Name: Shah, Dhrumi Date: 22/12/2023

NYU ID: ds7669

NET ID: N13402329

Name: Jain, Shraddha

NYU ID: sj4391

NET ID: N14396386

Course Section Number: 001

**Project Part IV - IV End-to-End Solution Integration and Data-Driven / Database Programming**

Total in points (100 points total): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Professor’s Comments:

Affirmation of my Independent Effort: Shraddha Jain, Dhrumi Shah

**Business Use Case:**

To provide our Insurance Policy seeking clients with an approximate cost based on their health conditions like Chronic Diseases. Here we have developed an ML model which helps us identify at-risk patients which will help choose the right policy package for them. Here we have delved deep into identifying one particular segment of a Chronic Disease, Cancer and more specifically we are focusing on the branch of Lung cancer. This ML model will identify the patients at Low(0), Medium(1) or High(2) risk of lung cancer and then the policy agent helps the clients to choose the right Health insurance policy which is customized specially for their needs.

**Steps and Workflow:**

1. Users are given a form where they have to fill in the details about their health, various parameter and level-based questions are asked in order to find out whether the user is an at-risk patient or not.
2. This data is being fed to the ML Model and then the model runs tests to predict if the user is an at-risk patient of Lung Cancer.
3. This data is further stored into the database created by us for future uses to identify more patients based on the indicators.
4. After finding out the health concerns of the individuals the health insurance policy and its estimated cost is given to the client.

**BPNM notation:**

A black arrow pointing to a white grid

Description automatically generated

The steps in the notation are as follows:

* The client will enter the info into the form that is presented by the insurance form.
* The data is being interpreted and the at-risk patients are being identified.
* The decision whether the client is an at-risk patient or not.
* Then finally a quotation on the basis of the health condition of the user is identified.

Our ML model and application risk will also help clients identify if they are at high, medium or lower risk of Lung Cancer.

**Creating an application to identify whether someone is at-risk of Lung Cancer.**

**Front-end application for retrieving Health data**

A screen shot of a computer

Description automatically generated

**Basic html web page creation**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Our front-end page where the data will be fed:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**The prediction result will be displayed as below in the left bottom screen**

A screenshot of a computer

Description automatically generated

**The dataset:**

The dataset we have is the one interpreted from the lung Cancer dataset.

The dataset is as follows:

A screenshot of a computer

Description automatically generated

We have uploaded this dataset in BigQueryCloud in order to run it and query it for creating the ML models based on the information provided.

This has mentioned all the health indicators that may help us identify the at-risk patients that might be affected with Chronic Disease.

**The data visualization of the dataset is as follows:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

Various data visualizations are being shown here that depict various indicators and contents of the dataset.

