Proposal-Predicting Baldness Risk and Hair Loss Severity Based on Personal Characteristics Using Machine Learning

Team Members

Full name	Proposal-role	Project-role	Major	email
Linshan Yu	Baseline models	Contributors to Idea Development, Designs and trains machine learning models	Computer science	lyu4@wpi.edu
Junjie Chu	Proposed adjustments	Contributors to Idea Development, Data cleaning,Preprocessing, and feature engineering	Computer science	jchu1@wpi.edu
Ping Li	Project objective	Idea generator,Designs and trains models,Evaluation	Computer science	pli2@wpi.edu
Yongli Bao	Dataset In formation	Contributors to Idea Development, Data collection, Evaluation & visualization	Computer science	ybao4@wpi.edu
Zongyu Mu	Project background	Contributors to Idea Development, Data collection, Coding - review,Designs and trains machine learning models	Computer science	zmu@wpi.edu

Background

Hair loss is a global health and aesthetic concern that significantly affects individuals’ physical and psychological well-being. Around 42% of men worldwide experience some form of noticeable hair loss or baldness during their lifetime.

This project aims to apply machine learning techniques to predict the risk and extent of hair loss based on multi-dimensional personal health data. These data include demographic attributes, lifestyle factors, biological indicators, and psychological traits. The ultimate goal is to develop predictive models that assist medical professionals in early detection, risk assessment, and treatment planning for individuals susceptible to hair loss.

By analyzing a dataset containing almost 8000 real-world samples, we aim to develop both classification and regression models that can predict whether a person is at risk of developing baldness (binary classification and estimate the expected severity of future hair loss (regression on a scale from 0 to 100).

Accurate and interpretable predictions from these models could have meaningful applications in personalized healthcare, helping dermatologists identify underlying causes and optimize treatment strategies for patients vulnerable to hair loss. It is beneficial to both individuals and the public's awareness of anti-baldness.

Dataset Information

1. Dataset Link

- The dataset adopts the baldness probability dataset from kaggle. This dataset is anonymized. The link to the dataset is below
- https://www.kaggle.com/datasets/itsnahm/baldness-probability

2. Dataset Description

Dimensions	Details
Data type	Tabular data, with each row representing one individual sample and each column corresponding to one feature, totaling 12 columns
Sample size	A total of 7,917 rows of samples; there are a total of 9 numerical columns and 5 category columns. The target column bald_prob has 7,838 rows of non-missing samples.
Features	• Basic features: age = Age of people in data gender = Male or female weight = weight of people body in data height = height of people body in data education = Education level of people job_role = Job roles of people who in data salary = salary each month is_married = Married status (1 = Yes and 0 = No) • Gene-related indicators：

	is_hereditary = Is the bald based from hereditary? (1 = Yes and 0 = No) • Pressure level： stress = Stress level of people in range 1 (lower) to 10 (highest) • Living habits： is_smoker = Is the people a smoker? (1 = Yes and 0 = No)
Target variable	• bald_prob = Probability score of bald which occur to people in range 0 to 1. The higher probability score means the higher probability to baldness occurred and vice versa.
Distribution	• The male sample in the dataset was slightly more than the female sample (approximately 53.4% : 46.6%), and the overall age distribution was concentrated in the 25-55 age range. • The average value of the target variable, hair loss probability, is 0.574, the median is 0.568, and the standard deviation is approximately 0.21. • Group mean values: 0.605 for males and 0.507 for females; Smokers: 0.619, non-smokers: 0.528; 0.724 for those with a genetic history and 0.535 for those without. Overall, it shows an upward trend with the increase of age, stress, smoking and genetic history.
Missing values/outliers handling	• Missing overview (proportion) : job_role 16.64%; The remaining columns range from 0.71% to 1.12% (bald_prob missing 1.00%). Suggestion: ① Fill in the numerical column with the median /KNN; ② Use the mode for category columns or combine them into "Unknown"; ③ Check for unreasonable values (such as age <5, etc.) and perform truncation or robust scaling. ④ For high-base categories (such as province), frequency merging or target encoding can be performed; ⑤ Before training, confirm that the bald_prob constraint is [0,1].

Project Objective

1. Project Objective

This project aims to apply supervised machine learning techniques to predict two related outcomes concerning hair health:

- Classification goal: Predict if an individual is at high risk or low risk of baldness.
  Target performance: accuracy $\geqslant$ 80%, recall $\geqslant$ 75%, precision $\geqslant$ 75%.
- Regression goal: Estimate the severity of hair loss on a continuous scale from 0 to 100.
  Target performance: MAE $\leqslant 10$, $R^2 \geqslant 0.6$.

The ultimate objective is to build interpretable and reproducible models that can identify key contributing factors to baldness, offering insights for personalized prevention and wellness recommendations.

2. Proposed Work and Plan

2.1 Validation: Scenarios, Data, and Preprocessing

Scenario and Relevance:
The project targets real-world wellness applications such as early detection of baldness risk and individualized lifestyle advice. Predictions are based on personal characteristics (age, gender, job, stress, hereditary factors, smoking, etc.) drawn from population-level data.

Data Source:

- Primary dataset: Kaggle Baldness Probability Dataset (~7,900 samples, 14 features).
- Supplementary synthetic data may be generated to balance distribution and enhance generalization.

