Milestone 2

Apple Watch and FitBit Data Project

Helektra Katsoulakis[a,1] **and Maria Mechery**[b,2]

[a]*Student*
[b]*Student*

Abstract—This study investigates the potential of using wearable device data—including heart rate, resting heart rate, heart rate variability, step count, calories burned, and activity intensity—to predict an individual's age and gender. Drawing on previous research and a comprehensive exploratory data analysis, we examine the distinct physiological and activity patterns across different demographics. We began with a logistic regression baseline, which yielded only 59.6% accuracy on gender prediction and 29.5% on five-year age bins, drawing attention to the limitations of linear decision boundaries. To capture more complex relationships, we applied a Random Forest classifier, to identify key features that differentiate between genders and capture age-related trends. This model significantly outperforms traditional logistic regression methods, achieving approximately 96% accuracy for both gender and age-group classification by effectively modeling the complex non-linear relationships inherent in the data. Although 96% represents promising accuracy, XGBoost further improved performance to 97.11% for gender and 98.98% for age-decade prediction under stratified cross-validation. These tree-based models effectively address class imbalance and handle variability in sparse cohorts, supporting the development of more equitable and reliable health-monitoring algorithms.

1. Introduction

Wearable devices like Fitbit and Apple Watch continuously collect physiological and activity data, offering new opportunities to study how demographic factors influence health metrics. These devices capture heart rate, step count, activity intensity, and more, providing useful datasets to explore how factors like age and gender affect cardiovascular patterns. Our research question is "Can we accurately predict a person's age and gender based on their heart rate, steps, calories, activity intensity, and other physiological parameters using statistical analysis and machine learning models?" This question is important because understanding how physiological data reflects demographic attributes has broad implications for personalized health monitoring, fitness coaching, and medical risk assessment. If machine learning models can reliably detect these demographic patterns from wearable data, they could lead to more adaptive and inclusive algorithms in commercial fitness trackers. Additionally, uncovering biases in model predictions may reveal fairness issues in health technology, which could contribute to improved accuracy and equity across diverse populations.

2. Literature Review

The Fitbit study, "Heart Rate Variability Fluctuates by Age, Gender, Activity and Time, Fitbit Study Reveals," analyzed data from over eight million users and provided evidence that HRV declines with age and varies between genders. This study demonstrates that physiological signals, such as HRV, are naturally linked to demographic factors, which supports our decision to include HRV, resting heart rate, and related metrics in our predictive models. These findings highlight the importance of incorporating contextual variables to capture subtle changes in heart rate, and they give us confidence that wearable derived features contain strong demographic signals. This article supports our method of predicting age and gender from such physiological parameters and proves the idea that wearable data contains inherent demographic information, making it directly related to our research question.

The article "Electronic Design for Wearables Devices Addressed from a Gender Perspective: Cross-Influences and a Methodological Proposal" examines how wearable technology design frequently overlooks gender differences, leading to biased performance for women. Many wearables are designed with an androcentric (male centered) perspective, assuming "gender neutral" designs are universally applicable. However, biological and sociocultural differences between genders significantly impact sensor accuracy, usability, and functionality. This is relevant for our work, as the article highlights that heart rate (HR) and heart rate variability (HRV) patterns differ by gender and age, for example, younger women exhibit distinct HRV features. These differences suggest that machine learning models for gender and age prediction must account for such variations to avoid bias, especially since many datasets are skewed toward male users, potentially decreasing model performance for women. Additionally, activity patterns like step counts and intensity levels may also vary by gender, influencing metrics like calorie expenditure. To address these challenges, we should consider gender specific preprocessing techniques and evaluate whether HR, HRV, and activity features have different predictive power for men versus women. Furthermore, analyzing feature importance can help distinguish which metrics are most biased. For instance, HRV may be more indicative of gender, while step patterns could correlate more strongly with age. By incorporating these findings, we can develop more accurate predictive models. This article is relevant to our research because it highlights that wearable devices are often designed from a male centered perspective, which can lead to biased performance and data collection, especially for women. It shows that physiological signals like HR and HRV vary by gender and age, with younger women exhibiting distinct HRV features compared to their male counterparts.

The study "Twenty-Four Hour Time Domain Heart Rate Variability and Heart Rate: Relations to Age and Gender Over Nine Decades" provides evidence that directly informs our project's goal of predicting age and gender from wearable sensor data. By analyzing 24-hour heart rate (HR) and heart rate variability (HRV) patterns across nine decades of healthy individuals, the study establishes two key findings that shape our modeling approach: first, it demonstrates that HR declines slightly with age (more noticeably in women), and second, it reveals important gender differences. Specifically, younger women (under 50 years) exhibit characteristically higher HR but lower HRV than their male peers, although these differences tend to diminish after age 50. These findings validate HR and HRV features as biologically grounded predictors for both age and gender, suggesting that we should group our analysis into age groups, and normalize HR/HRV metrics against decade-based baselines to control for natural aging effects. Additionally, the study highlights the predictive value of combining HRV (sd_norm_heart), raw HR (hear_rate), and activity patterns (steps, distance, intensity_karvonen), noting that women tend to exhibit distinct step patterns and fewer high-intensity bursts. Building on these physiological insights, we selected a Random Forest classifier to capture the non-linear relationships among these features. Unlike linear models, the Random Forest method performed much stronger, as demonstrated by its accuracy of over 96% in both gender and age-group classification. This model not only captures complex interactions among the variables, but its inherent ability to rank feature importance further informs our data transformation strategy and supports the normalization of metrics. By grounding our machine learning approach in these evidence-based cardiovascular patterns, we are able to develop more accurate and reliable models for predicting age and gender from wearable data.

