наименьшая температура (В Фаренгейтах)
* DewPointHighF – Наибольшая температура конденсации (В Фаренгейтах)
* DewPointAvgF - Средняя температура конденсации (В Фаренгейтах)
* DewPointLowF Наименьшая температура конденсации (В Фаренгейтах)
* HumidityHighPercent - Наибольшая влажность за день (%)
* HumidityAvgPercent - Средняя влажность за день (%)
* HumidityLowPercent - Наименьшая влажность за день (%)
* SeaLevelPressureHighInches – Наибольшее давление над уровнем моря (дюймы)
* SeaLevelPressureAvgInches - Среднее давление над уровнем моря (дюймы)
* SeaLevelPressureLowInches - Наименьшее давление над уровнем моря (дюймы)
* VisibilityHighMiles – Максимальная дальность видимости (мили)
* VisibilityAvgMiles – Средняя дальность видимости (мили)
* VisibilityLowMiles – Наименьшая дальность видимости (мили)
* WindHighMPH -Наибольшая скорость ветра (миль/час)
* WindAvgMPH - Средняя скорость ветра (миль/час)
* WindGustMPH - Наибольшая скорость порыва ветра (миль/час)
* PrecipitationSumInches – суммарное количество выпавших осадков (дюймы), “T” при отсутствии.
* Events – Неблагоприятные погодные явления, ‘ ’ (пробельный символ) при отсутствии.

Будет рассмотрена как бинарная, так и мультиклассовая классификация [3].

Первоначально необходимо загрузить исходные данные. Затем осуществляется оценка количества зафиксированных случаев каждого из возможных неблагоприятных погодных явлений, в результате фиксируется большая разрозненность классов в исходных данных, большая часть наблюдений не фиксирована негативных погодных явлений, в то время как “ Fog , Thunderstorm” и “Rain , Snow” были зафиксированы лишь единожды (Рисунок 6).

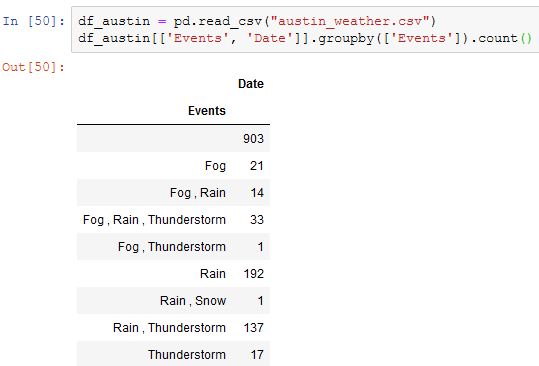


Рисунок 6 Распределение исходных данных по классам

Перед началом обучения моделей необходимо провести предварительное преобразования данных – будут удалены все строки, содержащие пропуски, класс “ ” свидетельствующий об отсутствии неблагоприятных погодных явлений будет переименован в “clear” и обозначенные как “T” значения нулевого уровня осадков будут заменены на 0. Затем будет добавлен новый столбец “Binary”, указывающий были ли негативные природные явления в определенный день. Binary принимает значение 1, если соответствующее значение признака “Events” равно “clear” и 0 во всех остальных случаях. (Рисунок 7).

Остался последний предварительный этап, большая часть исходных данных содержала пропуски и/или была представлена в виде строк, необходимо привести их к числовому представлению (Рисунок 8).

