Hair Loss Data Analysis and Factor Prediction

Tanner Wheeler
Computer Science Graduate Student
Utah State University
CS 6850
A01770306

***Abstract*—This document discusses the factors contributing to hair loss. The importance of determining the riskiest factors is discussed with personal histories. Analysis is completed on a Kaggle dataset to determine the biggest factors of hair loss. Four prediction models are created to determine if a person with multiple factors will have hair loss. It is determined this dataset does not match current medical statistics about hair loss. This dataset has been created and other entries should be gathered from real patients in order to have a more complete analysis the proposed methods.**

***Index Terms*—Hair Loss, Analysis, Prediction, Kaggle**

I. Introduction

A. Background

As a person ages hair loss is a common indicator a person's body is getting older. Compared to when in their youth, many people anticipate having less hair. This change in a person's body can cause anxiety, depression, low-self esteem, and fear of old age. I personally have experienced some of these emotions due to complete hair loss when needing chemotherapy treatments. During a nine week treatment cycle, I began to lose my hair at the end of week two. Months after my chemotherapy treatments were completed my hair returned. My wife has also felt the same emotions of anxiety and depression from hair loss. During and after high school my wife had an eating disorder. She would exercise more than two hours everyday and ate small amounts of food in order to control her weight. This caused her hair to become thin and fall out more frequently. After seeking professional and medical help for her eating disorder her hair began returning to normal.

Due to our experiences, my wife and I want to take care of our newly returned hair. My wife and I notice small changes to our hair because of past experiences. Completing a quick internet search you can find 12 different types of shampoo, e.g. natural, strengthening, shampoo for babies and children, etc., [1] with 35 different brands within the United States alone [2]. This is only one of 10 hair care products [3]. We can infer from these facts that people care about their hair.

Though people want to take care of their hair, there are factors causing hair loss even with good hair care being a daily routine. Some factors such as age and genetics cannot be avoided [4]. For age, every person will get older and their bodies will begin focusing on functions other than producing hair. With genetics, our DNA is composed of genes from both our mother and our father. If there is a history of baldness from both our mother and our father we are likely to experience baldness too. For other factors, we choose to endure, in order to protect our bodies, medications and medical treatments [4]. ”Hair loss can be a side effect of certain drugs, such as those used for cancer, arthritis, depression, heart problems, gout and high blood pressure. [4]” From personal experience, a person will begin to lose their hair during certain types of chemotherapy due to the treatment attacking fast growing cells in the body which include cancer cells, hair, a person's gums, and fingernails. Finally, some factors of hair loss are bi-products of their situations whether they are self inflicted or due to circumstances outside of their control. These factors include poor diets, stress, and hormonal changes. When a person experiences nutrient deficiencies, the body will begin to protect itself by shutting down less vital functions such as hair growth [6].

The purpose of this project is to complete an analysis of hair loss data and predict if a person will experience hair loss given different factors in their life. The aim of the analysis is to find which factors of hair loss are leading causes.

II. Method

A. The Data

The dataset for this project was downloaded from Kaggle [7]. This dataset is the sole source of data used in the project. The dataset includes 999 different entries with 13 different attributes:

- Medical **ID** of the Individual
- If the Individual has Experienced **Hormonal Changes**
- Experienced **Medical Conditions**
- Taken **Medications & Treatments**
- The Individual's **Nutritional Deficiencies**
- **Stress** Level of the Individual
- The Individual's **Age**
- If the Individual Practices **Poor Hair Care Habits**
- Exposure to **Environmental Factors**
- Does the Individual **Smoke**
- Significant **Weight Loss**
- Experiencing **Hair Loss**

Of the 13 attributes there are 11 Nominal attributes with 7 of these attributes being Yes or No binary entries. One attribute, Stress, is ordinal with values Low, Moderate, and High. Finally there is one ratio attribute, Age. In the dataset, the attribute Hair Loss is our binary classification. There are zero Null or empty entries.

