Name: Zhang, Zejun Date: 12/06/2023

NYU ID: zz4140

Course Section Number: CSCI-GA-2433

Project Part 3

**Total in points** (100 points total):

**Professor’s Comments**:

Physical database

The structured data, comprising confidential customer and company information, will be stored in local MySQL databases. Given the sensitive nature of this data, security was the paramount factor in my decision-making. Consequently, I opted to establish an SQL database locally, rather than relying on cloud storage, to ensure heightened security.

To enhance performance, indexing and partitioning techniques will be utilized. I will provide a detailed explanation of how these techniques are implemented, with specifics outlined in the accompanying SQL code.

Indexing

# Create index

CREATE INDEX CusIndex

ON Customer (SSN, C\_Name);

Create index on SSN, C\_Name can make look up by customer’s name and ssn faster. Name and ssn are commonly used to search a person within a database.

CREATE INDEX AccIndex

ON Account (Account\_name, Companycode);

Create index on Account\_name and Companycode can make look up by account name and company code faster. Company code can be commonly used for users to search an account within a business process.

CREATE INDEX ContractIndex

ON ManagerContract (Contract\_num);

Making look up by contract# faster when users search contracts.

CREATE INDEX AssIndex

ON Associate (Associate\_SSN, Associate\_name);

Same reason as the CusIndex, I created index for both name and ssn.

Partition

In the CUSTOMER and ACCOUNT tables, I implemented hash partitioning to segment the tables into four parts to enhance performance, as these two tables are expected to hold a larger volume of data compared to others.

Machine Learning Model

I switched the machine learning part to chronic kidney disease prediction as the advice by introduction of this part, and the dataset is from Kaggle: https://www.kaggle.com/datasets/mansoordaku/ckdisease/data

I employed a random forest regressor model to forecast the likelihood of specific chronic diseases, including chronic kidney disease. Given the necessity for numerous features in predicting chronic diseases, the random forest approach is well-suited for handling multifaceted features. Additionally, it offers the advantage of being able to indicate the importance of each feature in the prediction process.

Below are the features used to train the model. I have separately trained models for each type of disease to ensure more precise and tailored predictions.

age – age bp - blood pressure sg - specific gravity al – albumin su - sugar

rbc - red blood cells pc - pus cell pcc - pus cell clumps ba - bacteria

bgr - blood glucose random bu - blood urea sc - serum creatinine sod – sodium

The model will predict the probability, and then the quote will be calculated by each weighted probability.

Workflow

