GroupProjectBAIT509

Hi there, this is the group project of UBC MBAN 2024' BAIT509.

Project Overview

This project leverages the Chinese Health and Retirement Longitudinal Study (CHARLS) to explore the factors influencing the likelihood of middle-aged and elderly populations in China undergoing regular health examinations. The focus is on understanding how demographic, socioeconomic, and health-related factors contribute to this health-seeking behavior, with an aim to inform policy and improve participation in health screening programs.

Key Methods

- Used DummyClassifier, SVC, DecisionTree, KNN and Logistics Regression as Classifiers
- Used Parameters Optimization for better performance
- Used Cross Validation, Confusion Matrixs to ensure stability and accuracy.

Key Findings

- **Top Influencing Factors** : The decision tree model identified the type of HuKou, community where the respondent lives, and diagnosis of "Emotional and Mental Health Issues" as the top three factors.
- **Data Insights** : Exploration showed significant disparities in health examination participation, influenced by socioeconomic status, urban-rural divide, and specific health conditions.

- **Model Performance** : Logistic Regression emerged as the most effective model, demonstrating the best generalization ability and an acceptable balance between precision, recall, and accuracy.

Limitations and Future Directions

- The reliance on self-reported health indicators may introduce biases. Future studies could integrate objective health metrics for a comprehensive analysis.
- The analysis acknowledges the challenge of capturing the full spectrum of factors affecting health-seeking behavior due to data limitations and the complexity of healthcare access in China.

File Structure

- ---code/: Class proviede test scripts
- ---df.csv: essential database
- ---main_analysis.ipynb: main analysis jupyter notebook
- ---myenv.yml: conda env file

Things for reproduction

1. Using the following command to build the required Python environment: `conda env create -f myenv.yml`
2. Activate the env `conda activate bait509`
3. Read the truncated database through `df.csv` in `main_test.ipynb`
4. Run `main_test.ipynb`
5. Wait for the resuilt, it will take around 5 minutes based on M1 Pro Macbook, time may variate.

About the SQL database(For group members)

Because `ID` is a reserved variable in postgresql, it will be automately wraped with a quote which will cause problem when querying it. So besure to wrap any columns with `ID` as `"ID"` before quering it. And the quote is case sensitive so besure to type correct.

To avoid merge confict, naming the local file as: localtest_*

for instance:

`localtest_mine.ipynb`

this file will be ignored when pull requiest