B. Pre-Processing

In this project we create four different prediction models: Decision Trees, Random Forests, Naive Bayes, and Logistical Regression. In order to create and use these methods provided by Python for predictive modeling along with other Analysis, the data was modified into four different variations. First, we did an analysis on the attribute 'ID'. Four duplicate 'ID's were found in the data. It was determined not to remove the duplicate entries from the data as the attributes appeared to differ significantly. Rather, the duplicate 'ID's were determined to be a recording error. Finally, we removed the attribute 'ID' from the data.

Of the four different variations, one variation, which will be referred to as the Normal Data, was not modified from the base data except for the removal of the attribute 'ID'. The second and third variations of the data, which will be referred to as Binned Age Data and Variation Data, we created as the predictive models were not classifying the data significantly better than a coin flip. These two variations were created to improve this result. The Binned Age Data has the attribute 'Age' converted into nominal attributes for ages 18-20, 20-25, 25-30, 30-35, 35-40, 40-45, and 45-50 as the ages ranged from 18-50. The Variation Data is a dataset comprised of only the attributes with the highest variation: Medical Conditions, Medications & Treatments, Nutritional Deficiencies, Stress, and Age. These five attributes were not binary classifications. The final variation of the data, which will be referred to as One Hot Data, is the data with one hot encoding performed on all attributes except 'ID'. When working with the Normal Data, the Binned Age Data, and the Variation Data the values were converted to a numerical value using `pandas.get_dummies(...)` to build the Random Forest Classifier model.

C. Factor Analysis

Visualizations, Mutual Information, and Rules Analysis were used to determine which factors contribute the most to Hair Loss in the data. Factor Analysis, besides Rules Analysis, were completed before predictive modeling to help understand which variables would be the most helpful in determining Hair Loss.

For Visualizations, pandas was used on the Normal Data to separate the data into each category's possible values. The number of positive Hair Loss and Negative Hair Loss were then counted and used to create bar graphs representing which value has the highest number of positive Hair Loss and Negative Hair Loss for the attribute. From Figure 1 we can see variation in the 'Medical Conditions' and 'Age' for determining hair loss, but with 'Stress' and 'Genetics' we do not see a wide variety between the attributes. For the rest of the attributes, their graphs appeared similar to the graphs of 'Stress' and 'Genetics' with the exception of 'Medications & Treatments' and 'Nutritional Data' which are more similar to 'Medical Conditions' and 'Age'. This shows the highest variations of data for our classifier comes from 'Medical Conditions', 'Medication & Treatments', 'Age', and 'Stress'. Any other attribute will need to be combined with another attribute as to not give a 50% accuracy in classification.

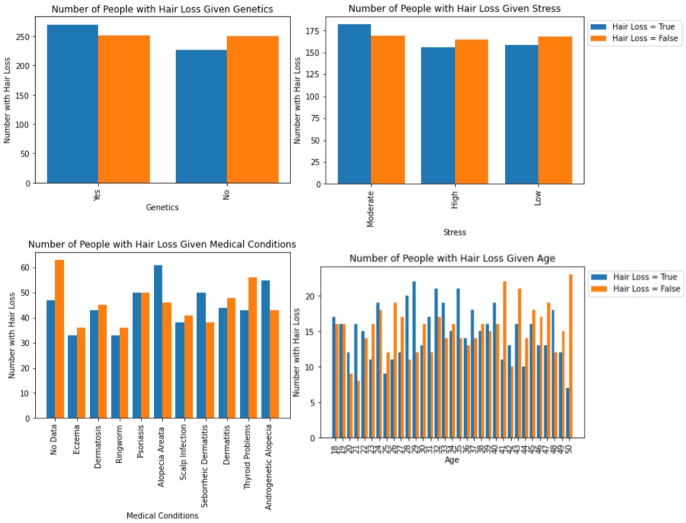

Fig. 1. Attribute Values and Hair Loss Counts

Calculating the Mutual Information between each category in the Normal Data we should be able to see which attributes have the highest correlation for these nominal values. The attribute with the highest correlation to our Hair Loss classification should also been seen. Looking at Table I the calculations mostly appear to be 0.0 for all mutual information and the anticipated information is not given. This will be discussed in the Experimental Results section of this paper.