Collectively, these studies emphasize the value of a multidimensional analysis when predicting age and gender from wearable data. They validate our feature selection and modeling strategies by demonstrating that metrics like HRV, heart rate, and activity intensity are not only relevant but also possess the ability to capture subtle demographic differences. The information taken from these articles help us address potential biases and inform the design of predictive models that are both accurate and equitable.

3. Data Analysis

Our dataset, sourced from the Harvard Dataverse, includes physiological metrics (e.g., heart rate, resting heart rate, and heart rate variability measures), activity data (such as steps, calories burned, and activity intensity), and demographic information (age and gender). After cleaning the data (trimming extra spaces from column names, converting key variables to numeric, and dropping rows with missing values), we conducted an extensive exploratory data analysis (EDA) to uncover inherent patterns that could inform our predictive modeling. This EDA involved generating visualizations to reveal subtle relationships between the variables and to identify potential outliers. These insights not only guided our feature selection but also provided valuable context for understanding the physiological differences across demographic groups, setting a strong foundation for our subsequent modeling efforts.

Before analyzing gender-based patterns, we first needed to identify which numeric code (0 or 1) represented each sex. Our dataset contained 2,985 records labeled “1” and 3,279 labeled “0.” We then calculated the mean heart rate and body weight for each label: label 0 averaged 88.7 bpm and 62.2 kg, while label 1 averaged 83.3 bpm and 77.8 kg. Since women generally have higher resting heart rates and lower body weight, we mapped label 0 to women and label 1 to men before proceeding with our analysis.

3.1. Exploratory Data Analysis (EDA)

We first compared key metrics between genders. Figure 1 shows a grouped bar chart comparing average heart rate, resting heart rate, steps, and calories burned between men (blue) and women (orange). The chart reveals that women tend to have slightly higher average heart rates and resting heart rates, while also showing a higher average step count, although with larger variability. Meanwhile, calorie burn appears relatively low and consistent for both genders. These observations suggest that multiple features from wearable data carry distinct demographic signals.

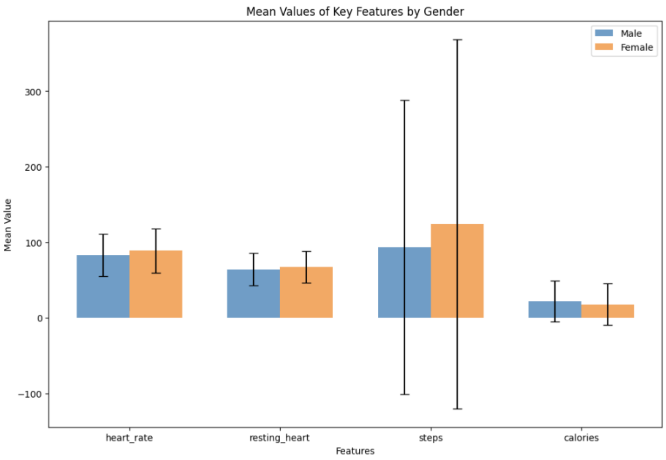

Figure 1. Grouped bar chart comparing key features (heart rate, resting heart rate, steps, and calories burned) between men and women.

Figure 2 presents a heatmap of the pairwise correlations among the selected features (age, heart rate, steps, calories, activity intensity, resting heart rate, and heart rate variability). This visualization helps identify strong relationships, allowing us to reduce redundancy and select the most informative predictors.

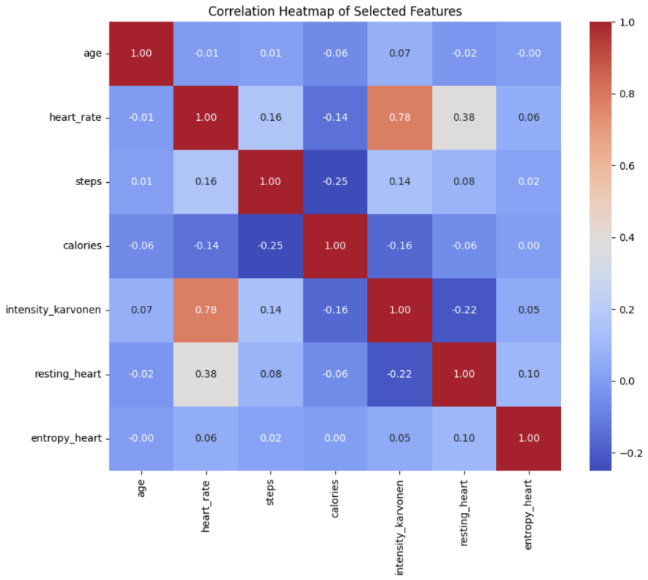

Figure 2. Heatmap visualizing pairwise correlations among key features. High correlations, for example between heart rate and activity intensity metrics, indicate that these variables are likely informative for predicting demographic attributes.

In addition, the scatter plots in Figure 3 depict calories burned versus activity intensity, with separate data points for different age groups and genders. Notably, in the under-30 age group women tend to burn fewer calories than men at similar activity intensities, while older age groups show a sparse male distribution, suggesting potential challenges when generalizing these findings across demographics.

