**Задание**

1. Обучение моделей и подбор параметров:
   1. K-ближайших соседей (KNN)
   2. Машина опорных векторов (SVM)
   3. Дерево решений ИЛИ Случайный лес
2. Оценка моделей
   1. Визуализация предсказанных значений
   2. Оценка качества прогноза (precision/recall/f1-score/ROC-AUC)
   3. Визуализация дерева решений ИЛИ Визуализация Feature Importance для случайного леса

**Теоретическая часть**

Random forest — это метод (основанный на подходе «разделяй и властвуй»), использующий ансамбль деревьев решений, созданных на случайно разделенном датасете. Набор таких деревьев-классификаторов образует лес.

KNN - для каждого объекта находятся k его ближайших соседей по тренировочной выборке. Объект относится к тому классу, который является наиболее распространённым среди k соседей.

Главная цель SVM как классификатора — найти уравнение разделяющей гиперплоскости в пространстве объектов, которая бы разделила два класса неким оптимальным образом.

**Выполнение работы**

Предварительно была произведена подготовка данных: перевод текстовых данных в числовые; замена поля «Дата» с соответствующим форматом в отдельные поля – день, месяц, год. Результат приведен на рис.1.

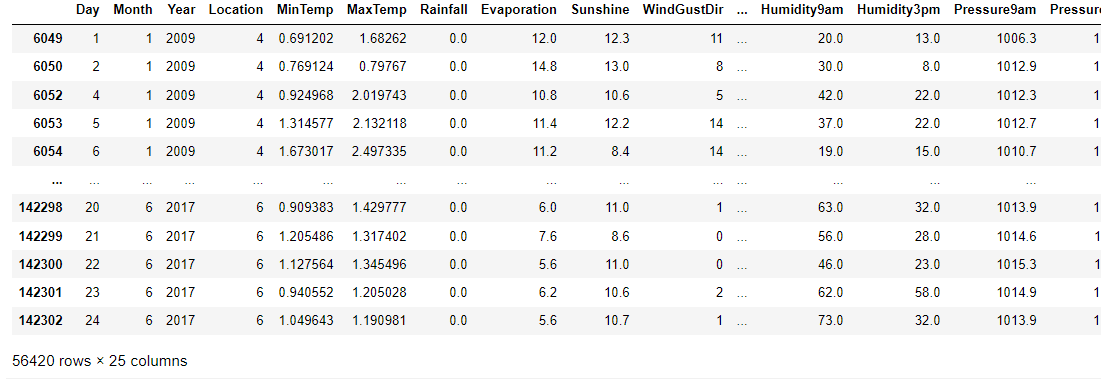


Рисунок 1

Выбраны данные для обучения и для результата классификации (рис.2).

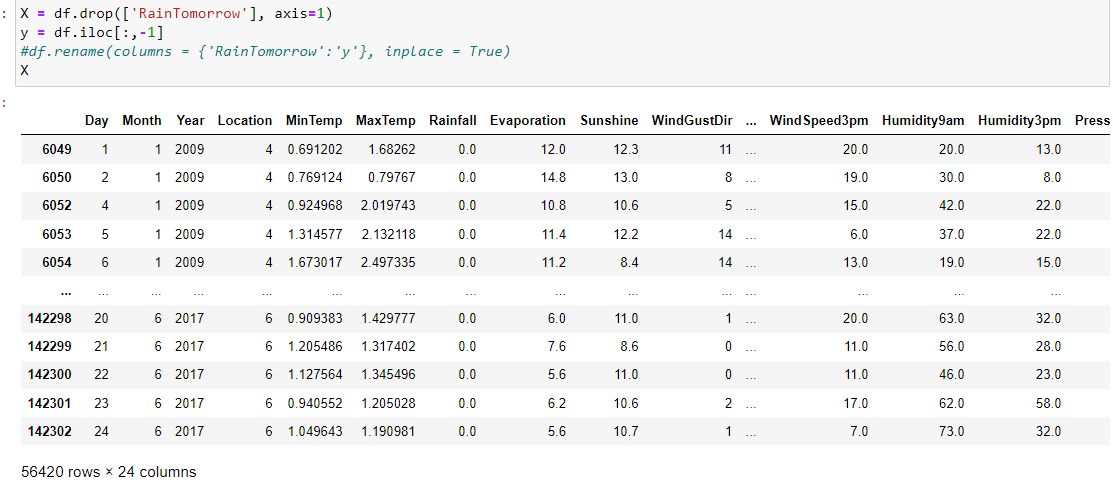
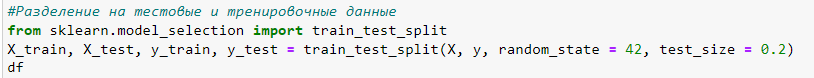


Рисунок 2

Для обучения модели данные были разделены на тестовые и тренировочные (рис.3) и приведены к целочисленному типу.



