Напишите программу на PySpark согласно вашему варианту для датасета “DS_2019_public.csv”. Датасет находится в папке English/Datasets на гугл диске Аксёнова С.В. 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». Также могут пригодиться книги из предыдущей лабораторной работы - «Изучаем Spark. Молниеносный анализ данных - 2015», «PySpark SQL Recipes» или интернетом.

Описание столбцов датасета (расширенное) находится в файле ниже на моём персональном сайте (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.

Вариант выбирается согласно вашему номеру в журнале, и выдаётся преподавателем лично на занятии. Также следует быть готовым ответить на вопросы по теме.

Вариант 1.

Напишите программу, реализующую модель линейной регрессии. Выберите несколько количественных признаков (не менее 5), и количественных признак, который вы будете предсказывать. Работы всех студентов в рамках варианта 1 должны предсказывать различные признаки.

Вариант 2.

Напишите программу, реализующую модель логистической регрессии. Выберите несколько количественных признаков (не менее 5). Предсказывайте одну из категорий Climate_Region_Pub – установите одну из категорий согласно вашему варианту задания на 1, а все остальные – на 0.

Вариант 3.

Напишите программу, реализующую модель случайных лесов. Выберите несколько количественных признаков (не менее 5). Предсказывайте одну из категорий Climate_Region_Pub – установите одну из категорий согласно вашему варианту задания на 1, а все остальные – на 0.