Predicting Hair Loss Risk Using Machine Learning

1 Introduction and Motivation

Hair loss affects millions of individuals worldwide and can have significant psychological and medical impact. Early detection enables preventive interventions, such as lifestyle modification, nutritional supplementation, and medical consultation. This project develops a machine learning model that predicts whether an individual is at increased risk of hair loss using self-reported survey features. The system aims to support early awareness rather than medical diagnosis.

Motivation. Many risk factors—such as hormonal changes, stress, and nutritional deficiency—are modifiable. Hence, predictive tools can provide value for public health and personalized wellness.

Target audience. Healthcare researchers, machine learning practitioners, and individuals seeking data-driven hair health assessment.

Goals.

- Build an end-to-end reproducible ML pipeline.
- Evaluate classical (Decision Tree, Random Forest) and modern methods (XGBoost).
- Prioritize recall to reduce false negatives.
- Identify top contributing factors influencing predictions.

2 Data

2.1 Dataset Source

The dataset is publicly available on Kaggle:

`https://www.kaggle.com/code/ayaabdalsalam/hair-loss-analysis/input?scriptVersionId=217216355`

2.2 Dataset Description

The dataset contains **999 samples** collected from a survey instrument. Each record includes:

- Demographic: Age
- Lifestyle: Smoking, Weight_Loss, Poor_Hair_Care_Habits
- Environmental: Environmental_Factors
- Biological: Hormonal_Changes, Genetics
- Medical: Medications_and_Treatments, Medical_Conditions
- Stress level
- Binary label: `Hair_Loss` (0 = No, 1 = Yes)

2.3 Preprocessing

- Cleaned column names (removed spaces, "&", and dashes).
- Converted Yes/No to 1/0.
- Filled missing values with 0.
- Engineered:
  - **AgeGroup**: 4-bin age mapping (0–25, 26–35, 36–45, 45+).
  - **Stress3**: maps textual/numerical stress levels to $\{0,1,2\}$.

2.4 Train/Test Split

A stratified 70/30 split produces:

- Train set: 699 samples
- Test set: 300 samples

3 Problem Formulation

We treat the task as a binary classification problem:

$$f_\theta(X) \to \hat{y} \in \{0, 1\}$$

Evaluation metrics:

- **Recall**: prioritized due to cost of false negatives.
- Precision
- F1-score

4 Methods

4.1 Models

Three supervised learning models were evaluated:

- Decision Tree (max_depth=5)
- Random Forest (300 estimators, class_weight=balanced)
- XGBoost (400 estimators, learning_rate=0.05, max_depth=4)

Random seed = 42 was fixed across all experiments.

5 Experimental Setup

Environment:

- Python 3.10
- scikit-learn 1.5
- XGBoost 2.0
- pandas, numpy

6 Results

6.1 Performance Comparison

Results below are computed using the **held-out 300-sample test set**.

Model	**Recall**	**Precision**	**F1**
Decision Tree	0.5436	0.4969	0.5192
Random Forest	0.5570	0.5253	0.5407
XGBoost	**0.6376**	**0.5556**	**0.5938**

Table 1: Predictive performance on test data.

6.2 Feature Importance

Feature	**Importance**
Hormonal_Changes	0.1736
Stress3	0.1468
Environmental_Factors	0.1396
AgeGroup	0.1353
Weight_Loss	0.1173
Smoking	0.1022
Poor_Hair_Care_Habits	0.0976
Genetics	0.0876

Table 2: XGBoost feature importances.

7 Reproducibility Checklist

Requirement	**Details Provided**
Data Source	Kaggle link included; dataset filename: `Predict Hair Fall.csv`
Feature Engineering	AgeGroup bins, Stress3 mapping, yes/no normalization
Hyperparameters Specified	All model hyperparameters listed for DT, RF, XGBoost
Random Seeds	Fixed seed = 42 for reproducibility
Train/Test Procedure	70/30 stratified split fully described
Saved Artifacts	`models/best_model.joblib`, `model_meta.joblib`, `feature_schema.json`
Instructions to Run	Single command: `python train_classification.py`

Table 3: Reproducibility compliance checklist.

8 Conclusion and Future Work

This project demonstrates that machine learning, particularly gradient boosting methods, can effectively predict hair loss risk based on lifestyle, environmental, and biological features. XGBoost achieved the highest recall and F1, validating the hypothesis that boosting methods outperform classical tree-based models on this dataset.

Future improvements include:

- Adding cross-validation for more robust estimates.
- Collecting richer biological and medical features.
- Exploring calibrated probability outputs.