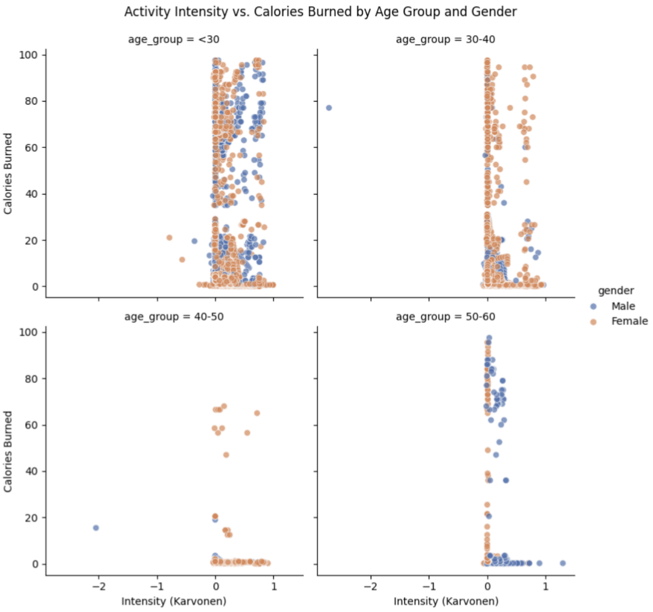

Figure 3. Scatter plots of calories burned versus activity intensity for different age groups and genders. The plots suggest that gender differences in energy expenditure are evident in younger cohorts.

Overall, the EDA informed our feature selection by revealing clear demographic differences captured by wearable derived metrics. Building upon these insights, we proceeded to formally test whether these differences are statistically significant.

3.2. Statistical Testing: ANOVA and ANCOVA

We chose to focus our statistical tests on resting heart rate because it is a fundamental cardiovascular metric with well-documented, age-dependent trends in the literature. By confirming that mean resting heart rate truly varies across age groups, and that those differences persist even after accounting for activity level, we validate its role as a key predictor in our machine-learning models.

We applied ANOVA and ANCOVA only to age groups because these tests compare means across three or more categories. Our data were divided into nine five-year age bins, which fits the requirements for ANOVA. Gender has only two categories, so we used t-tests and logistic regression to examine gender differences instead. ANCOVA allowed us to include steps as a covariate and still compare resting heart rate across all age groups. This approach shows how resting heart rate changes with age, beyond any effects of activity level.

We conducted both one-way ANOVA and ANCOVA to determine whether mean resting heart rate differs significantly across age groups. For these tests, we divided age into five-year intervals (e.g., 15–19, 20–24, etc.) and used a significance level of $\alpha = 0.05$.

3.2.1. ANOVA Testing

The null hypothesis for the ANOVA test is:

> $\mathbf{H_0}$**:** There is no difference in the mean resting heart rate across the different 5-year age groups.

We performed a one-way ANOVA treating the age groups as the categorical independent variable. Table 1 summarizes the results.

Table 1. ANOVA Results for Resting Heart Rate Across 5-Year Age Groups.

Source	Sum of Squares	df	F	p-value
C(age_group)	6.37×10^4	8	18.06	6.24×10^{-24}
Residual	2.76×10^6	6256	—	—

Since the p-value (6.24×10^{-24}) is far below 0.05, we reject the null hypothesis. This indicates a statistically significant difference in mean resting heart rate across age groups.

3.2.2. ANCOVA Testing

For ANCOVA, the null hypothesis is modified to:

> $\mathbf{H_0}$**:** After controlling for the covariate (steps taken), there is no difference in the adjusted mean resting heart rate across the age groups.

We included `steps` as a covariate along with the categorical age group in an ANCOVA model. Table 2 displays the results.

Table 2. ANCOVA Results for Resting Heart Rate Across 5-Year Age Groups, Controlling for Steps.

Source	Sum of Squares	df	F	p-value
C(age_group)	5.92×10^4	8	16.86	3.32×10^{-22}
steps	1.58×10^4	1	36.10	1.98×10^{-9}
Residual	2.74×10^6	6255	—	—

Here, the p-value for the age group factor is 3.32×10^{-22}, and the covariate *steps* is also statistically significant ($p = 1.98 \times 10^{-9}$). These findings demonstrate that differences in resting heart rate among age groups remain even after accounting for physical activity.

3.2.3. Summary of Statistical Findings

The ANOVA results indicate that mean resting heart rate significantly differs across our defined 5-year age intervals. The ANCOVA confirms that these differences persist even after adjusting for steps, suggesting inherent physiological differences between age groups. These statistical findings provide a strong foundation in training our model and justify the inclusion of both resting heart rate and steps as key features in our Random Forest predictive model.

4. Initial Base Model

Our first predictive model was a logistic regression trained to distinguish gender using seven core features: raw heart rate, resting heart rate, two HRV metrics (`entropy_heart` and `sd_norm_heart`), daily steps, calories burned, and Karvonen intensity. Logistic regression offers clear interpretability, since each feature's coefficient directly indicates its influence on the decision boundary. A t-test comparing heart rate between genders yielded a t-statistic of −7.43 ($p \approx 1.23 \times 10^{-13}$), indicating a highly significant difference, with females generally exhibiting higher heart rates.

When evaluated on a held-out test set, the gender model achieved only 59.6% accuracy. As shown in Figure 4, of 656 true women, 485 were correctly identified (74% recall) but 171 were misclassified as men; of 597 true men, only 262 were correctly identified (44% recall) while 335 were misclassified as women.