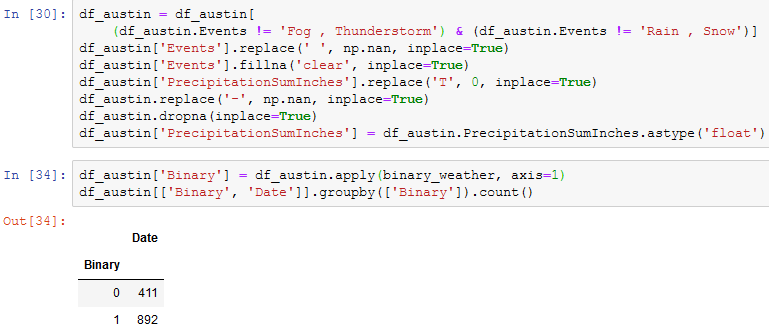


Рисунок 7 Преобразование исходных данных



Рисунок 8 – Приведение исходных данных к числовым типам

После завершения предварительной обработки данных можно приступить к построению моделей машинного обучения. В первую очередь будет рассмотрено применение деревьев решений для задач бинарной классификации, где в качестве целевого столбца будет выступать Binary, соответственно будет прогнозироваться наличие/отсутствие неблагоприятных погодных явлений. Затем будет рассмотрена задача мультиклассовой классификации, когда необходимо будет спрогнозировать возникновение конкретного негативного погодного явления или их отсутствие, соответственно целевым столбцом для мультиклассовой классификации выступит атрибут Events.

В первую очередь будет построена корреляционная матрица, указывающая на зависимости между целевыми атрибутами и остальными признаками (Рисунок 9).

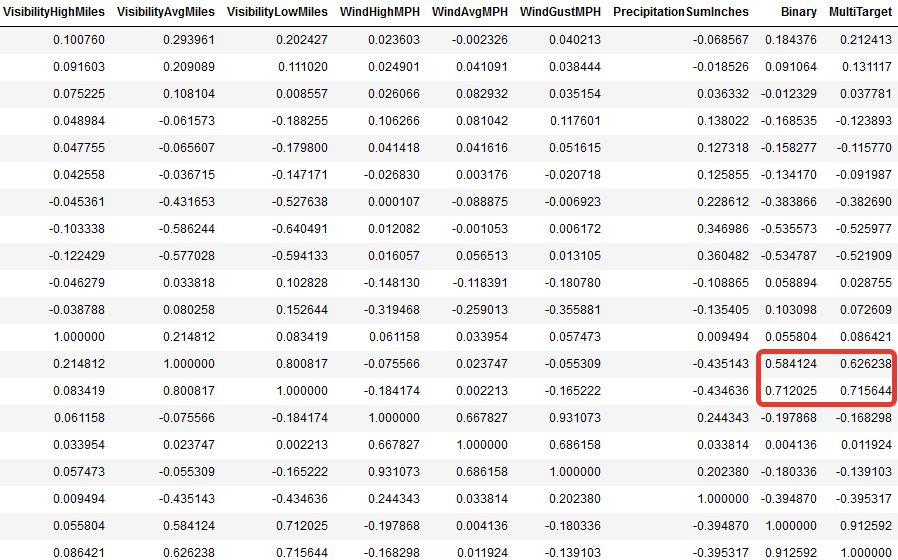


Рисунок 9 – корреляционная матрица (фрагмент)

Корреляционная матрица демонстрирует наибольшую взаимосвязь между целевыми атрибутами и признаками “VisibilityLowMiles” и “VisibilityAvgMiles”, отражающими наименьшую/среднюю видимость соответственно. В дополнении к ним можно выдвинуть гипотезу, что существует взаимосвязь между признаком PrecipitationSumInches, отражающим количество выпавших осадков и наличием/отсутствием прогнозируемых негативных погодных явлений. В результате при обучении и тестировании модели как для бинарной, так и для мультиклассовой классификации будут использованы только вышеперечисленные три признака, все остальные признаки будут проигнорированы.

## **2.2 бинарная классификация погодных явлений**

В первую очередь будет обучена модель для бинарной классификации. В результате точность на обучающей выборке составить 92.9%, точность на тестовой выборке 94,5% (Рисунок 10).

Но, так как разбиение на тестовую и обучающую выборки осуществляется случайным образом возможно ситуация когда высокая точность на тестовых данных является результатом удачного разбиения данных на две выборке, а не следствием хорошей обобщающей способности модели и как следствие качество может существенно снизиться при работе с другими данными. С целью исключения подобных ситуаций применяются дополнительные способы оценки обобщающей способности модели машинного обучения, к числу которых относиться k-блочная проверка.

