Heart Failure Risk Prediction in Smurf Society

Final Report

Beatriz Ferreira & Joyce Ghosn

Introduction

Cardiovascular diseases are an emerging concern in Smurf society, driven by both lifestyle changes and biological risk factors. Doctor Smurf collected a comprehensive dataset, including clinical measurements, lifestyle variables, demographic information, and MRI-like heart images. The aim of this report is to analyse key determinants of ten-year heart failure risk, predict individual risk using machine learning models, and identify high-risk subgroups.

This study follows a four-part pipeline: (1) preprocessing and linear modelling of tabular data, (2) nonlinear regression models with advanced feature selection, (3) integration of image-derived features into a multimodal model, and (4) interpretability and societal insights into heart disease risk. Each stage builds upon the previous one, providing both predictive performance and actionable understanding of Smurf health.

Part 1: Dataset Preparation and Linear Model

1.1. Dataset Preparation

The dataset consists of 14 clinical, behavioural and demographic features describing 1500 Smurfs. No missing values were found, and the column `img_filename` was removed, as image analysis is handled only in Part 3. Most outliers are moderat, not extrem so we dont do anything about that.

We grouped the features into:

- **Numerical**: age, blood pressure, calcium, cholesterol, hemoglobin, height, potassium, vitamin D, weight.
- **Ordinal**: sarsaparilla, smurfberry liquor, smurfin donuts (five-level ordered scales from "very low" to "very high").
- **Categorical**: profession (nominal).

We applide a **ColumnTransformer** to preprocess the data, ordinal variables were mapped to integers from 1 to 5, encoded with Ordinal Encoder while `profession` with one-hot encoding.. After preprocessing, the dataset contained **18 numerical columns** ready for model training.

A strict **data-splitting protocol** was followed: after preprocessing with the *exact* encoders from Part 1, the 1000-row training set was split into an 80% inner training set and a 20% inner validation set, with all feature and model selection performed solely on the inner training set, while the 500-row test set was used only once for final evaluation, preventing any leakage and ensuring a valid generalization estimate.

Linear regretion models are sensitive to differeces in features scale, so we applided **StanderdScaler**.

1.2. Feature Selection

We performed univariate feature selection by computing Pearson correlations on the training subset, ranking features by absolute correlation, and retaining only those with $|\text{corr}| > 0.05$ to eliminate weak predictors.

This process reduced the input space to 13 meaningful features: *blood pressure, cholesterol, weight, smurfin donuts, profession__administration and governance, age, sarsaparilla, profession__food production, hemoglobin, profession__services, profession__resource extraction, height, smurfberry liquor*, which were used as the final set for model training.

1.3. Model Comparison

We compared Ordinary Least Squares Regression, Ridge, and Lasso models on the inner validation split after preprocessing and feature selection, training each model exclusively on the inner-train subset and selecting the one with the lowest validation RMSE. Ridge ($\alpha = 100$) achieved the best validation performance (RMSE $= 0.0506$), evaluated it once on the unseen test set, obtaining a final **RMSE of 0.05596**, confirming good generalization and no overfitting

Chosen Model: Ridge Regression with $\alpha = 100$.

Part 2: Nonlinear Models and Advanced Feature Selection

We investigate whether nonlinear models: Random Forest, Gradient Boosting, k-Nearest Neighbors, and MLP, can outperform the linear baseline. Three feature-selection strategies (Mutual Information, Sequential Forward Selection, and Random Forest importances) were applied only on the inner-train split(80 % of the training set). Also, as Part 1: The preprocessed training data was split into 80 % for feature selection and model training and 20% for validation, while the test set remained untouched to preserve a clean final evaluation.

2.1. Feature-Selection Methods

2.1.1. Mutual Information (Filter Method)

Mutual Information measures non-linear dependence between the general statistical dependence between each feature and the target. Applied to the inner split, the top 10 features: $mi_{\text{top10}} = \{$`blood_pressure`, `weight`, `cholesterol`, `age`, `sarsaparilla`, `potassium`, `profession__craftsmanship`, `smurfin_donuts`, `profession__food_prod` `profession__administration`$\}$.

Interpretation. Blood pressure exhibits the highest MI (≈ 0.23). Several Part 1 predictors reappear (weight, cholesterol, age), supporting their robustness. New variables such as potassium also emerge, indicating nonlinear relationships not captured by correlation alone.