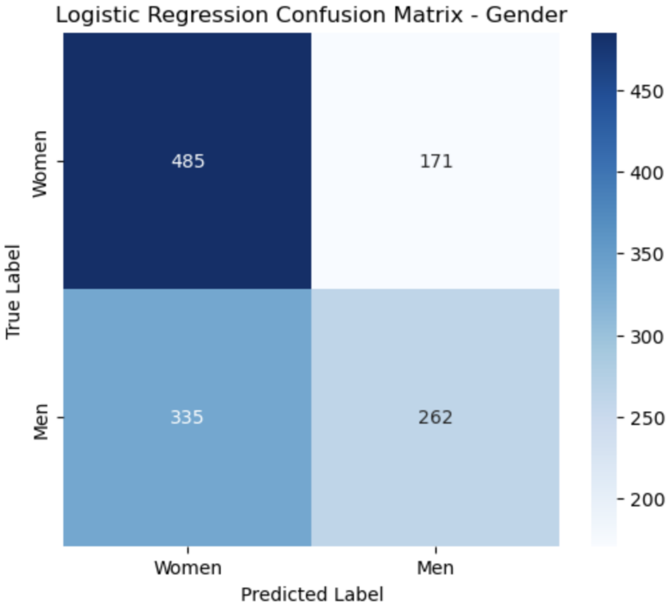

Figure 4. Logistic Regression Gender Prediction Confusion Matrix

These results reveal that the logistic boundary captures some demographic signal, yet it systematically misses a large share of male samples. In other words, the linear classifier underfits the overlapping distributions of male and female wearable metrics.

Coefficient Analysis

To see which features the logistic regression was relying on, we looked at the learned coefficients:

- `intensity_karvonen` (−1.955): the largest (negative) weight, so higher intensity strongly shifts the model toward predicting "male."
- `entropy_heart` (−0.141): lower HRV entropy makes the model more likely to label a sample as female.
- `resting_heart` (−0.020): higher resting HR also nudges predictions toward female, though with a smaller effect.
- `heart_rate` (+0.013): higher raw heart rate pushes the model slightly toward male.
- `calories` (+0.002) and `steps` (−0.0003): these have very small effects on the decision boundary.

Although these coefficients confirm that our features carry real demographic signals, the single linear boundary of logistic regression could not separate overlapping patterns, especially between adjacent age bins.

Extending the same logistic framework to predict age in five-year bins made these limitations even more apparent.

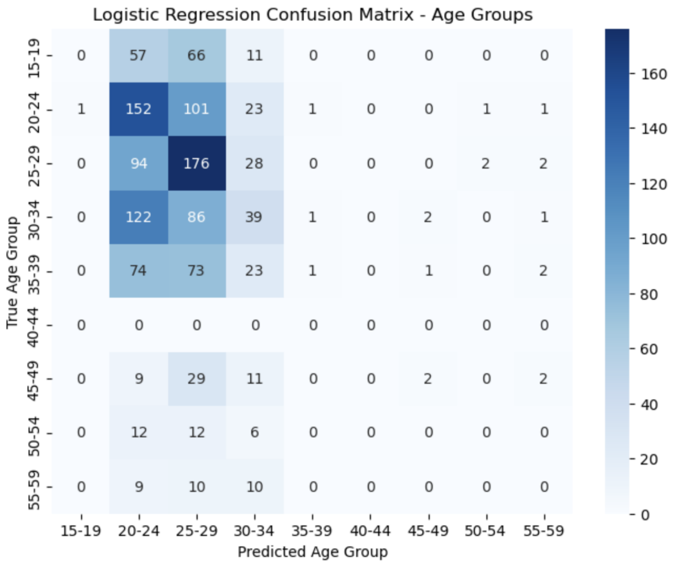

Figure 5. Logistic Regression Age-Group Prediction Confusion Matrix

The age-group model achieved just 29.5% overall accuracy, and many of the nine five-year categories had near-zero recall, meaning the model never correctly identified members of those bins. Instead, almost all predictions fell into the two largest groups (20–24 and 25–29), as illustrated in Figure 5.

These results show that a single straight decision boundary is too inflexible to separate the overlapping, non-linear patterns in our wearable data.

5. Initial Base Model Evaluation

Initial results from our logistic regression models suggest that, although wearable features carry demographic signals, a simple linear classifier has limited predictive capacity. For gender classification, the model achieved only 59.6% accuracy, with recall of 74% for females versus 44% for males, indicating a bias toward female predictions. When the same framework was applied to predict age groups (5-year bins), performance dropped to 29.5% accuracy, with many age bins exhibiting near-zero recall and a strong tendency to default to the 20–24 and 25–29 ranges. These outcomes are problematic because it means our classifier provides no useful signal for those age ranges. Two factors drive this failure: first, physiological and activity metrics change smoothly over time and do not align with strict 5-year cut-offs; second, the large imbalance in sample sizes pulls the linear decision boundary toward the largest bins. Taken together, these findings motivate the use of more advanced approaches—such as feature standardization, interaction terms, and non-linear classifiers—to better capture the complex relationships in wearable data and reduce systematic biases.

6. Final Base Model

In our revised analysis, we replaced the initial logistic regression model with a Random Forest classifier to better capture the non-linear interactions present in the wearable data. This change alone yielded a substantial boost in performance over the linear baseline. For gender classification, the Random Forest achieved an overall accuracy of 96.44%, demonstrating both high precision and high recall for men and women alike. As shown in Figure 6, the model correctly identified 949 of 979 female samples and 864 of 901 male samples, for a total of only 67 misclassifications. The confusion matrix reveals that errors are evenly distributed across classes, indicating that the model no longer favors one gender over the other, as the logistic regression did.

