

SmartPremium: Using Fitbit Data to Predict Insurance Premiums

Team members	Contributions
Priyadarshini Dhanaraj Muthamil Selvi, Tanishq Dahiya, Sai Srikanth Devasani, Lavesh Eknath Desai	All four members contributed equally to all project components, including design, data processing, modelling, analysis, visualisation, and report writing.

The link for the code:

https://gitlab.computing.dcu.ie/tanishq.dahiya2/smartpremium/-/blob/892e082d52e245f0cb9e35e6bc0994ec6212eba8/Group_10_Final.ipynb

Abstract:

This project develops a health risk scoring framework using real-time Fitbit data (steps, sleep, heart rate), including statistically grounded synthetic demographics. Using these features, K-Means clustering uncovers natural risk groups (Gold/Silver/Bronze), which are subsequently modelled using Logistic Regression and Random Forests. The models achieve strong predictive performance (0.94 and 0.92 accuracy) with clear, interpretable separation between tiers. Results demonstrate the viability of wearable-derived behavioural signals for early health-risk stratification. This approach offers practical value for both insurance underwriting and personalised preventive health guidance

1. Dataset & Problem:

- The objective is to build an interpretable, data-driven health-insurance risk stratification system using wearable data. Unlike traditional actuarial models, wearables provide continuous behavioural signals steps, intensity minutes, resting heart rate (RHR), BMI, sleep hours, and sedentary time that have established links to cardiovascular, metabolic, and mortality outcomes.
- **Dataset:** Public Kaggle Fitbit dataset for 33 individuals of 30 days. Metrics aggregated to user-level: AvgSteps, StdSteps, AvgDistance, AvgCalories, AvgVeryActive, AvgFairlyActive, AvgLightlyActive, AvgSedentary, AvgHeartRate, AvgBMI, AvgSleepHours. Because the original sample was small, we had to generate 1000 synthetic users using a **GaussianCopulaSynthesizer** to preserve feature correlations. Demographic attributes (Age, Sex, Smoker, Children) were simulated using scientific rules from recent studies(e.g., activity declines with age; smokers have poorer sleep/HR; parents sleep less).
- The final modelling dataset contained 1033 users & 17 variables

2. Methods:

1.1. Preprocessing

- Missing values were imputed using science-based conditional rules, avoiding mean imputation's variance shrinkage.
 - Resting Heart Rate: Inverse association with steps, with noise to reflect overtraining cases.
 - Sleep: moderate correlation with activity, while accounting for active individuals with poor sleep.
 - Body Mass Index: activity-linked but allowing metabolically healthy obese cases

1.2. Risk Score Construction

- A composite Health Risk Score was computed from validated factors:
 - Steps, VeryActive minutes, Heart rate, Sleep, BMI, Sedentary time.
- Each was normalised to risk factors (RF=0 best to RF=1 worst). A final insurance-style score added multiplicative modifiers for Age, Smoking, and Sex to produce a FinalRiskScore. This provides an interpretable ground proxy for unsupervised tiering.

1.3. Clustering (Tier Generation)

- We used K-Means ($k=3$, standardised features) identified natural groupings in health behaviour. Clusters were ranked by mean FinalRiskScore and mapped to:
 - Gold (Low)
 - Silver (Medium)
 - Bronze (High)
- Pairplots show visible though moderate separation, especially for HR, BMI, and Steps

1.4. Supervised Models

- To predict tiers, we trained:
 - Logistic Regression
 - Random Forest
- Train/test: 80/20 stratified split.
- Validation: 5-fold stratified CV (accuracy + weighted F1).
- Feature importance was analysed via logistic coefficients and Random Forest Gini importances.

3. Results:

3.1. Model Performance

Model	Accuracy	F1 - weighted
Logistic regression	0.94	0.94
Random Forest	0.92	0.92

3.2. Cross-validation confirms stability:

- Logistic Regression accuracy $\approx 0.95 \pm 0.019$
- Random Forest $\approx 0.935 \pm 0.012$

3.3. Interpretability

- Logistic regression coefficients reveal:
 - Higher AvgVeryActive lowers risk (positive toward Gold).
 - Higher BMI, RHR, and sedentary time push users toward Bronze.
 - Age modestly increases risk tier odds.
- Random forest feature importance reveals that these are the top 3 features according to the model:
 - AvgSedentary 0.292320
 - AvgVeryActive 0.248533
 - AvgSleepHours 0.198202

4. Analysis & Discussion:

4.1. Cluster Characteristics

- Gold users show high steps, good sleep, low sedentary time, healthy BMI, and lower RHR.
- Silver users occupy the broad mid-range.
- Bronze users have poorer metrics across steps, activity, BMI, HR, and sedentary behaviour

4.2. Model Behaviour

Analysis & insights Confusion matrices show that: -

- Most errors occur between Bronze and Silver tiers, which is expected because medium- and high-risk users can have overlapping metric ranges (e.g., borderline steps or BMI). Gold (Low) tier is relatively well separated, with fewer misclassifications into Bronze. This aligns with its more extreme values: high steps, good HR, good BMI, good sleep, good sedentary time.

