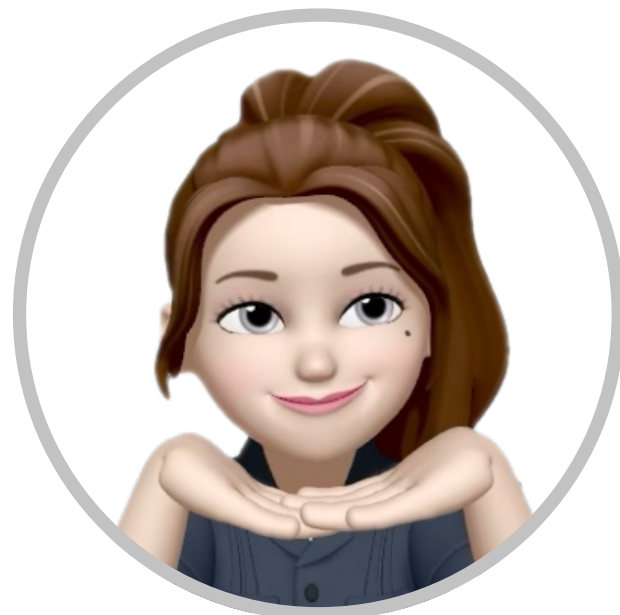


EWHA WOMAN'S UNIVERSITY

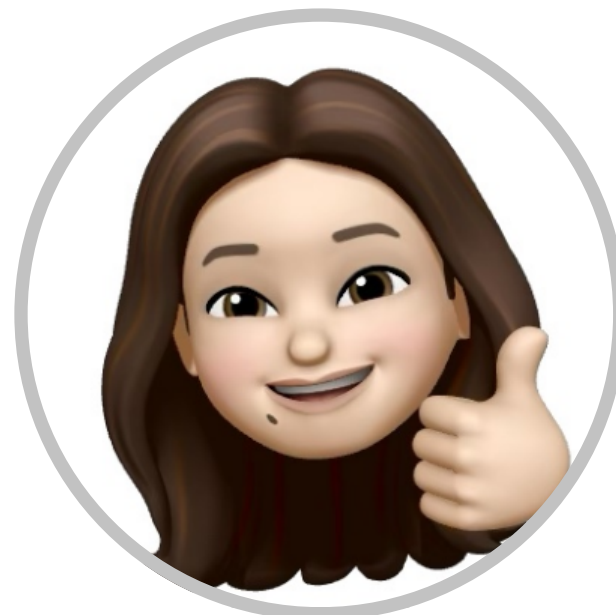
## Data Science Team Project

# Hair-Loss Prediction Model

Team 10



2391007 Minjeong Kim



2391016 Suahn Lee



2391017 Yoonji Lee

# Before we start...



**Chulsoo(31, male)**

Graduate Student

- hesitates to visit the hospital for treatment
- believes that it might not actually lead to baldness

## Input

---

- Genetics → NO
- Hormonal Changes → YES
- Medical Conditions → No data
- Medication → Accutane
- Nutritional Deficiencies → Magnesium
- Stress → High
- Poor Care Habits → YES
- Environmental Factors → YES
- Smoking → YES
- Weight Loss → YES



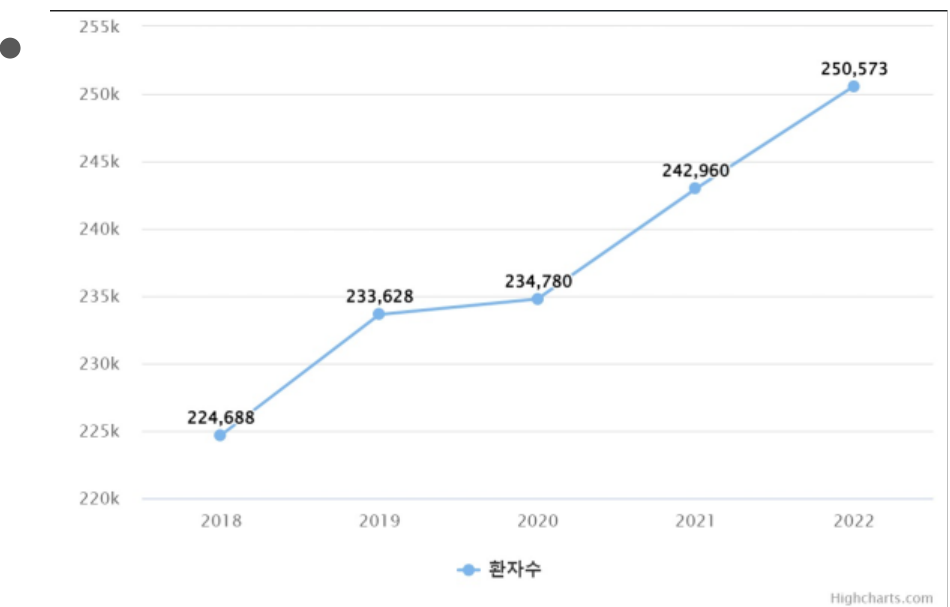
Probability of  
being bald

**??%**

# | Background

## Increasing patients

- nearly 250,000 people sought medical treatment for "pathological alopecia"