**Histogram:**

A group of blue bars

Description automatically generated

**Pie chart:**

A pie chart with numbers and a few percentages

Description automatically generated

**Scatter Plots:**

A graph with blue dots

Description automatically generated

A graph with red dots

Description automatically generated

**Heatmap:**

A green chart with black text

Description automatically generated with medium confidence

**The ML models that are being run by us to identify the at-risk patients are as follows:**

1. **KNN**

A screenshot of a computer code

Description automatically generated

1. **Decision Tree**

A screenshot of a computer program

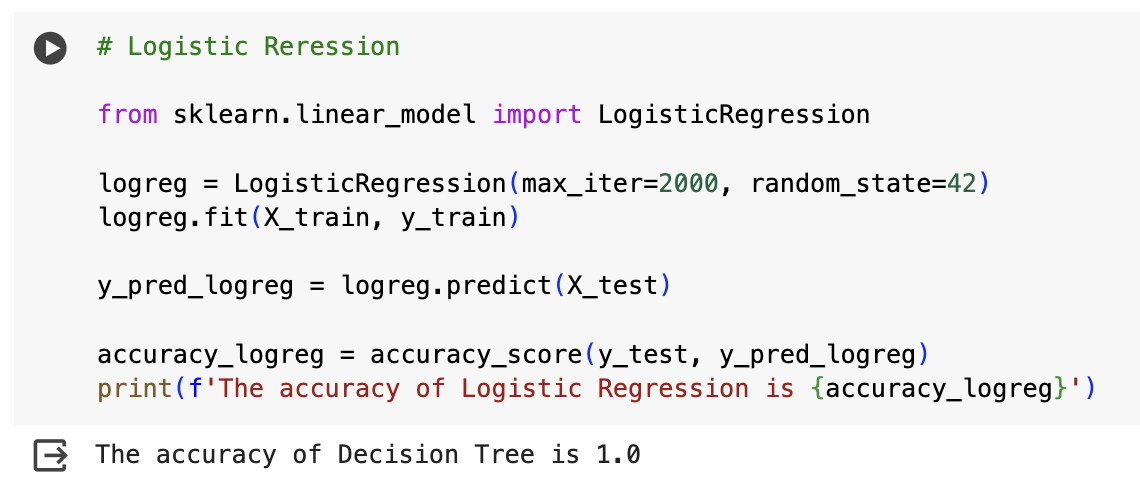
Description automatically generated

1. **NB**

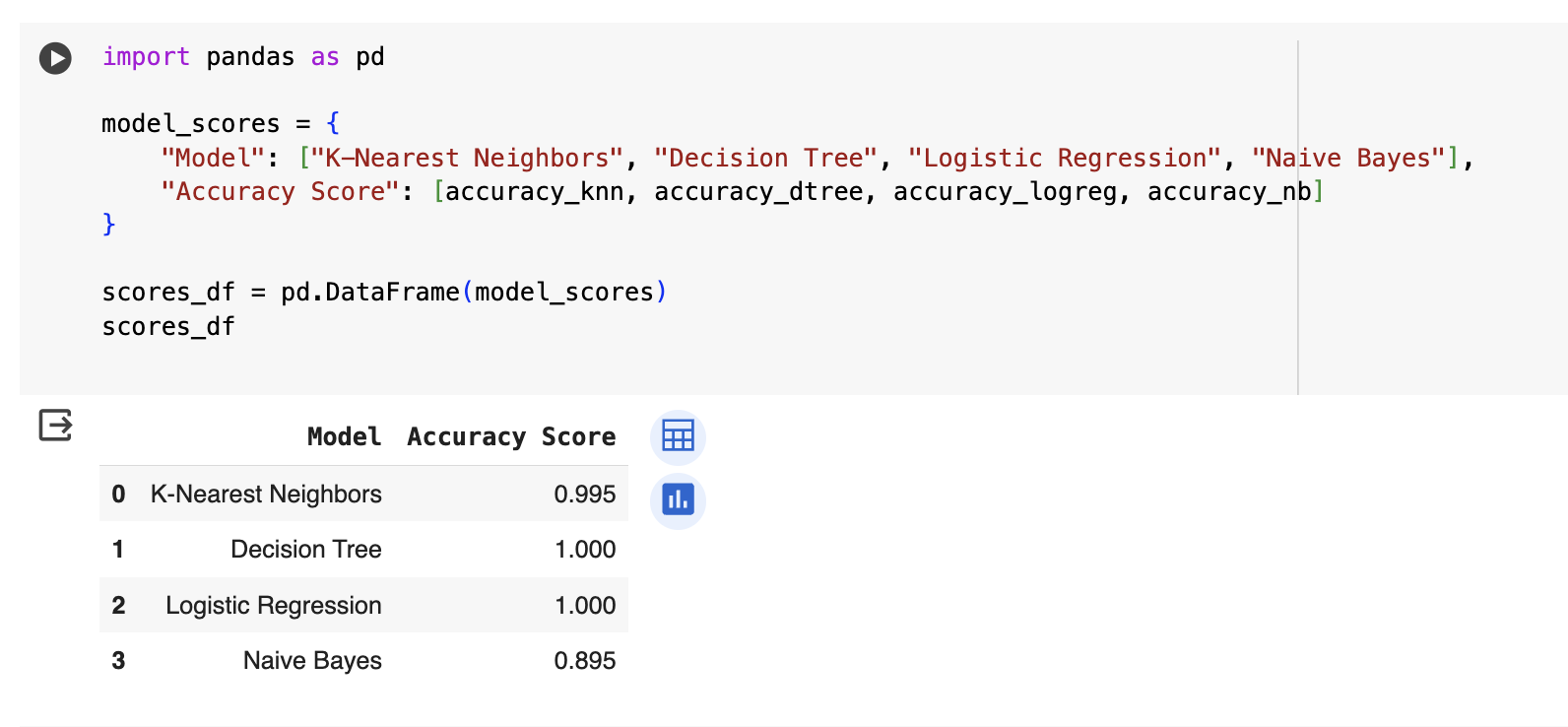
A screenshot of a computer code

Description automatically generated

1. **Logistic Regression**



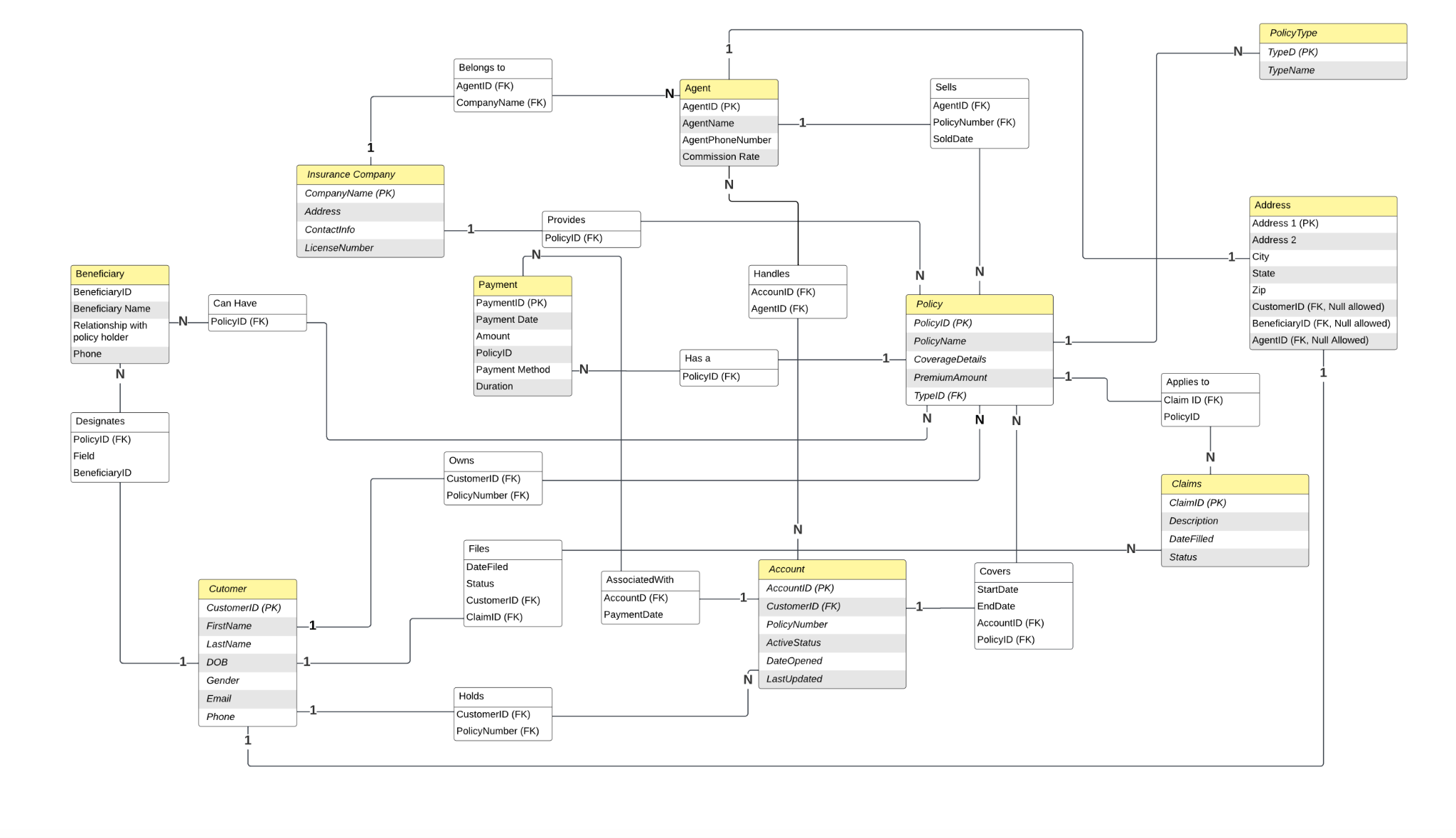
We also tested the model’s accuracy and compared them to identify our overall ML model accuracy



A screenshot of a graph

Description automatically generated

**Data Pipeline**



All three normal forms are satisfied by the above-described schema. Below is the justification for the same:

