THE UNIVERSITY OF HONG KONG

DEPARTMENT OF COMPUTER SCIENCE

AI-POWERED INSURANCE RECOMMENDER FOR WEARABLES

LEVERAGING HEALTH DATA TO PERSONALIZE AUTO INSURANCE PLANS

PROJECT PROGRESS UPDATE 1

Date of submission: 7th April 2025

Mentor: Dr. H.F. Ting

Group: MSP24016

Members:

Chan Yik Lun (Alan), 3036035893

Cheng Karen (Karen), 3036199089

Tai Kwong Chuen (Karl), 3036199467

Wong Tak Hei (Aidan), 3036196922

Table of Content

1. Achievements

We have successfully completed the data collection phase of our project, gathering two key datasets: the CASdataset (brvehins1), which provides detailed information on auto insurance, and the Kaggle health kit dataset, which offers important health metrics from Apple Watch and Fitbit users. Following this, we conducted thorough Exploratory Data Analyses (EDAs) for both datasets. For the CASdataset, we performed data cleaning, addressed missing values, and generated summary statistics that highlighted key trends in auto insurance claims and premiums. Similarly, we cleaned and analyzed the health kit dataset, uncovering valuable insights into user health metrics. Across both analyses, we created various visualizations, including histograms, box plots, and correlation matrices, to effectively illustrate relationships between the data points. These efforts allowed us to identify significant correlations that will inform our predictive modeling for insurance premiums moving forward.

1.1. EDA on the brvehins1 Dataset

The brvehins1 dataset originates from CASdatasets and is a Brazilian vehicle insurance dataset, providing detailed records by policy. It includes various risk features related to vehicle insurance but excludes city codes. The dataset contains 1,965,355 vehicle insurance policies from 2011. After addressing missing values, the total number of valid records is reduced to 1,655,869.

Summary of Insights from EDA_brvehins1.ipynb (see attached for details)

a) Vehicle and Premium Information

- The **average year of vehicles** covered by the auto insurance policies is **2005**.
- The **average total premium** for these vehicle insurance policies is **R$3,755**, with most premiums concentrated below **R$1 million**.

b) Claims Analysis

- The most common **types of claims** were:
  - **Partial collision**: 355,598 cases with a total claim amount of approximately R$1.3 million.
  - **Robbery**: 45,892 cases with a total claim amount of approximately R$1.1 million.

c) Correlation Insights

- **Total exposure** shows a relatively high correlation of **0.93** with the total premium.
- Claims related to:
  - **Partial collision** has a correlation of **0.72** with the total premium.

- **Robbery** has a correlation of **0.51** with the total premium.

d) Variability in average premiums

- On average, **males** paid higher premiums **(R$4,113**) compared to females **(R$3,843**).
- Drivers aged **36 to 45** had the highest average premiums at **R$4,493**, surpassing other age groups.
- Excluding other brands, policyholders of **Ferrari** had the highest average premium at **R$9,784**, followed by **Hyundai** at **R$6,348**.
- Policyholders in **São Paulo**, **Rio de Janeiro**, and **Bahia** tended to have relatively high premiums, averaging around **R$5,000** or more.

1.2. EDA on the Kaggle Apple Watch and Fitbit Dataset

The Kaggle health kit dataset includes 6,264 records of individuals who use Apple Watch or Fitbit to track their health metrics, such as the number of steps taken, heart rate, calories burned, distance traveled, heart rate variability, and activity intensity. When trying to establish a correlation between health data and auto insurance premiums, the goal is to identify metrics that reflect risk-taking behavior and those that indicate lifestyle factors. After processing the data and handling missing values, we derived the following insights:

Summary of Insights from EDA_Kaggle.ipynb (see attached for details)

a) User Demographics

- The age distribution shows a strong concentration of younger participants, with **44.68%** aged **18-25** and **34.90%** aged **26-35**, resulting in an average age of **29**. In contrast, older age groups are underrepresented, with only **11.49%** aged **36-45**, **6.63%** aged **46-55**, and **2.30%** over **55**, while there are **no participants under 18**, highlighting a significant focus on younger demographics.
- Gender distribution: approximately **48% male** and **52% female** participants.

b) Heart Rate Trends

- **Young Adults (18-25):**
  - Heart rates increase with physical activity intensity. For instance, the heart rate rises from 79.91 bpm (Lying) to 97.32 bpm (Running 7 METs).
- **Middle-Aged Adults (26-35):**
  - Similar trends are observed, with heart rates increasing significantly during higher intensity activities (e.g. 92.81 bpm for Running 7 METs).

