**Lab 9 - BDAS PySpark Algorithms and Visualisations**

**Lab objectives:** In this lab, you will:

1. Understand the differences between regression and classification.
2. Understand the differences between supervised and unsupervised learning.
3. Apply linear regression and logistic regression in a PySpark context.
4. Understand the different evaluation metrics available for the different algorithms.

Please connect to your EC2 instance (either using SSH or Connect) and go through the Jupyter Notebooks listed below. Once you have access, you can view the lab materials on GitHub by clicking the links below. We recommend interacting with these notebooks directly by cloning the [aws-instance repository](https://github.com/shohil-kishore/aws-instance).) in your EC2 instance. **Note that the datasets are also available on the GitHub repository.** Detailed instructions are available in the “Lab 7 – Walkthrough” document.

We recommend starting the tutorials in the following order:

1. [Linear Regression Documentation Example](https://github.com/shohil-kishore/aws-instance/blob/master/Spark%20Algorithms%20-%20Linear%20Regression%20(Documentation%20Example).ipynb), followed by the [Advanced Example.](https://github.com/shohil-kishore/aws-instance/blob/master/Spark%20Algorithms%20-%20Logistic%20Regression%20(Advanced%20Example).ipynb)
   1. Evaluation metric details are available at the bottom of the documentation example.
2. [Binomial Logistic Regression Documentation Example,](https://github.com/shohil-kishore/aws-instance/blob/master/Spark%20Algorithms%20-%20Logistic%20Regression%20(Documentation%20Example).ipynb) followed by the [Advanced Example.](https://github.com/shohil-kishore/aws-instance/blob/master/Spark%20Algorithms%20-%20Logistic%20Regression%20(Advanced%20Example).ipynb)
   1. Evaluation metric details are available at the bottom of the documentation example.
   2. Visualisation examples are available at the bottom of the advanced example.
3. [Tree Methods Documentation Example](https://github.com/shohil-kishore/aws-instance/blob/master/Spark%20Algorithms%20-%20Tree%20Methods%20(Documentation%20Example).ipynb), followed by the [Advanced Example.](https://github.com/shohil-kishore/aws-instance/blob/master/Spark%20Algorithms%20-%20Tree%20Methods%20(Advanced%20Example).ipynb)

Please do your best to understand the material instead of copying and pasting. Ask your tutors, speak to your classmates or use Google. **Please note that next week’s lab does not introduce any new content. It provides you with some additional time to ask questions and work through tutorials.**

**Troubleshooting:**

* **Why do I keep getting a 404 error when I click on the links?** That means that you don’t have access to the repository yet (as it’s a private repository, you need to request for access). To get access, simply follow Shohil on GitHub (<https://github.com/shohil-kishore>) and he will provide you with access to the repository. If you need the files urgently, they’re also available on Canvas.
* **I want something simpler, can I use VirtualBox instead?** Yes, you can. While we recommend that you use AWS, if you get locked out of your account or don’t have stable internet access, VirtualBox is a viable alternative. Follow the instructions [here](https://docs.google.com/document/d/1OagPKXZ8Lk2Ahvpkv1wF81YGoz4SCIzGAVOkk5cNkpY/edit?usp=sharing).