# AIOPS Assignment 3

Submitted by **Diana Varghese**

(dianavarghese100@gmail.com)

## 1. What is DVC, and why is DVC used?

Data Version Control or DVC is a command line tool and VS Code Extension to help you develop reproducible machine learning projects. DVC builds upon Git by introducing the concept of data files – large files that should not be stored in a Git repository, but still need to be tracked and versioned. It leverages Git's features to enable managing different versions of data, data pipelines, and experiments.

## 2. How is DVC different from git and GitHub?

DVC is a data and ML experiment management tool that takes advantage of the existing engineering toolset that we are familiar with (Git, CI/CD, etc.). DVC is meant to be run alongside Git but it is not fundamentally bound to Git and can work without it (except versioning-related features).

Git is employed as usual to store and version code. Git is a software while GitHub is a service.

## 3. Which command can be used to initialise a DVC project?

```
dvc init [-h] [-q | -v] [--no-scm] [-f] [--subdir]
```
Initialize a DVC project in the current working directory.

## 4. In what all use cases DVC can be used?

- Version your data and models. Store them in your cloud storage but keep their version info in your Git repo.
- Iterate fast with lightweight pipelines. When you make changes, only run the steps impacted by those changes.
- Track experiments in your local Git repo (no servers needed).
- Compare any data, code, parameters, model, or performance plots.
- Share experiments and automatically reproduce anyone's experiment.

## 5. Which command can be used to reproduce the entire pipeline?

```
dvc repro [-h] [-q | -v] [-f] [-i]
          [-s] [-p] [-P] [-R] [-m]
          [--downstream] [--force-downstream]
          [--pull] [--dry]
          [--glob] [--no-commit] [--no-run-cache]
          [targets [<target> ...]]
```
This can reproduce complete or partial pipelines by executing commands defined in their stages in the correct order.

## 6. Which DVC command can be used to check metrics?

```
dvc metrics [-h] [-q | -v] {show,diff} ...
```

This command is used to see certain output metrics for the ML project such as AUC, ROC, false positives, etc.

**7. Can we store a large amount of Data on GitHub? Justify**

GitHub limits the size of files allowed in repositories. If you attempt to add or update a file that is larger than 50 MB, you will receive a warning from Git. The changes will still successfully push to your repository but there could be a performance impact. If you add a file to a repository via a browser, the file can be no larger than 25 MB. GitHub blocks all files larger than 100 MB. To track files beyond this limit, you must use Git Large File Storage (Git LFS).

Git recommends that repositories remain small, ideally less than 1 GB as smaller repositories are faster to clone and easier to work with and maintain.