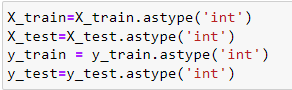


Рисунок 3

Классификация методом К ближайших соседей. Код обучения и предсказания классификатором приведен на рис.4.

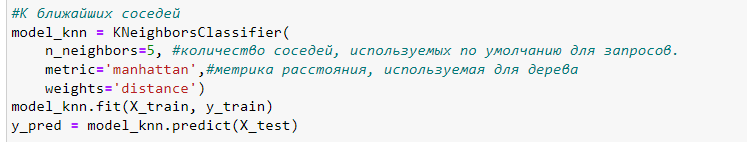


Рисунок 4

Результаты классификации по метрикам приведены на рис.5.

Accuracy = Correct Predictions / Total Predictions (доля объектов, в которых предсказанный класс совпал с реальным).

Precision = TruePositive / (TruePositive + FalsePositive) (вероятность того, что объект, отнесенный моделью к классу, окажется объектом этого класса).

Recall = TruePositive / (TruePositive + FalseNegative) (вероятность того, что объект класса будет отнесен моделью к этому классу).

F-Measure = (2 \* Precision \* Recall) / (Precision + Recall)

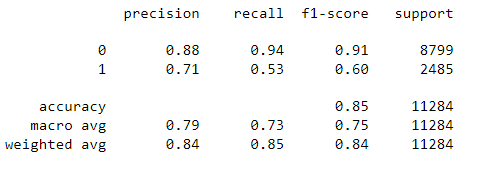


Рисунок 5

Классификация методом опорных векторов. Код приведен на рис.6.

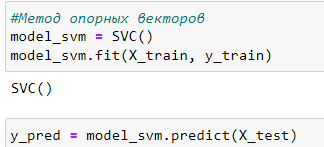


Рисунок 6

Результаты классификации по метрикам приведены на рис.7. Видно, что в результате классификации подавляющее большинство объектов были отнесены к классу 0, включая многие объекты из класса 1 (recall (1) -> 0).

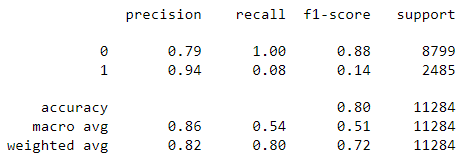


Рисунок 7

Классификация методом «Случайный лес». Код приведен на рис.8.

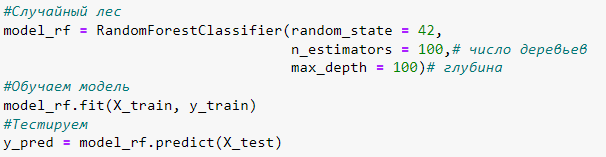


Рисунок 8

Результаты классификации приведены на рис.9 и рис.10.

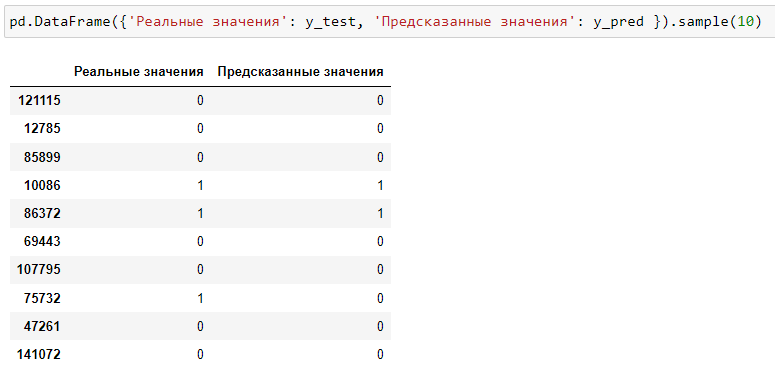


Рисунок 9

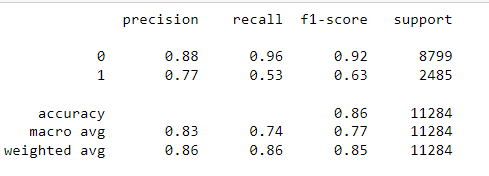


Рисунок 10

Проведем сравнение между результатами классификации и оценим качество прогноза, используя ROC-кривые (рис.11).