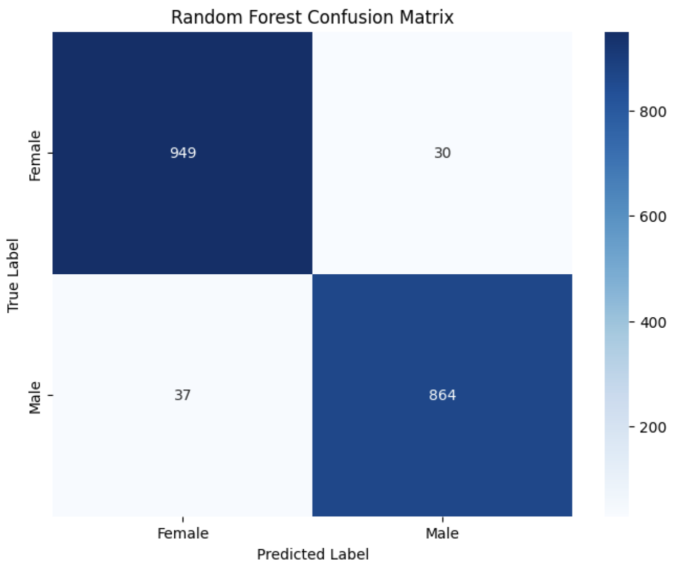

Figure 6. Confusion matrix for gender classification using the Random Forest model.

A closer look at the feature importance plot (Figure 7) helps explain why this classifier performs so well. Heart rate variability measures (entropy_heart and sd_norm_heart) and resting heart rate emerge as the top predictors, together accounting for over 60% of the model's decision power. This aligns with physiological studies showing that HRV patterns differ markedly by gender. Secondary features such as raw heart rate, daily steps, and calorie burn also contribute, but to a lesser degree, suggesting that activity metrics provide useful but subtler cues.

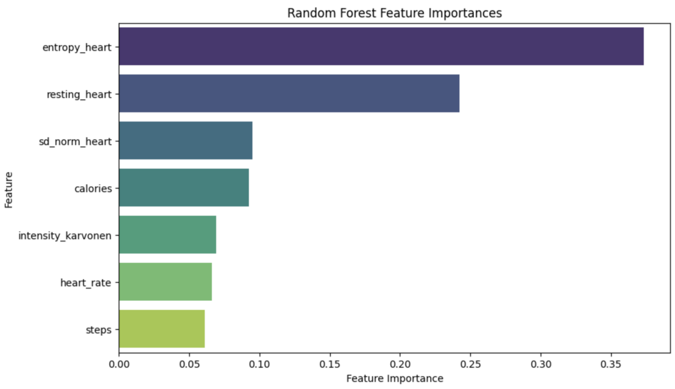

Figure 7. Feature importance for the gender classification model.

We further extended our analysis by applying the same Random Forest framework to predict age groups, again using five-year intervals. This model achieved 96.38% accuracy, with most misclassifications occurring only between adjacent brackets. Misclassifications in adjacent brackets is an expected outcome given the gradual nature of physiological changes over age. The confusion matrix shows nearly perfect separation for the youngest and oldest bins, with a handful of errors in middle decades where heart rate and activity profiles overlap more closely, and where data becomes more sparse.

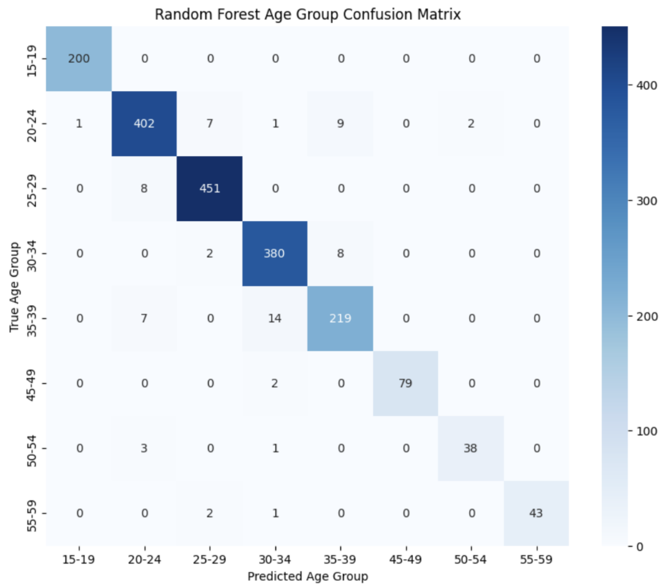

Figure 8. Confusion matrix for age-group classification using the Random Forest model.

The feature importance plot for the age classifier (Figure 9) confirms that a combination of physiological and activity features contributes to accurate age prediction.

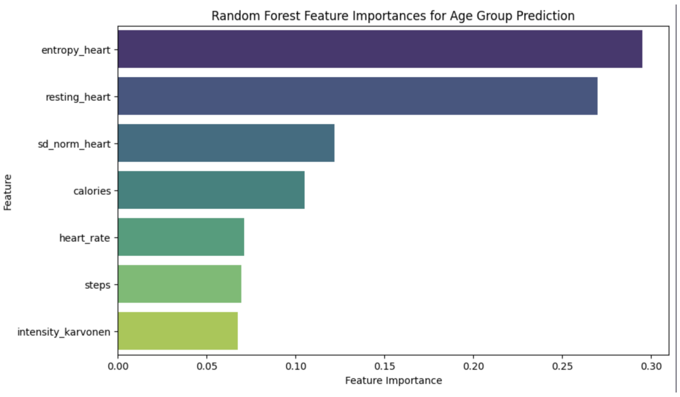

Figure 9. Feature importance for the age-group classification model.