Market Basket Analysis or Association Rule Mining was used for our Rule Analysis. This method used the dataset One Hot Data. Using the `apriori` and `association_rules` methods from MLXtend.Frequent_Patterns in Python we were able to generate rules with 0.2 Support and 0.5 Confidence. With the resulting rules, we extracted the rules with the consequent of 'Hair Loss' as True. We generated these rules with the following Support (S) and Confidence (C):

- Smoking = No → Hair Loss = Yes (S=0.25, C=0.53)
- Weight Loss = Yes → Hair Loss = Yes (S=0.25, C=0.52)
- Genetics = Yes → Hair Loss = Yes (S=0.27, C=0.52)
- Poor Hair Care Habits = No → Hair Loss = Yes (S=0.26, C=0.52)
- Environmental Factors = No → Hair Loss = Yes (S=0.25, C=0.51)

TABLE I
MUTUAL INFORMATION VALUES TABLE

	Genetics	Hormonal Changes	Medical Conditions	Medications & Treatments	Nutritional Deficiencies	Stress	Age	Poor Hair Care Habits	Environmental Factors	Smoking	Weight Loss	Hair Loss
Genetics		-0.0002	0.0	0.0	0.0	0.0	0.0	0.0016	-0.0002	0.0013	0.0225	0.0
Hormonal Changes			0.0	0.0	0.0	0.0	0.0	0.002	0.0012	0.0014	0.0211	0.0
Medical Conditions				0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Medications & Treatments					0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nutritional Deficiencies						0.0	0.0	0.0	0.0	0.0	0.0	0.0
Stress							0.0	0.0	0.0	0.0	0.0	0.0
Age								0.0	0.0	0.0	0.0	0.0
Poor Hair Care Habits									0.0186	0.0169	0.0032	0.0
Environmental Factors										0.0014	0.0639	0.0
Smoking											0.0065	0.0
Weight Loss												0.0

The highest mutual information calculated is Environmental Factors and Weight Loss

- Hormonal Changes = Yes → Hair Loss = Yes (S=0.26, C=0.50)

D. Prediction Models

With the Normal Data, the Binned Age Data, and the Variation Data we created models using Decision Trees, Random Forests, Naive Bayes Classification, and Logistical Regression. For testing our classification models 90% of the data was split into the training set and 10% into the testing set. For Decision Trees, we fed multiple tree depths to the model to determine the best depth for the model. From Figure 2, we see the best depth is 5 resulting in a 53% testing accuracy and 60% training accuracy from the Normal Data. This value improved with the Binned Age Data and the Variation; however, the max depth is better at 6 for these datasets. These datasets both achieved 56% accuracy with the training accuracy at 65% and 62% respectively.

For the Random Forest classification the max depth of the trees was set at 1. Multiple values for the number of estimators were tried with the classification. In Figure 3, 20 estimators resulted in the highest testing accuracy of 52.5% and a training accuracy of 56.5%. Again, the testing accuracies were improved with the Binned Age Data and the Variation Data datasets; however, the training accuracies were farther away from the testing accuracies.

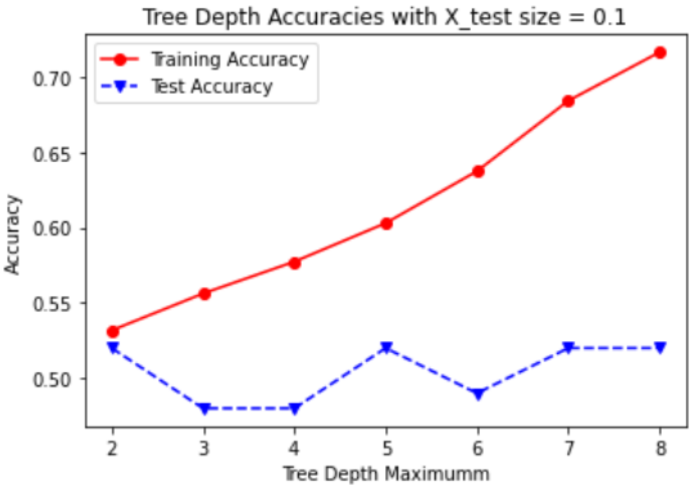

Fig. 2. Decision Tree given different max-depths

From our attribute analysis we know the attributes in the data are not distributed over a Gaussian distribution. To accommodate this knowledge, we used the `CategoricalNB` Naive Bayes method from SKLearn which works with categorical attributes. From this model we achieved a high testing accuracy of 47% and 59% training accuracy with the Binned Age Data.