- **Older Adults (36-45 and 46-55):**
    - The heart rates are generally higher for similar activities compared to younger groups, particularly during high-intensity exercises (e.g. 111.02 bpm for Running 7 METs in the 36-45 group).
- **Elderly (>55):**
    - Heart rates in older age groups are lower than those in younger individuals, with the highest recorded heart rate being 77.09 bpm during Running at 5 METs. This seems unusual, we should closely examine the dataset in the next phase for any inaccuracies or outliers that might affect our analysis.

c) Correlation Insights

- The results revealed meaningful correlations among height, weight, and gender.
    - **Height and Gender (0.74)**: This strong correlation suggests that height may significantly differ between genders, likely reflecting biological differences.
    - **Weight and Gender (0.58)**: This indicates a moderate correlation, suggesting that weight also varies by gender, though the relationship is less pronounced than height.
    - **Height and Weight (0.69)**: This strong correlation implies that taller individuals tend to weigh more, which is a common expectation in many populations.

2. Challenges

2.1. Challenge 1: Identifying a Usable Dataset for Predicting Auto Insurance Premiums

One of the primary challenges we faced was identifying a dataset that provides sufficient insights for predicting auto insurance premiums while adhering to privacy regulations. Despite initial difficulties, we successfully sourced a sufficiently large dataset that includes claim history and policyholder demographic details. However, this dataset is geographically constrained to Brazil, which limits the generalizability of the insights to other regions—a limitation we have decided to accept as a compromise for the project.

2.2. Challenge 2: Merging Datasets Without a Unique Identifier

Another significant challenge has been the lack of a unique identifier to merge the brvehin1 dataset with the Kaggle Apple Watch and Fitbit dataset. This limitation prevents a

straightforward integration of the two datasets. Combining these datasets is critical to adding value and enhancing the prediction of total premiums by incorporating health-related data as an additional feature for users. To address this challenge, our next step involves leveraging common data points available in both datasets, such as **driver age** and **gender**. Specifically, we propose using population-level aggregation which involves the following steps:

- Grouping the data into buckets based on age and gender categories.
    - For Health Kit Data: Calculate average health metrics (e.g. median steps, heart rate variability) for each demographic group.
    - For Insurance Data: Calculate average premiums or claim rates for the same demographic groups.
- Merging the aggregated datasets using shared demographic groupings (e.g. "males aged 30-35").
- Training a predictive model by aggregating health metrics for the demographic group as additional features

This approach will help us establish a statistical relationship between the datasets, enabling us to combine them indirectly. By doing so, each bucket, defined by specific combinations of age and gender, will provide statistically significant data that can be used to derive insights and improve premium predictions. While the approach offers a practical solution, it sacrifices individual-level insights and may introduce ecological fallacy (i.e. group-level trends may not accurately reflect individual behavior).

3. Plans for the upcoming weeks

By the next Progress Update scheduled for May 5, we will focus on two critical areas: data preprocessing and the evaluation of the AI model.

During the data preprocessing phase, we will address the challenges of merging and cleansing the datasets, ensuring that all datasets are thoroughly cleaned, normalized, and structured to facilitate accurate analysis. This will include handling missing values, outlier detection, and feature scaling.

In addition, we will conduct a comprehensive evaluation of the AI model to assess its performance and effectiveness. This will involve testing various algorithms, fine-tuning hyperparameters, and validating the model using appropriate metrics. Our goal is to ensure that the model is robust and reliable, capable of delivering insights from the processed data.

Task	Date	Key Milestones	Progress
Detailed Proposal	Mar 10	Finalize project scope, datasets, define model architecture, and outline wearable app scope	Done
Progress Update 1	Apr 7	Complete data collection & EDA	Done
Progress Update 2	**May 5**	**Preprocessing (merging and cleansing), training and evaluating AI models (insurance product classification and premium prediction)**	**In progress**
Interim Report & Presentation	Jun 1	Develop wearable app prototype (Apple Watch/Fitbit interface)	Pending
Progress Update 3	Jun 16	Integrate backend API with wearable app	Pending
Progress Update 4	Jul 7	Conduct user testing and app refinement	Pending
Project Webpage	Jul 15	Deploy webpage with demo, documentation, and results	Pending