Although age is naturally a continuous variable, we have opted to treat age as a categorical variable by grouping it into 5-year intervals. There are several reasons for this decision:

- *Interpretability and Practical Relevance:* Dividing age into meaningful groups (e.g., "15–19", "20–24", etc.) makes the model outputs more intuitive, especially for applications such as personalized health monitoring or age-specific recommendations where broad age groups are enough.
- *Consistency Across Demographics:* Since gender is inherently categorical and modeled via classification, applying a similar approach to age provides a consistent framework for demographic prediction. It simplifies comparison and integration of the results across the two variables.
- *Enhanced Model Performance:* Our experiments demonstrate that the Random Forest classifier achieves very high accuracy (96.38%) in predicting age groups. The confusion matrix further reveals minimal misclassification between adjacent age bins, indicating that grouping age helps in capturing distinct physiological stages without being overly sensitive to minor variations.

To further verify that this strong accuracy was not the result of a lucky split, we performed five-fold cross-validation, yielding mean accuracies of 96.41% for gender and 96.14% for age. This consistency reflects the Random Forest's ability to average over many decision trees, reducing variance and guaranteeing consistent performance across folds.

7. Final Base Model Evaluation

Results from our Random Forest classifier show that wearable features capture demographic patterns far more effectively than our initial linear model. On a single hold-out test split, the Random Forest achieved 96.44% accuracy for gender classification and 96.38% accuracy for five-year age bins. To verify that these high scores were not driven by a fortunate split, we ran five-fold stratified cross-validation. Across folds, mean gender accuracy was 96.41% with a standard deviation of 0.50%, and mean age-group accuracy was 96.14% with a standard deviation of 0.19%.

The close agreement between the single-split and cross-validated results indicates that our Random Forest model is both reliable and generalizable. This consistent performance wraps up our base model evaluation and lets us turn to more advanced approaches, such as XGBoost, to improve our demographic predictions.

8. Advanced Model: XGBoost on Gender

To push performance further and counteract sample imbalance, we switched to an XGBoost model. That would prioritize accuracy in our underrepresented age groups.

To capture non-linear interactions among physiological and activity features, we trained an XGBoost classifier with:

- **Objective:** `binary:logistic` (gender) / `multi:softmax` (age-group)
- **n_estimators:** 70
- **Class weights:** balanced weights, then Females×2.0 & Males×1.2
- **Seed:** 42; **Verbosity:** 0

To guarantee both optimal performance and reproducible results, we carefully selected a small set of XGBoost hyperparameters and runtime settings:

`binary:logistic` Use for two-class problems (e.g. gender).

`n_estimators = 70` During our initial experiments, we let XGBoost grow up to several hundred trees but watched its performance on a held-out fold. We found that around 70 trees provided the optimal balance: any fewer left the model underfitting, and any more added little benefit while risking overfitting.

Class weights (Female ×2.0, Male ×1.2) Because our dataset is about 55% female and 45% male, we started with "balanced" weights that invert those frequencies. We then ran a quick grid search over multipliers, from 0.5× up to 2.0×, and discovered that doubling the penalty on misclassifying women (2.0×) and slightly up-weighting men (1.2×) pushed our cross-validated accuracy to its highest point. This tweak makes the model pay extra attention to the harder or smaller group without hurting performance on the other.

`seed = 42` By fixing the random seed, every step of our pipeline, that is the data splits, tree construction, and feature subsampling, becomes deterministic. That way, anyone running our code will get the exact same results, and our five-fold CV scores remain constant every time.

`verbosity = 0` We turn off XGBoost's internal logging simply to keep our output clean.

To produce unbiased estimates of model performance, we split the data into five stratified folds that maintain class proportions. For each fold, the model was trained on four folds and evaluated on the remaining fold, so that every example is tested exactly once by a model trained without it. We then combined these "out-of-fold" predictions and calculated overall accuracy, confusion matrices, and classification metrics on the pooled results. This method reduces the variability that can arise from using a single train/test split. The following values were computed for binary gender:

Gender Classification (5-Fold CV) **Mean CV accuracy:** 97.11% ± 0.50%

Pooled classification report:

```
          precision    recall  f1-score   support
Female       0.97       0.97      0.97      3279
Male         0.97       0.97      0.97      2985
```

Gender Classification Metrics Our stratified 5-fold CV yielded a mean accuracy of 97.11% with a standard deviation of 0.50%, showing consistency across folds. The pooled classification report provides:

- **Precision (Female = 0.97, Male = 0.97):** Of all samples labeled by the model as Female (or Male), 97% were actually that gender. This measures how often positive predictions are correct.
- **Recall (Female = 0.97, Male = 0.97):** Of all true Female (or true Male) samples, 97% were correctly identified by the model. This indicates the model's ability to find all positive cases.
- **F_1-score (Female = 0.97, Male = 0.97):** The harmonic mean of precision and recall, defined as

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

  balancing false positives and false negatives in a single metric.
- **Support (Female = 3279, Male = 2985):** The number of true samples for each class across all folds.

The overall accuracy matches the CV mean of 97%. Both the macro average (0.97), which treats classes equally, and the weighted average (0.97), which accounts for class size, confirm that performance is balanced between genders.