SKLearn's `LogisticRegression` method was used for building the Logistic Regression classification model. Different combinations of variables for the model were used: LBFGS sovler with L2 Penalty, Newton-Cholesky Solver with L2

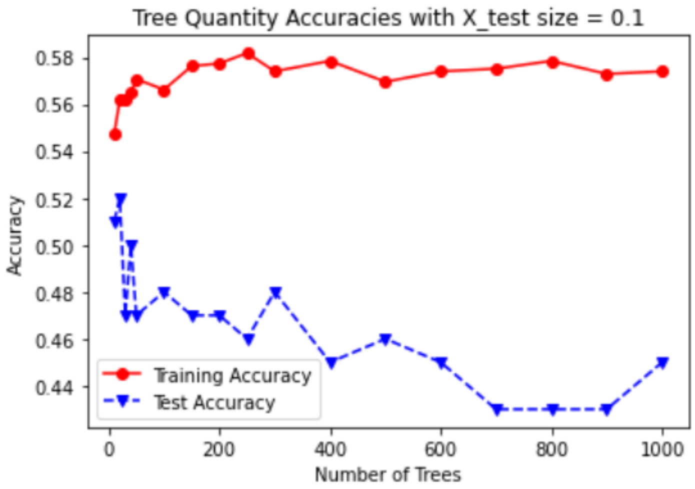

Fig. 3. Random Forests given different estimator quantities

Penalty, Liblinear Solver with L2 Penalty, and Liblinear Solver with L1 Penalty. For this classification model, the Variation Data performed the best with the LBFGS Solver and the L2 Penalty. This combination had a 53% testing accuracy with a 54% training accuracy.

To improve the results of the Decision Tree and Random Forest models, we hyper-tuned the parameters using SKLearn's `GridSearchCV` algorithm. The best parameter settings for Decision Tree classification did not improve the results for training of testing accuracies. The best test accuracy results for Random Forest classification were improved to 57% for the Normal Data, but the testing and training accuracy difference increased with a training accuracy of 69%.

III. Experimental Results

A. Factor Analysis Results

Looking at the results for the Factor Analysis completed, there appears to be something strange with the data used in this project. We know that medical conditions, medication & treatments, and genetics should be attributing to Hair Loss [4]. We do see from our rule analysis genetics will contribute to hair loss, but only 50 percent of the time can we be confident in this rule. This is the same for all of the computed rules. If we see one of the rules we would have the same amount of confidence in the result as a coin flip. With chemotherapy it is reported patients have a 65% chance of losing their hair [7]. We should be seeing this in our rules, but 'Medication & Treatments' do not appear in our rules.

For our Mutual Information table there should be mutual information between some attribute and hair loss; however, from Table I no attribute shares any mutual information with 'Hair Loss'. Few of the attributes even share mutual information with one another. This could be a coding error due to a lack of knowledge, but the code was compared with previously used code to verify this error did not occur. We should be seeing mutual information between 'Medications & Treatments' and 'Weight Loss' as side effects to certain medications can cause weight loss.

B. Prediction Model Results

The highest results for each prediction model have been compiled into Table II. If we needed the model with the smallest difference between testing and training accuracies we would choose the Logistic Regression model on the Variation Data. If we wanted the model with the highest test accuracy we would choose the Random Forest model on the Normal data. Even with an accuracy of 57%, we should not be satisfied with the results. With our classification being a binary classification, a 50% accuracy is achieved by randomly assigning the test data to a class. Like discussed above in Factor Analysis Results, randomly assigning could be completed by a coin flip, and 57% is not satisfying enough to justify this model in the classification of the data.