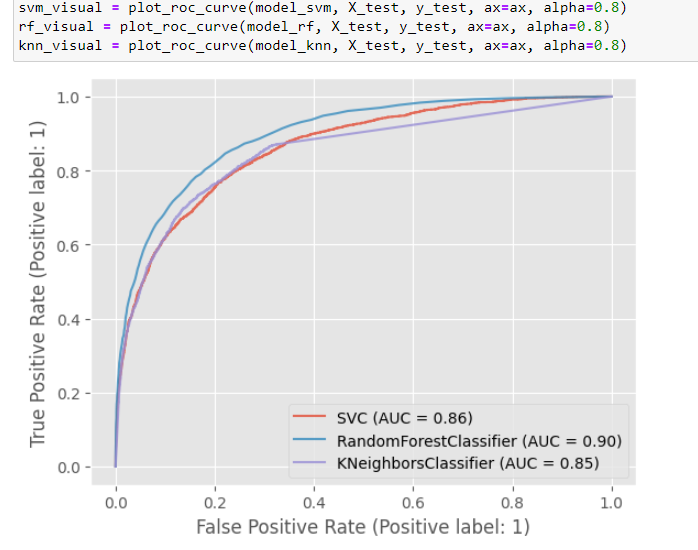


Рисунок 11

Визуализация Feature Importance (определение наиболее значащих признаков) для случайного леса приведена на рисунке 12.

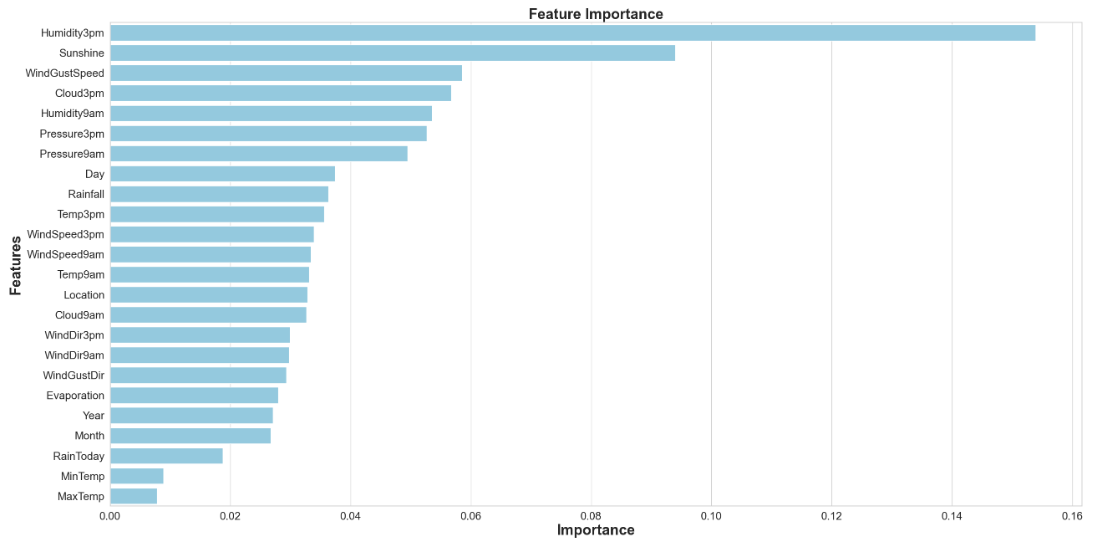


Рисунок 12

Визуализация предсказанных значений для KNN приведена на рис.13.

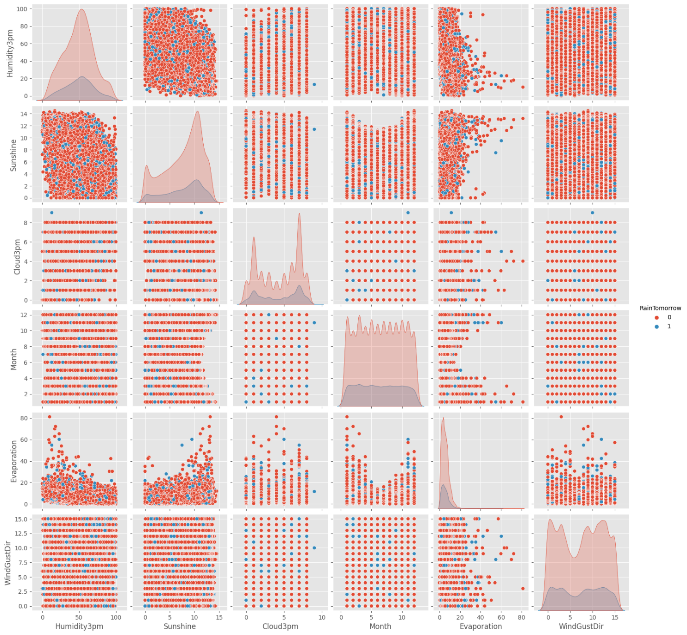


Рисунок 13

Ввиду сложного распределения значений RainTomorrow, работа по классификации данных может дать непредсказуемый результат.