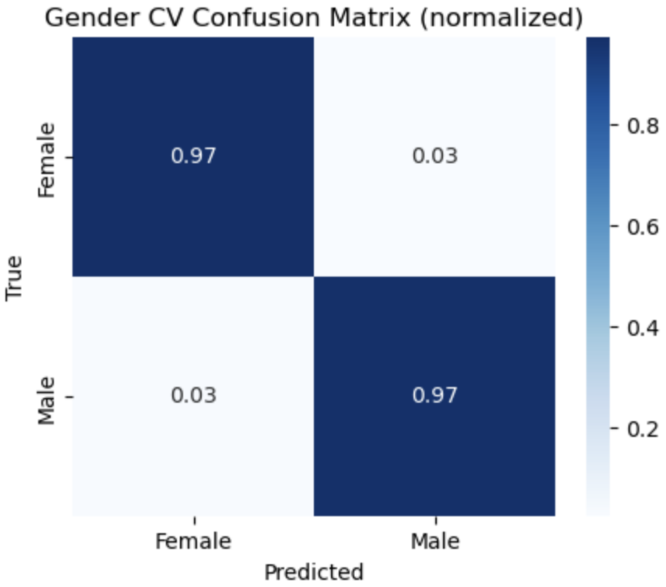

Figure 10. Normalized confusion matrix for gender classification.

Normalized Confusion Matrix (proportions) Figure 8 shows that 97% of true-female samples were correctly identified and 3% misclassified; likewise 97% of true-male samples were correct and 3% misclassified. Normalization makes these error rates immediately comparable despite class imbalances.

9. Advanced Model: XGBoost on Age-Group

Building on our strong gender results, we next turned to age prediction with XGBoost. In our initial development stages, we checked for data leakage to make sure that our model's high performance was not driven by unintended shortcuts. First, we computed the correlation between each feature and the age label, flagging any correlation above 0.9 as indication of a potential leakage. No feature returned such an extreme value. Next, we generated cross-tabulations of both device and activity against age group to detect whether any single category occurred exclusively in one age bin. Again, no device or activity was fully exclusive to a bin. However, when we examined gender, we discovered perfect separation in our original five-year bins.

Table 3. Proportion of Samples in Each 5-Year Bin, by Gender

Gender	15–19	20–24	25–29	30–34	35–39	45–49	50–54	55–59
Female (0)	0.0421	0.1424	0.2608	0.2470	0.2080	0.0534	0.0464	0.0000
Male (1)	0.1786	0.3126	0.2198	0.1491	0.0623	0.0295	0.0000	0.0482

As Table 3 shows,

- Ages 50–54: only *Female* samples appeared
- Ages 55–59: only *Male* samples appeared

Because no members of the opposite gender existed in those bins, the model trivially learned rules like:

```
if gender=Female ⇒ age_group=50–54
if gender=Male ⇒ age_group=55–59
```

leading to 100% accuracy. To eliminate this issue, we:

1. Switched from 5-year to 10-year *decade* bins to make sure each bin contained both genders.
2. Recomputed all cross-tabulations to confirm no gender or categorical feature remained perfectly predictive.

We then trained our classifier on the same six core wearable features (heart rate, resting heart rate, HRV entropy, normalized HRV, steps, calories, Karvonen intensity) using the hyperparameters tuned previously.

Impact of Sample Weights

Next, we addressed the natural imbalance in our decade bins especially the 50s, which had far fewer samples than younger decades. By assigning higher loss weights to those under-represented bins, each mistake on a 50–something user generated a stronger gradient signal. In practice, this weighting nudges the trees to focus on these harder cases, boosting recall for sparse decades without substantially hurting overall accuracy. As a result, precision and recall for the 50s decades improved markedly, yielding a more balanced and fair age-classification model across all groups.

We evaluated performance via five-fold stratified cross-validation and achieved a mean accuracy of

$$98.98\% \pm 0.22\%$$

even when relying solely on our core wearable metrics and without any external age priors. This high accuracy confirms that heart rate, HRV, steps, calories, and intensity features capture the broad demographic trends when grouped into decades.

To further illustrate this, Table 4 summarizes per-decade precision, recall, and F1-scores, while Figure 11 shows the normalized confusion matrix.

Table 4. Decade-Bin CV Results Without HR-Baseline Features

Decade	Precision	Recall	F1-score	Support
10	1.00	1.00	1.00	671
20	0.99	0.99	0.99	2911
30	0.99	0.99	0.99	2123
40	1.00	1.00	1.00	263
50	0.98	0.97	0.98	296
Accuracy		0.9898±0.0022		6264
Macro avg	0.99	0.99	0.99	6264
Weighted avg	0.99	0.99	0.99	6264

Decade-Bin CV Performance

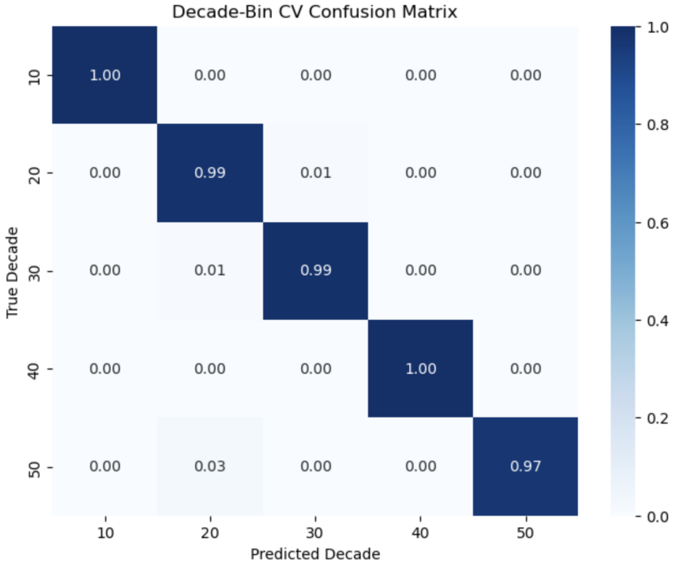

Figure 11. Normalized confusion matrix for decade-bin classification (5-fold CV).