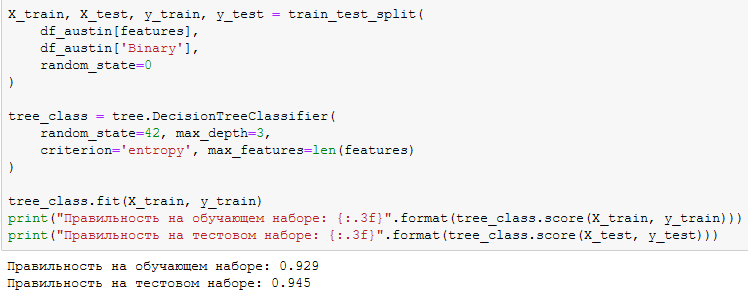


Рисунок 10 Модель для бинарной классификации

K-блочная проверка – разновидность перекрестной проверки, при которой исходный набор данных разбивается на несколько (обычно 5 или 10) приблизительно равных частей (блоков), затем строится несколько моделей по следующему принципу – например, если при пятиблочной перекрестной проверке исходные данные разделены на пять блоков, то первая модель будет использовать первый блок в качестве тестовой выборки, а все остальные в качестве обучающей выборки. Соответственно вторая модель воспользуется вторым блоком для тестирования, а первый, третий, четвёртый и пятый блоки станут обучающей выборкой и так далее. В результате можно будет оценить как точность модели для каждого из блоков, так и усредненное качество всех моделей, обученных данным способом. В библиотеке sklearn k-блочная проверка представлена в виде функции cross\_val\_score [1].

Теперь можно оценить обобщающую способность модели для бинарной классификации с помощью перекрестной проверки. K-блочная проверка с пятью блоками подтверждает высокое качество модели и демонстрирует, что точность модели может незначительно снизиться при другом разбиении исходных данных (Рисунок 11).

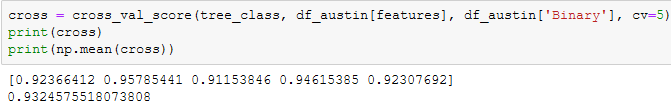


Рисунок 11 Перекрестная проверка для бинарной классификации

Следующим шагом станет оценка значимости признаков. В результате наибольший вклад демонстрирует признак PrecipitationSumInches (84.03%), существеено меньший результат показывает VisibilityLowMiles (14.85%). Так как VisibliAvgMiles демонстрирует наименьший (1.12%) вклад в итоговый результат, в дальнейшем он не будет использоваться. (Рисунок 12)

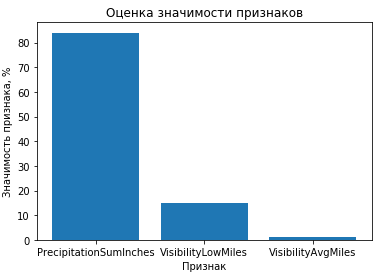


Рисунок 12 Визуализация доли вклада признаков в классификация

## **2.3 Мультиклассовая классификация погодных явлений**

В заключении будет продемонстрировано модель на базе дерева решений для мультиклассовой классификации погодных явлений, соответственно модель должна спрогнозировать будут ли негативные погодные явления и, если да – какие именно?

В целом модель для многоклассовой классификации соответствует своему аналогу для бинарной блассификации, с тем исключением, что в качестве признаков используются только “VisibilityLowMiles” и “PrecipitationSumInches” и целевым атрибутом является “MultiTarget”, значения которого являются закодированным числовым представлением атрибута “Events”. В результате точность на обучающей выборке составляет 86%, а точность на тестовой – 88,3%, что говорит о не идеальном, но достаточно высоком качестве модели[[3]](#footnote-3) (Рисунок 13).

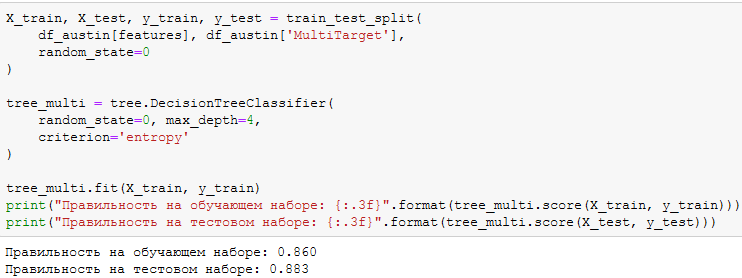


Рисунок 13 Модель для мультиклассовой классификации

Теперь необходимо удостовериться в качестве модели. В первую очередь, по аналогии с бинарной классификацией, осуществляется пятиблочная перекрестная проверка, результат которой подтверждает обобщающую способность модели, пусть и на уровне ниже (в среднем 85%), чем при проверке на тестовой выборке (Рисунок 14)

