**Задание 4 – Классификаторы (Apache Spark)**

Напишите программу на PySpark согласно вашему варианту для набора данных “DS\_2019\_public.csv” с помощью трёх алгоритмов, из которых нужно взять и логистическую регрессию, случайный лес и дерево решений. Напишите небольшой отчёт, который должен содержать титульный лист, вариант задания, код программы и полученные вами результаты. Для помощи в выполнении задания следует пользоваться книгой «Machine Learning with PySpark -2019.pdf» и «PySpark SQL Recipes».

Описание столбцов набора данных (расширенное) находится в файле recs2009\_public\_codebook.xlsx. В отчёте обязательно должно быть описание выбранных признаков, предсказываемых признаков/категорий, почему вы сделали такой выбор и выведена точность работы полученной модели. Если точность модели получилась невысокой, внесите изменения в выбор признаков, чтобы повысить точность, и отобразите это в отчёте.

Задача по полученному варианту.

Для этого необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки:

Метка класса – Climate\_Region\_Pub. Так как необходимо создать бинарные классификаторы а возможных классов – 5, то сначала необходимо изменить значение метки Climate\_Region\_Pub.

Для этого необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки:

если последняя цифра 0 или 5: метку 0 заменить на класс А, метки 1, 2, 3, 4 заменить на класс В;

если последняя цифра 1 или 6: метку 1 заменить на класс А, метки 0, 2, 3, 4 заменить на класс В;

если последняя цифра 2 или 7: метку 2 заменить на класс А, метки 0, 1, 3, 4 заменить на класс В;

если последняя цифра 3 или 8: метку 3 заменить на класс А, метки 0, 1, 2, 4 заменить на класс В;

если последняя цифра 4 или 9: метку 4 заменить на класс А, метки 0, 1, 2, 3 заменить на класс В.

Выбрать признаки, использующиеся при обучении, и, если необходимо, выполнить их предобработку. Разделить выборку на обучающую и тестовую. Написать короткий отчет по работе, включив в него программу с комментариями, значения качества моделей TP, TN, FP, FN.