Preprocessing:

- Handle missing values via median (numeric) or mode (categorical) imputation.
- Standardize numerical features; encode categorical ones using one-hot or target encoding.
- Detect and cap outliers (e.g., unrealistic age or weight).
- Split dataset into 70% training, 15% validation, and 15% testing, using stratified sampling to maintain class proportions.

2.2 Standardized Implementation (Replicability)

Framework and Environment:
Python 3 with pandas, numpy, scikit-learn, and matplotlib in VS code or Google Colab.

Model Configuration:

- For baldness risk classification: Logistic Regression, Random Forest, and XGBoost Classifier.
- For hair loss severity prediction: Linear Regression, Ridge/Lasso Regression, Random Forest Regressor, and XGBoost Regressor.
- Apply GridSearchCV for hyperparameter tuning and 5-fold cross-validation for performance estimation.
- Fix random seeds for full reproducibility.
- Analyze feature importance and visualize relationships for model interpretability.

3. Evaluation: Metrics and Experiments

Prediction Target	Primary Metric	Secondary Metrics	Rationale
Baldness Risk (Yes/No)	Recall	Precision, F1-score, ROC-AUC, Accuracy	Recall is prioritized to minimize false negatives—ensuring high-risk individuals are not missed. Other metrics ensure overall balance and robustness.
Hair Loss Severity (0–100)	MAE (Mean Absolute Error)	RMSE, R^2 Score	MAE provides interpretable deviation, while RMSE and R^2 capture variance explanation and residual quality.

Experimental Procedures:

- Use 5-fold cross-validation on training data.
- Evaluate on the independent test set.
- Visualize confusion matrices and ROC curves for classification; residual plots for regression.
- Rank and compare models according to primary and secondary metrics.
- Perform feature importance and sensitivity analysis (e.g., effect of stress, heredity, and age).

4. Model Selection and Comparison

For baldness risk prediction, models will be ranked by Recall (primary) and F1-score, with Logistic Regression as a baseline for interpretability.
For hair loss severity estimation, models will be ranked by MAE and R^2, starting with Linear Regression as baseline and comparing improvements from regularized and ensemble models.
The final models will balance accuracy, generalization, and explainability.

5. Expected Outcomes

- Two complete, reproducible ML pipelines (classification + regression).
- Target performance:
  - Baldness risk: Accuracy $\geqslant 80\%$, Recall $\geqslant 75\%$.
  - Severity score: MAE $\leqslant 10$, $R^2 \geqslant 0.6$.
- Ranked model comparison table and interpretability visualization (e.g., SHAP or feature importance).

- Analytical insight into the strongest predictors of hair loss, such as heredity, stress, or lifestyle.
- Practical demonstration of applying ML to personalized health prediction.

Baseline Models

Task Type	Baseline Models	Rationale (Adaptability & Justification)
Classification (Predict Baldness: Yes/No)	1. Logistic Regression 2. Decision Tree Classifier	1. Logistic Regression is chosen as a simple and interpretable linear baseline. It quantifies how each individual feature (e.g., age, stress level, heredity) affects the probability of baldness. The model' s coefficients directly reveal feature influence, making it suitable for medical interpretability. 2. Decision Tree Classifier serves as a non-linear counterpart that can capture hierarchical feature interactions (e.g., how stress and heredity jointly increase risk). It provides transparent decision rules, complementing the linear model by modeling non-linear patterns.
Regression (Predict Hair Loss Severity: 0–100)	1. Linear Regression 2. Ridge / Lasso Regression	1. Linear Regression acts as a fundamental baseline to quantify linear relationships between features and hair loss severity. It is simple, explainable, and forms a benchmark for more complex models. 2. Ridge and Lasso Regression extend the linear model with regularization to control overfitting and enhance stability. Ridge addresses multicollinearity among correlated health factors (e.g., age and stress), while Lasso performs feature selection, highlighting the most influential predictors for interpretability.

Proposed Adjustment

Expected Risks, Assessment, and Mitigation

1. **Data-Related Risks**

- Risk: Insufficient or imbalanced samples (e.g., bald vs. non-bald groups).
- Assessment: May reduce model generalization, prolong data collection, and require more preprocessing.

- Mitigation: Apply data-augmentation or resampling (SMOTE), collect additional balanced samples, and ensure standardized feature encoding.

2. **Model Development Risks**

- Risk: Overfitting in small-sample scenarios or unstable training due to noisy features.
- Assessment: Could extend model-tuning time and increase computational cost.
- Mitigation: Introduce regularization (Ridge/Lasso), apply cross-validation, and remove redundant variables.

3. **Testing & Evaluation Risks**

- Risk: Performance metrics fluctuate across different folds or unseen test sets.
- Assessment: Inconsistent results may delay validation and comparison across models.
- Mitigation: Use k-fold cross-validation and maintain a fixed random seed for reproducibility.

4. **Deployment & Adaptability Risks**

- Risk: Model trained on lab data may not generalize to new populations.
- Assessment: Impacts real-world applicability and requires additional domain adaptation.
- Mitigation: Continuously collect feedback data, retrain periodically, and test robustness under new demographic subsets.