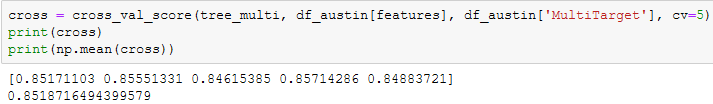


Рисунок 14 Перекрестная проверка для мультиклассовой классификации

Так как решается задача решается задача мультаклассовой классификации в условиях существенного преобладания одного из классов (clear), необходимо определить точность предсказания с точки зрения каждого из имеющегося класса. С этой целью будет построена матрица ошибок, в первую очередь для дальнейшего удобства сгенерирована таблица соответствия наименования класса и его закодированного обозначения (Рисунок 15).

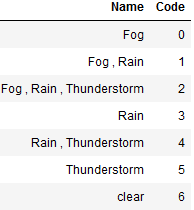


Рисунок 15 Таблица соответствий названий (Name) и закодированных обозначений (Code) классов

Теперь можно построить матрицу ошибок, по оси Y которой представлены фактические метки классов, по оси X – прогнозируемые моделью значения. Соответственно числа на главной диагонали матрицы означают, сколько объектов были успешно классифицированы, а любые значения вне главной диагонали указывают на ошибку классификации. Например, присутствовало 240 значений с меткой “clean”, 227 из них были успешно классифицированы, при этом один объект был ошибочно отнесен к классу “Thunderstorm”, два объекта ошибочно отнесены к классу “Rain, Thunderstorm” и 14 объектов ошибочно отнесены к классу “Rain”. При этом стоит отменить, что наихудший результат показывает класс “Rain”, который представляют 42 объекта, из которых правильно классифицированы только 29 (Рисунок 16)

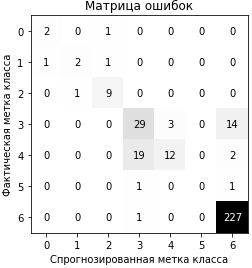


Рисунок 16 Матрица ошибок при мультиклассовой классификации

# **Выводы**

В результате выполнения работы продемонстрировано применение моделей машинного обучения с учителем на базе деревьев решений для задач бинарной и мультиклассовой классификации. Были затронуты такие темы как предварительная обработка данных и метрики оценки обобщающей способности модели после обучения.

В качестве примера был изучен набор данных о погодных явления в Остине за 2015-2017 годы, в результате бинарной классификации факта наличия/отсутствия негативных погодных явлений получена точность 94.5% при проверки на тестовой выборке, 93,2% при перекрестной проверке. В результате мультеклассовой классификации наличия и конкретного типа негативных погодных явлений получена точность 88,3% при проверке на тестовой выборке и 85.1% при перекрестной проверке.

Для модели мультиклассовой классификации была построена матрица ошибок, указывающая что большинство ошибок классификации связаны с классом “Rain” (дождь).

# **Список источников**

1. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — СПб.: Диалектика, 2017. — 480 с.: ил.
2. Iris Species Classify iris plants into three species in this classic dataset URL: <https://www.kaggle.com/uciml/iris>
3. Austin Weather Historical temperature, precipitation, humidity, and windspeed for Austin, Texas URL: https://www.kaggle.com/grubenm/austin-weather

1. Параметр модели машинного обучения, задающийся перед обучением [↑](#footnote-ref-1)
2. 100% точности в данном случае получено благодаря удачному разбиению на обучающую и тестовую выборку, при другом разделении исходных данных на данные для обучения и тестирования точность может незначительно снизиться. [↑](#footnote-ref-2)
3. С точки зрения учебного проекта, для прикладного решения и точности в 99% моет оказаться недостаточно [↑](#footnote-ref-3)