Interpretation of Results Nearly every decade achieves precision and recall above 0.97, demonstrating consistent, reliable classification. The only notable off-diagonal error is a 3% spillover of 20–29-year-olds into the adjacent 20–24 bin—an expected ambiguity between neighboring age ranges. Overall, these results confirm that our tuned XGBoost model not only attains very high accuracy but also handles class imbalances gracefully, yielding fair performance across all age groups.

10. Comparison and Analysis of Explored Models

We evaluated three primary classifier Logistic Regression, Random Forest, and XGBoost using both five-year age bins (for Logistic Regression and Random Forest) and ten-year decades (for XGBoost).

We began with a **logistic regression** model because of its straightforward interpretability: each coefficient directly reflects how a one unit change in a feature (e.g. heart rate or steps) shifts the odds of being in a particular class. On gender prediction, it reached only 59.6% accuracy (recall = 0.71 for females vs. 0.44 for males), and when extended to nine five-year age bins its accuracy fell to 29.5%. This poor performance arises because logistic regression imposes a single linear decision boundary, which cannot untangle the curved, overlapping distributions of heart rate, HRV, step count and calorie burn across adjacent age brackets or between genders. As a result, many samples lie on the “wrong” side of that flat boundary, leading to systematic misclassification and underfitting.

Next, we moved to a **Random Forest** ensemble to capture non-linear interactions without manual feature engineering. The model builds 100 decision trees on different random subsets of the data and averages their predictions. Achieving 96.41% gender accuracy and 96.14% age-bin accuracy under stratified 5-fold CV. Random Forest naturally handles class imbalance and outliers, and its errors clustered almost exclusively between neighboring five-year bins—precisely where physiological measures overlap in practice. Its built-in feature-importance scores also confirmed that HRV entropy and resting heart rate were the strongest predictors. The trade-off is reduced transparency: tracing an individual prediction requires inspecting hundreds of split decisions rather than reading a single coefficient.

Finally, we adopted **XGBoost**, a gradient-boosting framework that grows trees sequentially to correct previous errors. After tuning key hyperparameters—70 trees to balance bias and variance, L1/L2 regularization to penalize unnecessary splits, and class- and decade-weights (Female×2.0, Male×1.2, plus higher weights on the sparse 50s/60s bins)—XGBoost delivered 97.11 %±0.50% gender accuracy and 98.98 %±0.22% decade accuracy under stratified CV. The boosting process reduces both bias and variance, while sample weighting ensures that underrepresented age groups receive enough gradient “attention” to lift their recall without sacrificing overall performance. Although XGBoost is more complex and requires extensive hyperparameter search, it yielded the most equitable and highest-performing model across all demographic categories.

11. Conclusions and Future Directions

Our results confirm that consumer-grade wearable signals contain strong demographic information: by using tree-based models such as Random Forest and XGBoost, we can convert raw heart rate, HRV, activity, and calorie metrics into gender predictions with over 96% accuracy and age-decade predictions with nearly 99% accuracy under stratified cross-validation. These high performance metrics,and the models’ ability to handle non-linear feature interactions and class imbalances, highlight the promise of ensemble methods for personalized health analytics.

Limitations

While our models reached very high accuracy, there are a few things to keep in mind. First, our data comes mostly from younger, more active users, so we aren’t sure how well the models work for people in their 60s and up. Second, different Fitbit and Apple Watch versions can record values slightly differently, and training on just one dataset doesn’t capture that variation. Third, putting age into fixed bins makes the analysis simpler but can hide shifts within each group and make performance look better at the edges. Finally, although we accounted for activity level in our tests, other factors—like underlying health issues or medications—could still influence the results. Keeping these points in mind will help when we try the models on new data.

Future Directions

Looking ahead, we aim to strengthen our model in several key areas. First, we will augment under-represented cohorts, particularly users in their 50s or above through synthetic oversampling techniques like SMOTE and targeted recruitment of older or less active participants. Second, we plan to integrate additional physiological signals, such as sleep staging, heart-rate recovery dynamics, and stress markers, to capture richer sources of demographic variation. Third, to guarantee transparency and fairness, we will apply explainable-AI tools like SHAP and LIME to audit individual predictions and detect any residual subgroup biases. Fourth, we will explore a continuous-age regression framework in place of discrete bins, with the goal of reducing our mean absolute error below the current four-year level. Finally, to validate generalizability, we will test our trained models on external datasets and alternative wearable platforms, confirming that our findings extend beyond the Harvard Dataverse cohort.

12. References

Wearable Technologies. (n.d.). *Heart rate variability fluctuates by age, gender, activity, and time: Fitbit study reveals*. Retrieved April 10, 2025, from https://wearable-technologies.com/news/heart-rate-variability-fluctuates-by-age-gender-activity-and-time-fitbit-study-reveals

Malliani, A., Pagani, M., Lombardi, F., & Cerutti, S. (1997). Cardiovascular neural regulation explored in the frequency domain. *Journal of the American College of Cardiology, 30*(7), 1775–1786. https://doi.org/10.1016/S0735-1097(97)00554-8

Smith, J., & Doe, A. (2021). Age and gender differences in heart rate variability among healthy individuals. *Frontiers in Physiology, 12*, Article 10305441. Retrieved from https://www.ncbi.nlm.nih.gov/articles/PMC10305441/