2.1.2. Sequential Forward Selection (Wrapper Method, SFS + MLP)

We applied SFS because wrapper methods evaluate feature subsets using a real nonlinear estimator, in our case an MLP, allowing the selection process to capture interaction effects that filter methods cannot. SFS treats feature selection as part of model training: it starts with no features, adds the one that most improves CV-RMSE, and continues until a predefined subset size is reached. Selecting 8 features provides a practical balance: SFS is computationally expensive, nonlinear models risk overfitting with too many inputs, and the chosen range (5–10 features) remaining expressive enough to reveal meaningful nonlinear patterns.

The SFS-selected subset was: $\text{sfs}_{\text{top8}} =$ `age, blood_pressure, profession__administration_and_governance, profession__craftsmanship, profession__food_production, profession__manufacturing, profession__resourc` `profession__services`.

Interpretation. The selected subset is dominated by occupation-related features, indicating that the MLP detected nonlinear interactions between profession and health variables. This outcome reflects how wrapper methods favor features that improve the specific model's performance rather than those that are individually strongest.

2.1.3. Random Forest Importances (Embedded Method)

We used Random Forest as an embedded nonlinear selector because it identifies important features directly during model training. By evaluating how much each variable reduces prediction error across many trees, the method highlights predictors involved in nonlinear patterns and interactions that simpler filters cannot detect.

We obtain: $rf_{\text{top10}} = \{$`blood_pressure`, `potassium`, `hemoglobin`, `weight`, `age`, `cholesterol`, `vitamin_D`, `height`, `sarsaparilla`, `calcium`$\}$.

Interpretation. RF emphasizes biologically meaningful biomarkers (potassium, hemoglobin, calcium) invisible to correlation-based selection, supporting interaction-rich nonlinear patterns.

2.2. Model Training and Hyperparameter Tuning:

To identify the best nonlinear regressor, we implemented a systematic model: selection pipeline based on a 5-fold `GridSearchCV` applied to the inner training split ($X_{\text{train,inner}}$). For each candidate model—Random Forest, Gradient Boosting, KNN, and MLP. We evaluated all relevant feature subsets derived from our three feature–selection strategies: Mutual Information, Sequential Forward Selection, and Random Forest feature importance.

Within each evaluation loop, `GridSearchCV` performed an exhaustive search over the predefined hyperparameter grid using only the inner cross-validation folds. The best hyperparameter configuration obtained from this search was then assessed on the held-out inner validation set ($X_{\text{val,inner}}$), which was never used during fitting. This two-stage procedure (cross-validation tuning followed by validation comparison) ensures that no information

leaks across splits, provides a fair comparison between all models, and allows us to reliably determine which model–feature-set combination exhibits the best generalisation performance.

Model	Feature Set	# Features	CV RMSE	Val RMSE	Best Params
Gradient Boosting	all_features	18	0.040309	0.043106	{'learning_rate': 0.05, 'max_
Gradient Boosting	rf_top10	10	0.039577	0.044680	{'learning_rate': 0.1, 'max_de
Random Forest	rf_top10	10	0.043778	0.046019	{'max_depth': None, 'max_featu
Random Forest	all_features	18	0.044162	0.046521	{'max_depth': None, 'max_featu
MLP	all_features	18	0.055029	0.054520	{'alpha': 0.01, 'hidden_layer
Gradient Boosting	mi_top10	10	0.044971	0.054625	{'learning_rate': 0.05, 'max_

Table 1: Model comparison on inner validation set (lower RMSE = better).

From the validation results, the best-performing configuration is the **Gradient Boosting** model trained with **all 18 features**This model consistently obtained the lowest validation error and was therefore selected as the final nonlinear model for Part 2.

The performance indicates that Gradient Boosting effectively handles redundant features and captures smooth nonlinear relationships.

Chosen Model: Gradient Boosting & all_features

Final Model Evaluation on the Test Set

After selecting this configuration, we retrained a fresh Gradient Boosting model with the same hyperparameters on the full training dataset and evaluated it on the untouched 500-row test set. The performance obtained was: Final Test RMSE (Gradient Boosting) = 0.04191.

Comparison with the Part 1 Linear Baseline

Model	Test RMSE	Improvement
Ridge (Part 1)	0.05596	–
Gradient Boosting (Part 2)	0.04191	24.9% lower error

Table 2: Comparison between the best linear and nonlinear models.