**1NF -** *When all attribute values are atomic, meaning they cannot be subdivided into smaller values, a relation is said to be in 1NF. None of the values in the aforementioned schema can be divided into more than one value or object; all values must be atomic.*

**2NF -** *A relation is said to be in 2NF if all of the information in the candidate key is dependent upon any non-key column. The ids of each table serve as the main key in the schema design. Additionally, the primary key of the table allows for the unique identification of each non-primary key. A partial primary key does not depend on any non-primary key.*

**3NF-** *When every non-prime attribute depends directly (as opposed to transitively) on the full candidate key, a relation is said to be in 3NF. In the tables above, there isn't a single column that is transitively dependent on the primary key.*

Therefore, our schema follows all the normal forms for the given dataset.

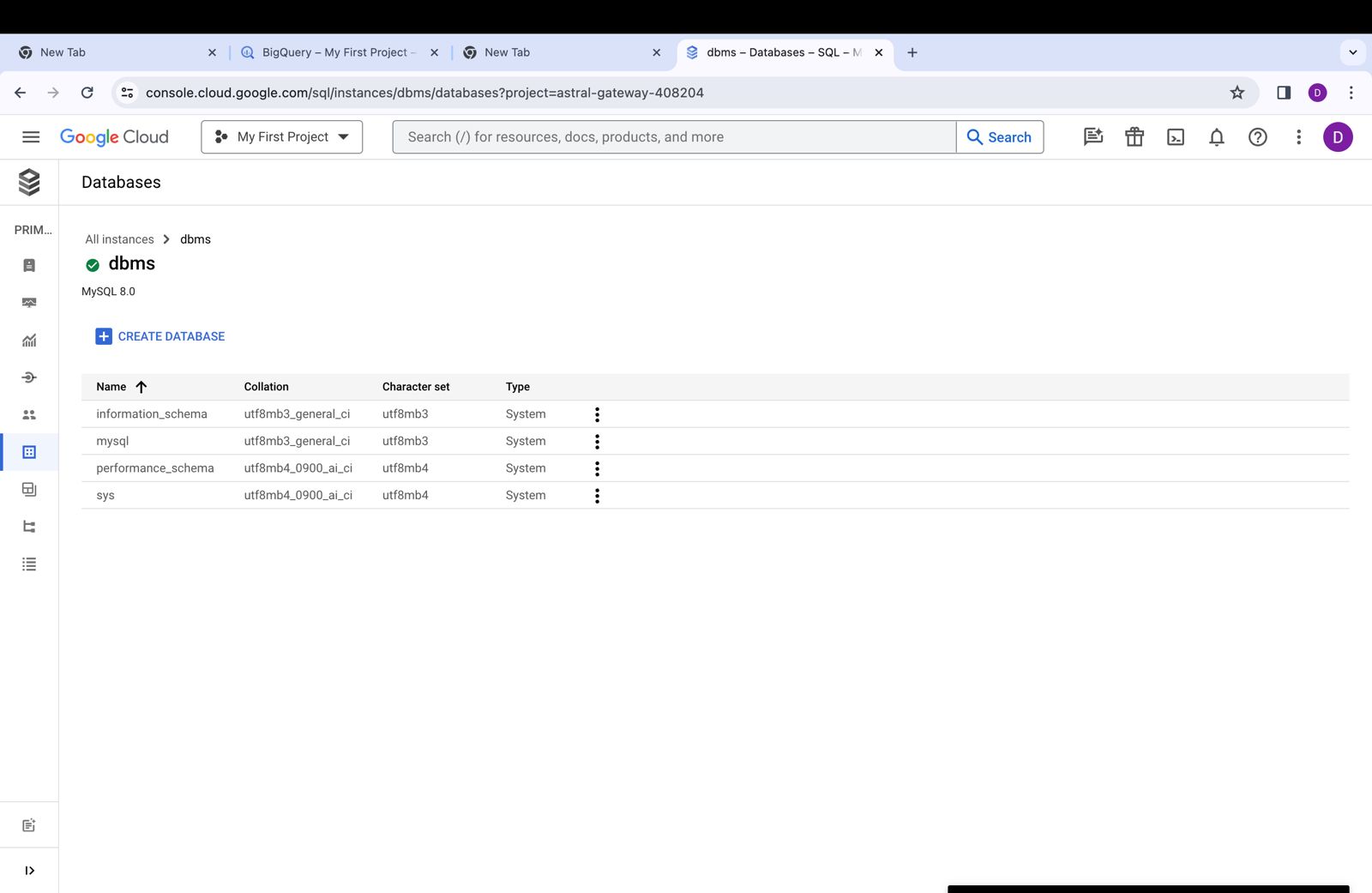
**Storing data in the Cloud Database**

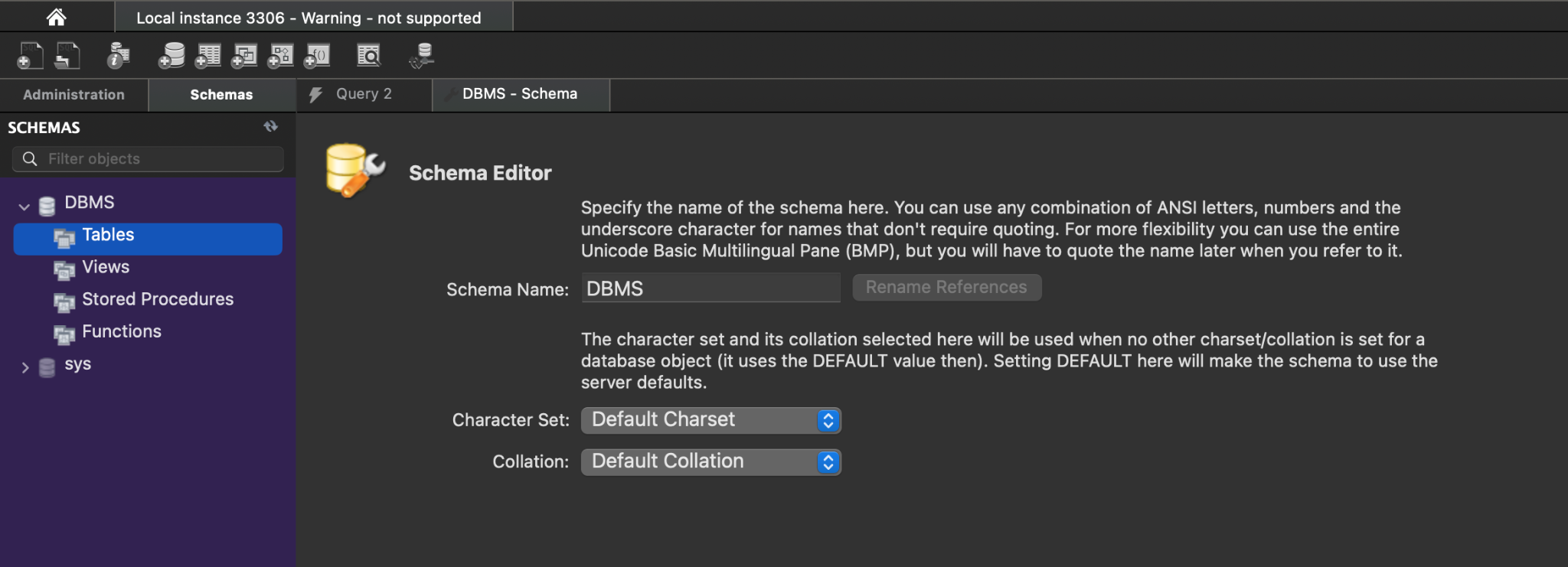
We have used Google cloud to store our dataset using MySQL.