C. Discussion

From our factor analysis and prediction models we can see something is different about this data. We should be seeing patterns in the data providing Mutual Information and rules matching medical statistics. These patterns, that do not exist, should help with classifying the data. Looking more in-depth into the dataset on Kaggle we can find other programmers that worked with this code did not have better results except for one programmer that used AdaBoost [6].

With the results of our analysis, we can suspect this data is a paired trial of patients, where specific patients are chosen, or data that was randomly generated. The latter is confirmed from further investigation on the Kaggle website. This dataset is synthetically created data rather than actual data gathered from patients [6]. This synthetic data would explain the replication of 'ID's in the data especially since the information of each replicated 'ID' is does not match previous histories based on their ages.

IV. Conclusion

In this project, I was unable to find which of the provided factors contributed more to hair loss. The Mutual Information calculated showed there are only small correlations between the given attributes. The rules generated do not have a strong enough confidence to be used in determining a factor of hair loss within the data. Of the four classification methods used, Random Forest produced the highest testing accuracy and Logistic Regression had the closest comparable testing and training accuracies, but these accuracies are not much better than a coin toss. To determine the biggest factors of hair loss and to create a better prediction model, more data should be collected from actual patients. The current dataset used does not contain real patterns of hair loss factors.

TABLE II
PREDICTION MODEL HIGHEST ACCURACIES

Data Set	Decision Tree		Random Forest		Naive Bayes		Logistic Regression	
	Test	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***
Normal	52%	60%	57%	69%	45%	61%	50%	56%
Binned Age	56%	65%	54%	71%	47%	59%	48%	57%
Variation	55%	61%	54%	71%	42%	60%	53%	54%

A. Future Work

Because the data is synthetically generated, I would be interested in determining if the model used to create the data can be found within the data. This could be done by creating models based on all different combinations of the attributes. There are 2046 different combinations. Using hyper-tuning we could find the best Decision Tree or Random Forest model to classify the data. The best results could potentially determine the model used to create the original dataset.

As stated above, another programmer was able to calculate a higher testing accuracy using AdaBoost. Further work can be done using AdaBoost to improve the predictive classification of the data.

B. Lessons Learned

When working with Decision Trees and Random Forests, I learned to monitor the testing and training accuracy results. From initial efforts the testing accuracy looked promising, but the training accuracy difference was too large compared to the testing accuracy. This showed the model was overfitting the data.

The biggest lesson learned is to verify the gathering of the dataset used. Most of the frustration about this project came from the lack of or little improvement made with multiple classification methods and hyper-tuning. Knowing the dataset was synthetic data beforehand, I would have changed the direction of the analysis.

V. ROLES

I completed the project for this class alone. I did receive dietetic information from my wife who is a Dietetics graduate program at Utah State University.

REFERENCES

[1] J. Morais, “Types of shampoo,” Joan Morais Cosmetics School, https://joanmorais.com/formulating-shampoo-part-1-types/ (accessed May 1, 2024).

[2] Published by Statista Research Department and F. 5, “U.S.: Brands of Shampoo used 2020,” Statista, https://www.statista.com/statistics/276927/us-households-brands-of-shampoo-used/ (accessed May 1, 2024).

[3] F. Du, “Hair products & hair care: Best shampoo, conditioner, hair mask and more (updated 2022),” Luxy® Hair, https://www.luxyhair.com/blogs/hair-blog/hair-products-101 (accessed May 1, 2024).

[4] “Hair loss,” Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/hair-loss/symptoms-causes/syc-20372926 (accessed May 1, 2024).

[5] M. Wheeler and T. Wheeler, “Personal Interview,” May. 1, 2024

[6] A. Kulkarni, “Hair health prediction,” Kaggle, https://www.kaggle.com/datasets/amitvkulkarni/hair-health/data (accessed May 1, 2024).

[7] P. Clarence D. Moore, “Hair loss due to cancer treatment,” Pharmacy Times, https://www.pharmacytimes.com/view/hair-loss-due-to-cancer-treatment (accessed May 1, 2024).