**Algorithm Understanding**

**How does the Gradient-Boosted Tree Algorithm work in Classification? How does Gradient Boost differ from AdaBoost and Logistic Regression?**

Gradient boosting is a method where you modify weak models to make them perform better.

AdaBoost was the first designed boosting algorithm, with a particular loss function. Gradient Boosting is a generic algorithm that that can be used with any differentiable loss function. This makes Gradient Boosting more flexible than AdaBoost.

**Interview Readiness**

**What is a Delta Lake and how does it offer a solution to building reliable data pipelines?**

Delta Lake is a new generation of data storage solutions, called Lakehouse platforms. Lakehouse platforms can store data files in open formats (structured, semi-structured and unstructured data), but it also includes a meta- data layer that adds transactions, versioning, and auxiliary data structures over the data files in an open format, and that can be queried directly with diverse APIs and engines.

This platform avoid working in tandem with data lakes and then doing ETL to extract data to a data warehouse (structured data) before being able to use it for BI, reports and data science.

**Interview Readiness**

**When working with Pandas, we use the class pandas.core.frame.DataFrame and when working with the pandas API in Spark, we use the class pyspark.pandas.frame.DataFrame, are these the same, explain why or why not?**

Spark allows you to do cluster computing to do parallel jobs on several machines. Pandas is one of the most used libraries to work with structured data for analysis, ML and data science, but is not designed to work on several machines. Pyspark.pandas does run on several machines and is written to run in Apache Spark.

While pyspark.pandas allow you to work with big data, it has some differences from pandas, like lazy execution, immutability and is more difficult to perform complex operations with.

Pandas will crash with very large amounts of data making it impossible to use in those cases to analyze the data.

**Interview Readiness**

**What is a Machine Learning Pipeline is and why it’s important? What are the steps in a Machine Learning workflow?**

In machine learning, it is common to run a sequence of algorithms to process and learn from data. A pipeline is a way of sequentially combining multiple algorithms into a single workflow, making it easy to repeat and adjust.

Thus, a ML pipelines consist of multiple sequential steps that do everything from data extraction and preprocessing to model training and deployment.

ML workflows include data extraction and analysis, data cleaning and processing and model training, model evaluation and validation, model optimization.