Визуализация предсказанных значений для SVM приведена на рис.14.

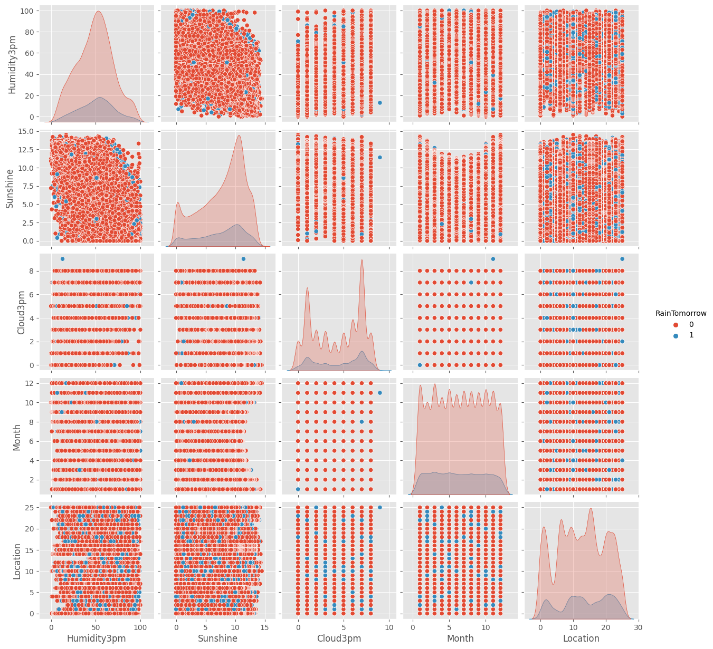


Рисунок 14

Визуализация предсказанных значений для RF приведена на рис.13.

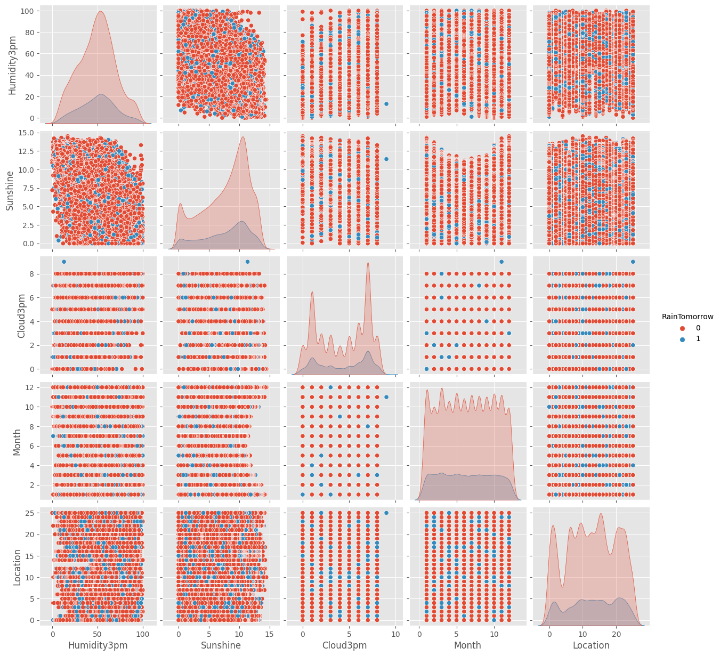
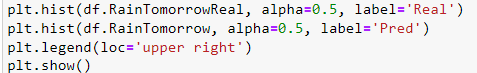
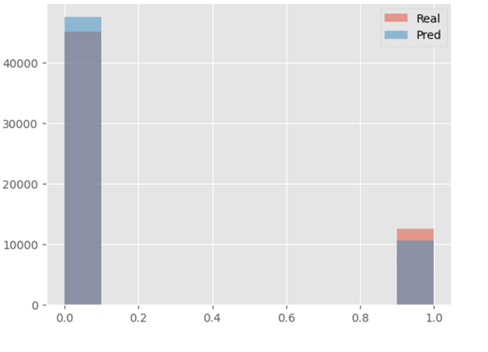


Рисунок 15

График, сравнивающий реальное значение «Будет ли дождь завтра» с предсказанным, приведен на рис.16.



