Introduction

Our dataset is a selection of columns from open source datasets for data exploration and analysis More detailed column description can find clues from the following url：

<https://archive.ics.uci.edu/ml/datasets/heart+Disease>

This full database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

We perform simple type conversion on the selected columns, and then perform data exploration and analysis.

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

To see Test Costs (donated by Peter Turney), please see the folder "Costs

analyis and feature enginering , choose my columns (target and feature column) to anlyis

''' high-dimensional datasets. In this case, it is better to visualize the correlation in a matrix. Here's how to achieve this using a visualization, I'm using our full Dataset ss you can see from the matrix above, there is a high correlation between heart disease and the first four variables, as well as between alcohol consumption and body fat percentage. This is also in line with our daily common sense: there is no randomness in the patient's community id (patientid), and gender may not necessarily be related to heart disease or diabetes, or if there is, it is not directly related; whether to take painkillers, It has little to do with the pain area and whether you have a heart attack, because it is something that has already happened, and there is no necessary correlation. It is a post-occurrence thing. '''

we need to double check which columns are useless

''' If our dataset has completely positive or negative properties, then the performance of the model is likely to suffer from a problem called "multicollinearity". Multicollinearity occurs when one predictor in a multiple regression model can be linearly predicted by other predictors with high prediction accuracy. This may lead to distorted or misleading results. Fortunately, decision tree and boosted tree algorithms are inherently immune to multicollinearity. When they decide to split, the tree selects only one fully related feature. However, other algorithms like logistic regression or linear regression also do not avoid this problem and you should fix it before training the model

Correlation between variables does not imply causation. Any highly correlated variables should be carefully examined and considered

So we then performed feature analysis on all the data columns, completely random columns, we can treat them as noise columns. We can see that heart disease, diabetes, bmi, drinking or not, and age show great regularity in the data (regularity here refers to non-random, and has certain characteristics of time, distribution, and non-complete randomness) '''

what problem we choose to explore?

''' In fact, this data set can analyze many problems, we choose the most likely problem in data analysis: the correlation between heart disease and other factors. The reason why we give up studying the relationship between diabetes and heart disease, or diabetes and other factors, is because statistical analysis, correlation analysis, and characteristic analysis of each column show the correlation of heart disease '''

Our dataset is over 300M records, and I want to do data explore and machine learning on it, the dataset I can save on google clouds’s Filestoe or Hadoop on Google Cloud. And the data anlysis instance we choose to use Colaboratory, which in our case ,with student identity, it can be free.

Technical Solutions:

Jupyter Notebook + Python/pyspark+ Spark SQL / MLlib + Matplot + pandas  
Cost estimation: I will choose to use Google Colaboratory/ dataset save at google clouds’s Filestoe or Hadoop on Google Cloud.

Architecture Design