**AIOPS Assignment 3**

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

DVC is built to make ML models shareable and reproducible. It is designed to handle large files, data sets, machine learning models, metrics, and code.

DVC is a system for data version control. It is essentially like Git but is used for data. With DVC, you can keep the information about different versions of your data in Git while storing your original data somewhere else.

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

Git: Fast, scalable, distributed revision control system. Git is a free and open-source distributed version control system designed to handle everything from small to very large projects with speed and efficiency

DVC: Open-source Version Control System for Machine Learning Projects. It is an open-source Version Control System for data science and machine learning projects. It is designed to handle large files, data sets, machine learning models, metrics, and code.

**3. Which command can be used to initialize a DVC project?**

To start the versioning process you have to create a git repository in the base folder of your data and initialize DVC afterwards through

git init

DVC init

**4. In What use cases DVC can be used?**

Data and model versioning is the base layer of DVC for large files, datasets, and machine learning models. Use a standard Git workflow, but without storing large files in the repo. Data is cached by DVC, allowing for efficient sharing. Think "Git for data".

Data and model access shows how to bring, explore, and access data artifacts from outside the project. This can help download a specific version of an ML model to a deployment server or import a dataset into another project, for example.

Data pipelines describe how models and other data artifacts are built, and provide an efficient way to reproduce them. Think "Makefiles for data and ML projects" done right.

Metrics, parameters, and plots can be attached to pipelines. These let you capture, evaluate, and visualize ML projects without leaving Git.

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

-p , --pipeline - reproduce the entire pipelines

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

A set of commands to display and compare metrics: show, and diff.

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

**GitHub blocks files larger than 100 MB**