Assignment A04: Machine Learning with SparkML

You should thoroughly read through the entire assignment before beginning your work! Don't start the cluster until you are ready.

Start your cluster

Start EMR Cluster and connect to it as in previous labs

Provide the Master Node and Cluster Metadata

Provide the instance metadata. Run this command in the master node.

```
curl http://169.254.169.254/latest/dynamic/instance-identity/document/ > instance metadata.json
```

Problem

In this assignment, you will work with the Hospital Inpatient Discharges Data.

The data is in CSV format. There are approximately 2.3 million patients in the dataset.

The data is accessible at
https://health.data.ny.gov/api/views/82xm-y6g8/rows.csv?accessType=DOWNLOAD

There are two parts to this homework, each in its own Jupyter notebook.

- transform-data: in this notebook you will load the dataset, explore the data, extract the features you wish to use and store this new dataset in **your S3 bucket**.
- model-data: in this notebook you will work with the transformed dataset you created in transform-data.ipynb and train one or more models to predict whether a hospital stay is going to exceed 3 days.

The data size is approximately 1GB uncompressed. Intensive operations should not take more than a few minutes each. **Even though Spark is relatively fast, it still takes time to process operations.**

In this assignment, **we are giving you directions on what to do, but not how to do it.**
In each notebook there is a section at the end where you need to provide specific information. Please do not edit the structure of these cells.

Some suggestions you should consider:

- Review both notebooks before starting the assignment, and perhaps starting your cluster
- Use AWS spot pricing for your cluster. With m4.xlarge machines, each machine costs $0.20/hr per machine time, plus the EMR fee of $0.06 for a total of $0.26 per machine per hour. You can save money with spot pricing (just on machine time, not on EMR cost.) Also, remember that spot instances can terminate without warning so use with caution.
- Refer to the PySpark Documentation and Spark Documentation
- Start early on this assignment, not two days before it is due. This assignment will take more time than previous assignments.
- Consider saving intermediate datasets in **your S3** buckets, in Apache Parquet format.
- Consider saving a model object in S3 after you train it, especially if training takes a while. To save a model object, use the following code: model.save("s3://[[your-s3-bucket]]/model_location/")
- When creating the Machine Learning pipelines, you may want to try it first on a small sample of your training data to make sure the pipelines work as planned. To create a tiny DataFrame, use the limit method: df.limit(100) (this creates a small DataFrame with the first 100 rows from df.)
- If you need to re-start your Jupyter notebook for any reason, make sure you close the Spark connection first **before** restarting the kernel. To do this, type either sc.stop() or spark.stop() in a cell. If you don't do this, YARN will not release resources previously allocated.

Submitting the Assignment

Make sure you submit only **files requested.**

The files to be submitted and uploaded to Canvas are:

- instance-metadata.json
- transform-data.ipynb
- model-data.ipynb

Grading Rubric

- We will look at the results files and/or scripts. If the result files are exactly what is expected, in the proper format, etc., we may run your scripts to make sure they produce the output. If everything works, you will get full credit for the problem.
- If the submitted results are not what is expected, we will look at and run your code and provide partial credit wherever possible and applicable.
- Points **will** be deducted for each the following reasons:
    - Instructions are not followed
    - Output is not in expected format (not sorted, missing fields, wrong delimiter, unusual characters in the files, etc.)
    - There are more files submitted than need to be
    - There are additional lines in the results files (whether empty or not) o Files are not the requested filename
    - Homework is late (unless you are using a late day and provide notice in advance)