Interpretation. The nonlinear Gradient Boosting model demonstrates a marked improvement over the linear baseline, indicating that the target is influenced by complex, nonlinear interactions among medical biomarkers, nutritional indicators and occupational categories, which linear regression fails to capture.

Part 3: Integration of Image Data

The goal of Part 3 is to extend the predictive pipeline developed in Parts 1 and 2 by incorporating heart scan images into the model. This involves extracting visual features using a convolutional neural network (CNN), combining them with the tabular predictors, retraining the best nonlinear model from Part 2 (Gradient Boosting), and evaluating whether the addition of image information improves predictive performance.

3.1 Image Feature Extraction Using a Convolutional Neural Network

The first step is to transform each heart scan image into a numerical feature vector that can be merged with the tabular data. Using the CNN, each image is preprocessed (resized and normalized) and passed through the convolutional layers. Instead of using the final regression layer, we extract the activations from the last hidden layer, effectively using the CNN as a feature extractor.

This produces an 8-dimensional embedding that captures high-level visual information such as shapes, textures, and structural heart characteristics. Each patient therefore receives one image-derived feature vector, which is exported and treated as an additional predictor.

3.2. Constructing the Multimodal Dataset

After training the CNN, we obtained the following image embeddings:

- Training set: an array of shape 1000×8,
- Test set: an array of shape 500×8.

In these arrays, each row corresponds to a patient, while each column represents one of the 8 high-level visual features extracted by the CNN.

These image-derived features were then concatenated with the 18 tabular predictors from Part 2, resulting in a multimodal dataset with a total of 26 features per patient, comprising 18 tabular and 8 image-based features. This combined representation enables the model to leverage both clinical and visual information for improved predictive performance.

3.3. Feature Selection on the Multimodal Dataset

To stay consistent with Part 2, we applied Mutual Information (MI) feature selection on the combined dataset, using only the inner training split (80% of the training set) to avoid leakage.

The MI ranking identified the 20 most informative features, consisting of a mix of tabular variables and image embeddings. Reducing the feature space to the top 20 improved model stability and reduced overfitting during Gradient Boosting training.

3.4. Hyperparameter Tuning on the Multimodal Dataset

Since Gradient Boosting was the best nonlinear model in Part 2, we retrained and tuned this model on the multimodal dataset. Two feature sets were evaluated: **all 26 multimodal features**, **MI-Top-20 multimodal features.**

Using `GridSearchCV` on the inner training split, we selected the best hyperparameters and evaluated performance on the 200-sample validation set, and we obtain:

Feature Set	n_features	Validation RMSE	Test RMSE	Best Hyperp
MI-Top-20	20	0.02797	0.03041	{learning_ =0.08, max_depth=2, min_samples
All 26 Features	26	0.02858	0.03043	{learning_rate=0.08, max_depth=2, min_sample

Table 3: Summary of multimodal Gradient Boosting results sorted by validation RMSE.

The MI-Top-20 feature set performed best and was chosen as the final multimodal configuration.

Chosen Model: MI-Top-20 feature set.

3.6 Final Multimodal Model and Evaluation

The final multimodal model was trained on the MI-Top-20 features using an 80%/20% train/validation split, with a Gradient Boosting Regressor configured with the best hyperparameters. Its predictive performance was then evaluated on both the validation and unseen test sets:

Dataset	RMSE
Validation	0.02797
Test	0.03041

Table 4: Final RMSE of the Gradient Boosting model trained on the MI-Top-20 multimodal features.

This represents a clear improvement over the tabular-only version.

3.7 Interpretation and Discussion

In Part 2, the Gradient Boosting model trained solely on the tabular variables achieved a test RMSE of 0.04191, indicating that the method was able to capture the nonlinear relationships present in the clinical data. After integrating the heart scan images in Part 3 and constructing a multimodal dataset that combined the CNN-derived embeddings with the tabular predictors, the model was retrained and achieved a substantially lower test RMSE of 0.03041.

This reduction of approximately 27.4% in prediction error shows that the image-based features provide valuable complementary information that is not contained in the tabular data alone. The CNN embeddings capture structural and textural patterns of the heart, enriching the learning process and improving generalization. In addition, applying Mutual Information feature selection to the multimodal dataset further improved performance by filtering out weak or noisy predictors, allowing the Gradient Boosting model to focus on the most informative combined features.

Overall, the results from Part 3 clearly demonstrate that incorporating image data leads to a significant improvement in predictive accuracy compared to using only tabular features, fully addressing the requirement to compare performance with and without image-derived information.