4.3. Interpretation

The strong performance of linear and ensemble models suggests that:

- Insurance risk from wearables is highly predictable from a small feature set.
- Activity and sedentary behaviour dominate risk tiering, consistent with WHO and BMJ mortality research.

- Combining HR, sleep, and BMI provides robust discrimination of higher-risk individuals.
- Synthetic demographic expansion did not distort patterns; the model remained stable across folds.

4.4. Example Prediction

- A sample profile (8000 steps, 30 min vigorous activity, BMI=24.5, HR=72, sleep=7.5h, sedentary=480 min, age=35) yields:
 - Logistic Regression → Silver (99%)
 - Random Forest → Silver (71%)
- Consensus: Silver (Medium)
- Recommendations align with clinical standards: increase steps, improve sleep, reduce sedentary time.

5. Ethical & Legal Considerations:

- Wearable-derived insurance modelling raises concerns:
 - Data protection: Even synthetic demographics must follow principles of transparency and fairness.
 - No real identities were used; synthetic expansion prevented re-identification.
- Bias: Health behaviour may correlate with socioeconomic status; fairness audits should be included in future work.
- Actuarial fairness vs societal fairness: Risk-based pricing must balance predictive accuracy with ethical constraints to avoid penalising groups with structural disadvantages.
- The project avoids using real claims or sensitive personal data, focusing instead on behavioural signals.

6. Limitations & future work:

- **Limitations**
 - Small original sample: Only 33 real users; synthetic expansion to 1033 amplifies existing patterns and cannot add genuinely new variation.
 - Multicollinearity: Some health features (e.g., AvgHeartRate, AvgBMI, AvgSedentary) exhibit high VIF scores, which can affect linear-model interpretability. Tree models handle this better, but it is still a modelling caveat.
 - Cross-sectional: The analysis is based on aggregated averages per user; it does not capture temporal changes or trajectories in health behaviour.
- **Future Work**
 - Acquire and link real outcome data (claims, diagnoses, events) to validate the constructed risk scores and tiers against actual insurance outcomes.

- Explore longitudinal modelling of daily time-series data (e.g., sequence models) for early detection of risk transitions.
- Investigate fairness and bias across demographic groups if used in any real-world decision-making context.

7. Conclusion:

- This project demonstrates that behavioural metrics from wearables can construct meaningful insurance risk profiles. The combined pipeline science-based risk scoring, unsupervised tiering, and supervised tier prediction achieves strong predictive performance with transparent interpretability.

8. Bibliography:

- [1] Munich Re. “Stratifying Mortality Risk Using Physical Activity as Measured by Wearables.” 2018.
- [2] BMJ Open Sport & Exercise Medicine. “Physical activity and the insurance industry.” 2021.
- [3] World Health Organization. “Physical Activity Guidelines.” 2020.
- [4] Zhang, G. et al. “Cumulative Resting Heart Rate Exposure and Risk of All-Cause Mortality.” Nature Scientific Reports, 2016.
- [5] Ekelund, U. et al. “Dose–response associations between accelerometry-measured physical activity and mortality.” BMJ, 2019.
- [6] Chaput, J-P. et al. “Sleep duration and health.” BMJ, 2020.
- [7] Munich Re. “Heart Rate and Mortality.” 2020.
- [8] Blüher, M. “Metabolically Healthy Obesity.” Endocrine Reviews, 2020.
- [9] Saint-Maurice, P. “Association of Step Count with Chronic Disease.” Nature Medicine, 2022.
- [10] WHO & Insurance BMI Guidelines.
- [11] Coversure. “Risk Assessment in Health Insurance.” 2025.
- [12] Society of Actuaries. “Risk Scoring in Health Insurance: A Primer.” 2016.
- [13] PMC8917048. “A semiparametric risk score for physical activity.”
- [14] Classification in Liability Insurance Using Machine Learning Models. arXiv:2411.00354.