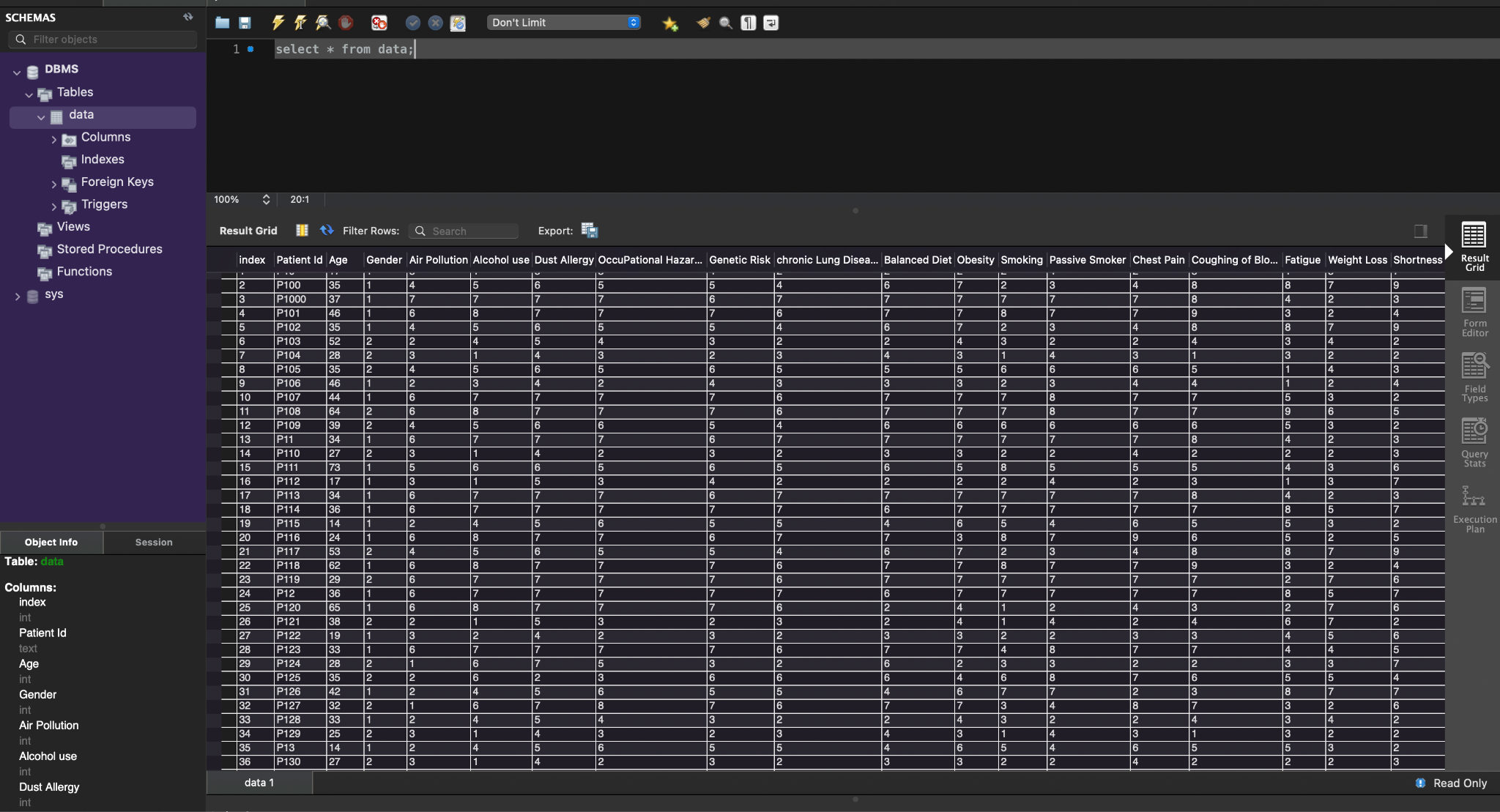
Added the local IP Address of MySQL under the connection details of Google Cloud MySQL

Added the connection details of MySQL Google Cloud in the local workbench

Post establishing the connection of MySQL hosted in Google cloud, we executed the below commands for storing the tables as per the dataset we analyzed.







**Reference Architecture for Hybrid Data**

A hybrid data architecture that integrates on-premises and cloud-based resources is what the insurance business ought to implement. This method offers the scalability and flexibility needed to handle different kinds of data and workloads.

Components of the Reference Architecture:

1. On-Premises Data Center: This is like a local storage space where a company keeps its important operational data and key applications.

2. Cloud Data Platform: It's like renting storage space in the cloud (online) for analyzing data, creating visual presentations, and storing huge amounts of data.

3. Integration Layer: This acts as a bridge, allowing data to move smoothly between the local storage (on-premises) and the cloud.

4. Data Governance Platform: This is like a set of rules and tools ensuring that data is secure, follows legal standards, and is consistent across both local and cloud storage.

Advantages:

1. Scalability: It's like being able to stretch or shrink your cloud storage space as needed, depending on how much data you're dealing with.

2. Cost-effectiveness: You only pay for what you use in cloud services, which helps save money.

3. Agility: Think of it as being able to quickly set up new online tools and applications for analyzing data.

4. Security and Compliance: Cloud services come with strong protection measures and meet legal standards.