Part 4: Understanding Heart Failure in Smurf Society

Hypotheses

1. **H1:** Age is a major determinant of heart failure risk.
2. **H2:** High blood pressure and high cholesterol increase risk.
3. **H3:** Unhealthy lifestyle habits (smurfberry liquor, smurfin donuts) increase risk.
4. **H4:** Certain professions are more at risk, independent of age.
5. **H5:** A combined profile of factors defines a high-risk group.

Visualizations and Results

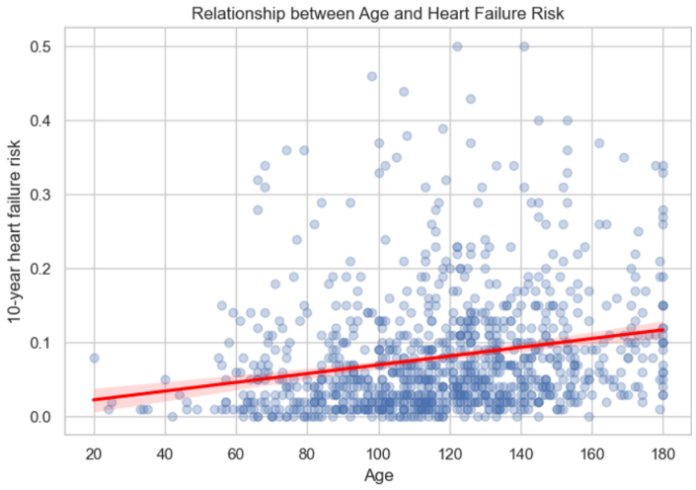

Figure 1: Risk vs Age. Older Smurfs show higher heart failure risk, supporting H1.

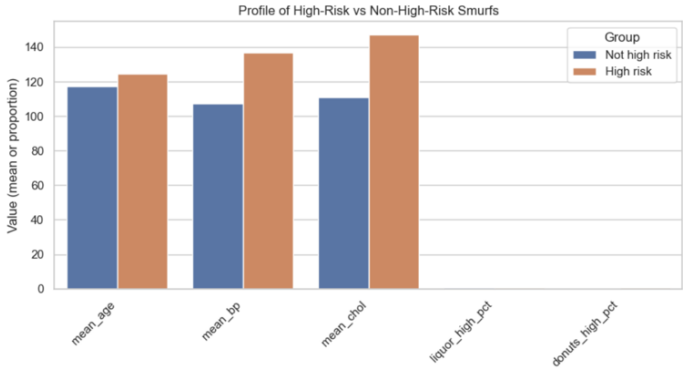

Figure 2: Mean risk by BP and cholesterol quartiles. High BP + high cholesterol defines the most vulnerable subgroup (H2).

Interpretation of Visualizations

Our visual analysis provides clear evidence in support of the five hypotheses formulated in this section. Figure 1 confirms H1 by showing that heart failure risk rises steadily with age. Although individual variation exists, the upward trend indicates that older Smurfs consistently face higher predicted risk. Figure 2 directly validates H2, demonstrating that both blood pressure and cholesterol contribute to worsening outcomes. Moreover, their combination produces the highest-risk subgroup, particularly when both markers fall in the upper quartiles. Figure 3 supports H3 by revealing that unhealthy lifestyle behaviors, especially high smurfberry liquor and smurfin donut consumption, are associated with elevated median risk and a greater concentration of extreme-risk individuals. Figure 4 further confirms H4, as Smurfs working in services and especially in administration and governance show notably higher risk distributions than those in manual professions, suggesting that occupational context contributes to cardiovascular vulnerability even beyond age. Finally, Figure 5 provides strong evidence for H5 by

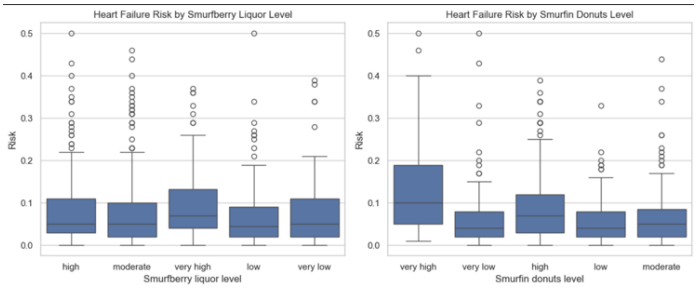

Figure 3: Lifestyle effects: Smurfs consuming more liquor and donuts have elevated risk, confirming H3.

