**1.** **Database infrastructure**

The whole infrastructure is constituted of three DBs. All the three DBs are stored in MongoDB. Apache Kafka is used to give insurance agents updates on new data coming in. And SendGrid is used to send emails to customers.

DB1 is the open reference data the insurance company can get from open sources. These reference data, such as the open diabetes patients data and the open cardiovascular disease patients data, are used for building machine learning models to make predictions in the future.

DB2 is the insurance company’s historical completed order data. This data contains all the quotes the company has given in the past, and will be used as a reference for the future quotes. For example, when a new customer comes in, and his or her situation is very similar with a past customer’s, then their quotes might be similar with the consideration of inflation.

DB3 is where the incoming new customers’ information are stored. It contains multiple fields regarding the new customers’ economic status, their interested insurance, and their health situations. The customers who complete the payments will be transferred to the DB2 as completed order data.

**2.** **Appropriate business use cases**

The insurance business has three types of work. The first kind of position is salesperson, who brings the customers into the insurance website. The second kind of position is an insurance agent who provides quotes to the customers. The third kind of position is research analysts, who constantly derive useful open chronic disease patients data, store the data into DB1 and build machine learning models.

When a customer is introduced by a salesperson and comes to the insurance website, he is asked to provide all his personal information, health situation, what kind of insurance product is looking for, and what’s his budget, etc. His data are stored into DB3, and Apache Kafka will notify insurance agents of new data in DB3. The website will tell the customer a quote will send to his email within 24 hours. The insurance agent will access DB3 data and start to estimate the quote with all the information.

The insurance agent will firstly get all the customer’s health data, take them as inputs to the machine learning models created by the company’s research analysts, and form a prediction on how likely this customer will get chronic diseases during the insurance period. Then the agent will compare this customer’s other information with the completed order information in DB2 to see if there’s any similar customers. Based on the similar completed order quotes, this customer’s chronic disease likelihood, and inflation, the insurance agent will finally update the quote to the DB3. Apache Kafka will detect when the quote is formed, which will invoke SendGrid to send the quote to this customer by email.

If the customer accepts the quote and pays for the insurance product, then his data entry will be removed from DB3 and put into DB2 with the purchase date. Otherwise, if the customer hasn’t replied within 1 month, the quote will be deactivated and his data entry in DB3 will be deleted.

**3.** **Machine learning models**

The ml\_model.ipynb contains two machine learning models regarding the prediction of diabetes and cardiovascular diseases. The open patients data are found in<https://github.com/kb22/Heart-Disease-Prediction> (Heart Disease Prediction) and<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download> (Pima Indians Diabetes Database) respectively.

The decision trees are used in both models. And multiple attempts show trees with the max-depth of 8 can provide the best prediction results for both diabetes and cardiovascular diseases predictions.

The final results are: for the diabetes prediction, the in-sample prediction accuracy is 91.8%, while the out-of-sample prediction accuracy is 74.0%, and the three most important indicators are “BMI”, “Pregnancies” and “Glucose”; for the cardiovascular disease prediction, the in-sample prediction accuracy is 100%, while the out-of-sample prediction accuracy is 76.9, and the three most important indicators are “ca”, “thalach” and “oldpeak”.

**4. Application document**

The software github link: <https://github.com/reallljanice/janice_zhou_final-project_su22>

The application is divided into six parts: (1) Initial DB set ups; (2) Machine learning models; (3) User interface; (4) Insurance agent analysis; (5) Quote update; (6) Order transfer

1. Initial DB set ups

At the beginning of the application, the DB2 and DB3 are set up. Since it is a demo, DB2 currently only contains three data entries, which represent successful insurance orders in the past. DB3 is empty, which is waiting for new customers to come in to ask for quotes. For simplification, all the DBs are in the csv.file format in this software demo, but the infrastructure design would be similar to real DBs.

1. Machine learning models

In part 3 of this project, I have developed two decision tree models for the prediction of diabetes and cardiovascular diseases, which are addressed in detail in the third bullet of this report. These two machine learning models are the foundation of this software, since all new customers’ health status will be assessed by these models and the quote prices will be different with different model results.

1. User interface

When new customers are brought by a salesperson into this application, they are required to fill out a form with multiple fields. The fields include the customer’s age, gender, bmi, and a variety of his or her health conditions like resting blood pressure. After the form is filled out, a new data entry will be recorded in DB3, with the “pending” quote, and the insurance agent will be notified by Apache Kafka to analyze this new data entry.

1. Insurance agent analysis

The insurance agent will access DB3 to get the latest new data entry with the “pending” quote. There are three steps in his or her analysis.

Firstly, the agent will predict this new customer’s probability of getting diabetes using the diabetes decision tree model “dt\_d”. For this new customer Clement, the model result shows he will have 100% probability of getting diabetes.

Secondly, the agent will predict this new custtomer’s probability of getting cardiovascular disease using the cardio decision tree model “dt\_c”. For Clement, the model result shows he will also have 100% probability of getting cardiovascular disease.

Finally, the agent will compare Clement’s information with the past completed orders that are stored in DB2. The agent will search if DB2 contains any data with the same age, sex, diabetes probability and cardiovascular probability as Clement’s. The analysis shows Jason has the same health condition as Clement, and he has purchased insurance on 06/16/2020. His insurance price was $1,500.

Considering the inflation and the insurance industry price adjustment, the agent believes $1,700 will be a fair price for Clement.

1. Quote update

Once the agent has concluded a quote of $1,700 for Clement, he or she will update Clement’s data in DB3, changing the “pending” quote fields to “1,700”. Apache Kafka will detect this change, triggering SendGrid to send out an email to Clement and let him know his insurance quote has been formed.

1. Order transfer

After Clement receives the email and sees this insurance quote, he may choose to accept it or decline it.

If he accepts the quote and wants to proceed with the purchase, after his payment is completed, his data in DB3 will be deleted. At the meantime, Clement’s transformed data with only name, age, sex, quote, purchase date, diabetes probability and cardiovascular probability will be transferred into DB2.

If Clement doesn’t accept the quote immediately, then his data will be maintained in DB3 for one month until the quote expires. After one month, his data will be deleted from DB3. If he wants to pursue this insurance again, he has to go back to the user interface and fill out the form again. A new agent will evaluate his situation and give him a new quote, which may be higher or lower than $1,700 depending on his updated health situation and the insurance market. And he will have another one month to decide.

If Clement declines the quote, his data will be deleted from DB3 and his insurance quote will be expired immediately.