Write a program in PySpark according to your task variant. Use the dataset “DS_2019_public.csv”. Dataset can be found in the folder English/Datasets on Google Drive of S.V. Axyonov. Write a small report which must contain the title page, task variant, program code and obtained results. To complete the task, use the book «Machine Learning with PySpark -2019.pdf». You may also use books from the previous lab work - «Learning Spark. Lightning-fast data analysis», «PySpark SQL Recipes» or the Internet.

The description of dataset columns (extended) can be found in file on my personal TPU website (below, named recs2009_public_codebook.xlsx). The report must contain the description of chosen features, features/classes you are going to predict, explanation of your choice and accuracy metrics (precision, recall etc. according to the book). If the accuracy ended up being not very high, make changes in feature selection to improve it, and reflect it in the report.

The number of variant/predicted climate region is given by the teacher on the lab lesson. You should also be prepared to answer some questions on the topic.

**Variant 1.**

Write a program that implements linear regression model. Choose a few quantitative features (no less than 5) and a quantitative feature to predict. All students of variant 1 must predict different features.

**Variant 2.**

Write a program that implements logistic regression classification model. Choose a few quantitative features (no less than 5). Predict one of the categories of Climate_Region_Pub – according to your task variant, one of the categories should be set to 1, and the rest to 0. I.e. if you need to predict region 3, all numbers 3 in column are set to 1, and rest to 0.

**Вариант 3.**

Write a program that implements random forest classification model. Choose a few quantitative features (no less than 5). Predict one of the categories of Climate_Region_Pub – according to your task variant, one of the categories should be set to 1, and the rest to 0. I.e. if you need to predict region 3, all numbers 3 in column are set to 1, and rest to 0.