Отметим, что в случае отсутствия информации в наборе данных о реальной погоде на следующий день, данные о дне уничтожаются для более точных данных при сравнении. По графику видно, что классификатор KNN отнес к классу 0 больше объектов, чем необходимо.

  
Рисунок 16

Код, позволяющий оценить реальную картину осадков на следующий день, приведен на рис.17.

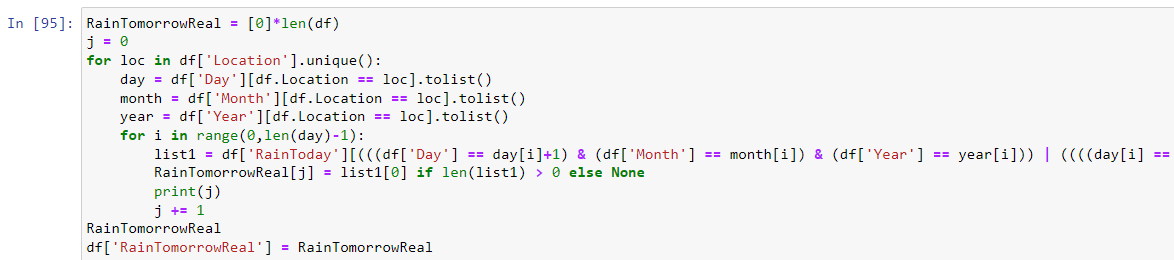


Рисунок 17

Приведем также гистограмму значений, предсказанных классификатором по методу случайного леса (рис.18) и методу опорных векторов (рис.19).

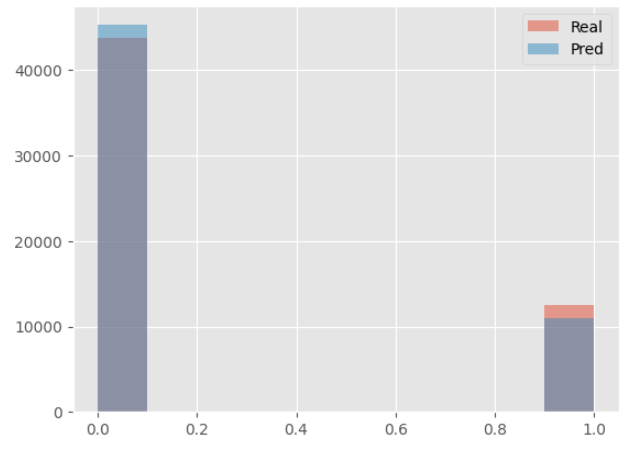


Рисунок 18

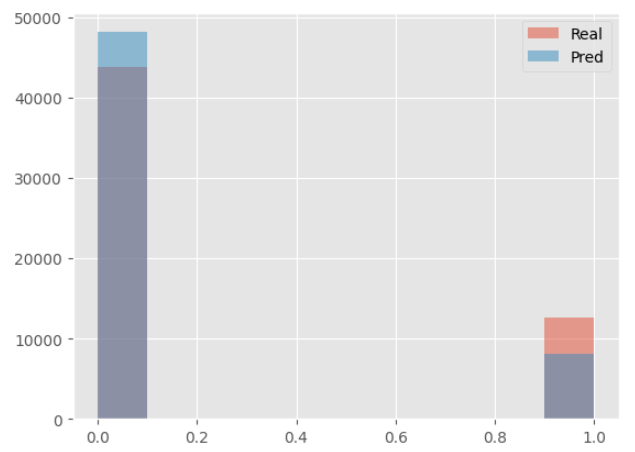


Рисунок 19

**Выводы**

В результате выполнения работы были получены и закреплены навыки предобработки данных и применения методов машинного обучения для решения задач классификации.

Были получены результаты прогноза «Будет ли дождь завтра?» тремя разными классификаторами: методом случайного леса, k ближайших соседей и методом опорных векторов. Наибольшую точность показал метод случайного леса (0.90). Результаты прогноза показали, что большая часть дней в Австралии пройдет без осадков. Визуализация прогноза оказалась сложной задачей ввиду большого количества данных и отсутствием прямой зависимости между признаками и результатом. Однако по гистограммам, отражающим распределение значений – реальных и прогнозируемых – можно сделать вывод о достаточно большой точности классификаторов. Наиболее близкие результаты были получены классификатором случайного дерева, графики это также доказывают.