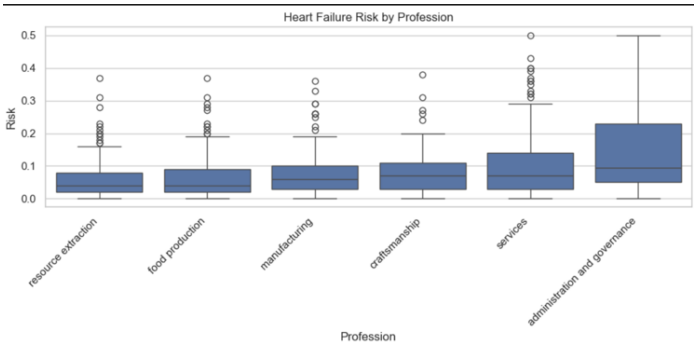

Figure 4: Risk by profession. Administration/governance and service roles are higher risk, supporting H4.

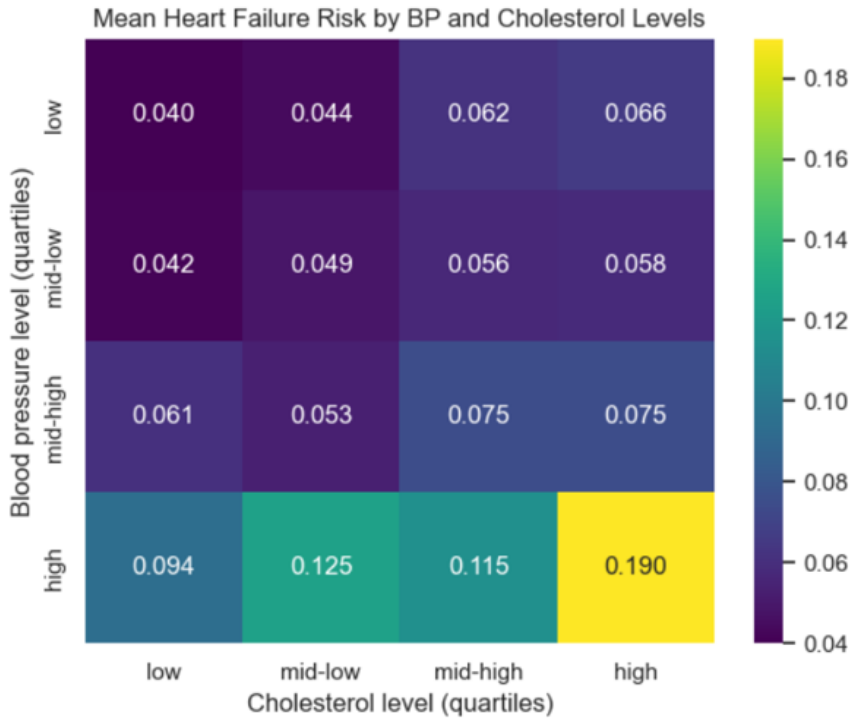

Figure 5: High-risk profile combining age, BP, cholesterol, and lifestyle habits. The group is older, hypertensive, hypercholesterolemic, and more likely to consume liquor and donuts (H5).

illustrating a clear high-risk profile: Smurfs with the highest predicted risk are simultaneously older, hypertensive, hypercholesterolemic, and more likely to engage in heavy liquor and donut consumption. Taken together, these visualizations consistently demonstrate that heart failure risk in Smurf society is shaped by a combined influence of demographic factors, clinical biomarkers, lifestyle habits, and occupational context.

Conclusion

Across the four parts of this project, we developed a full machine-learning pipeline to predict 10 year heart failure risk in Smurf society. Starting from careful preprocessing and a linear Ridge baseline, we showed in Part 2 that nonlinear models, particularly Gradient Boosting, substantially improved performance, reducing prediction error by nearly 25%. In Part 3, the integration of CNN derived image features led to our best model overall, achieving a test RMSE of 0.03041 and demonstrating the clear benefit of combining tabular and visual information. Part 4 complemented the modelling work by examining the factors most associated with elevated risk. The visual analyses revealed consistent patterns: older Smurfs, those with higher blood pressure or cholesterol, individuals with unhealthy lifestyle habits, and certain professions all exhibited higher predicted risk.

Taken together, the project provides both strong predictive performance and a clear understanding of the main drivers shaping heart failure vulnerability in Smurf society.