Case Study 4:

Abstract Statement

Alterimac is an organisation working on big data ecosystems for enhanced insight extractions. Recently, their client has included advanced machine learning as a requirement for data insights. This has reflected into an architecture which includes usage of Dataproc, BigQuery and Apache Spark ML to perform machine learning on a dataset.

Use any of the following tools:

- BigQuery, to prepare the linear regression input table, which is written to your Google Cloud project
- Python, to query and manage data in BigQuery
- Apache Spark, to access the resulting linear regression table
- Spark ML, to build and evaluate the model
- Dataproc PySpark job, to invoke Spark ML functions

We are particularly supposed to work on a natality dataset available on GCP which is an open dataset to experiment and test machine learning capabilities of the big query platform.

Use linear regression to build a model of birth weight as a function of five factors:

1. gestation weeks
2. mother's age
3. father's age
4. mother's weight gain during pregnancy
5. Apgar score