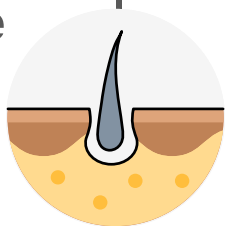


## Growing Hairloss Market

- KB Securities estimated the Korean domestic hair loss market to be worth 4 trillion won in 2023
- it is projected that the global market for hair loss-related products and treatments will reach approximately 400 trillion won by 2024.

## Importance of Early Treatment

- Medication treatment is only possible if the hair follicles are still alive
- The sooner the treatment is started, the more effective it will be
- When hair loss becomes severe, hair transplant surgery becomes the only viable option.  
+ also add to the emotional burden.



A service that aligns with these social currents is needed!

# | Background

## Increasing patients

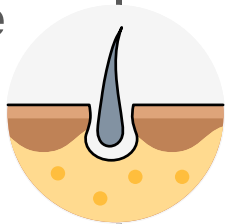
- nearly 250,000 people sought medical treatment for "pathological alopecia"
- A survey by Embrain (1000 Adults, 2023)
  - 17.2% in their 20s
  - 28.4% in their 30s
  - 29% in their 40s
  - 33.3% in their 50s

## Growing Hairloss Market

- KB Securities estimated the Korean domestic hair loss market to be worth 4 trillion won in 2023
- it is projected that the global market for hair loss-related products and treatments will reach approximately 400 trillion won by 2024.

## Importance of Early Treatment

- Medication treatment is only possible if the hair follicles are still alive
- The sooner the treatment is started, the more effective it will be
- When hair loss becomes severe, hair transplant surgery becomes the only viable option.  
+ also add to the emotional burden.



A service that aligns with these social currents is needed!

# | Problem Definition

predict the likelihood of becoming bald and induce users to visit dermatology clinics and to buy products that are relevant to their hair condition

# | Value

- aim to provide personalized product recommendations
- can keep track of the user's hair condition
- helpful when getting diagnosed for alopecia (since users have records of the progress)

# Data from Kaggle & Review

## Dataset Overview :

This dataset encompasses data on potential contributors to baldness in individuals. Each entry represents a distinct person, with columns detailing factors spanning genetics, hormonal fluctuations, medical ailments, treatments, nutrient deficiencies, stress levels, age, hair care practices, environmental exposures, smoking habits, weight fluctuations, and the presence or absence of baldness.

## Purpose of the Dataset :

This dataset is designed for exploratory analysis, modeling, and predictive analytics endeavors, aiming to decipher the interplay between diverse factors and the probability of baldness in individuals.

## Data Review :

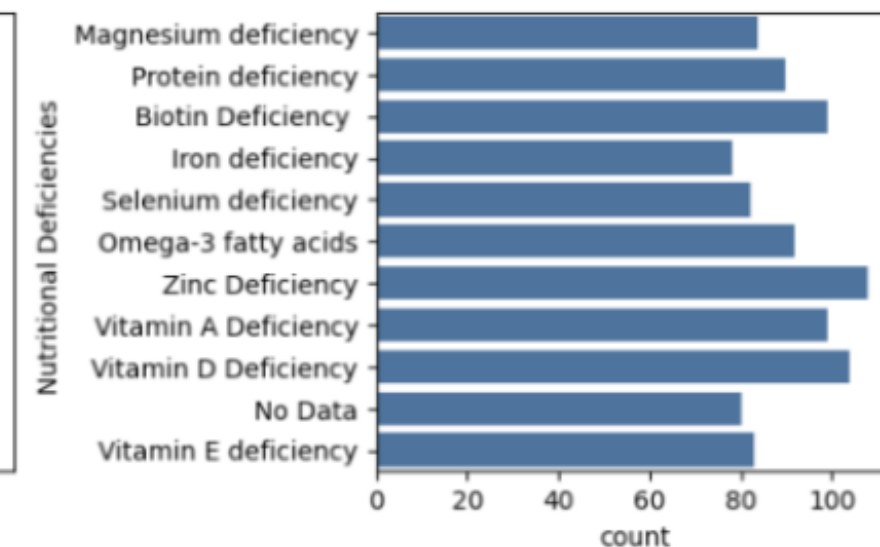
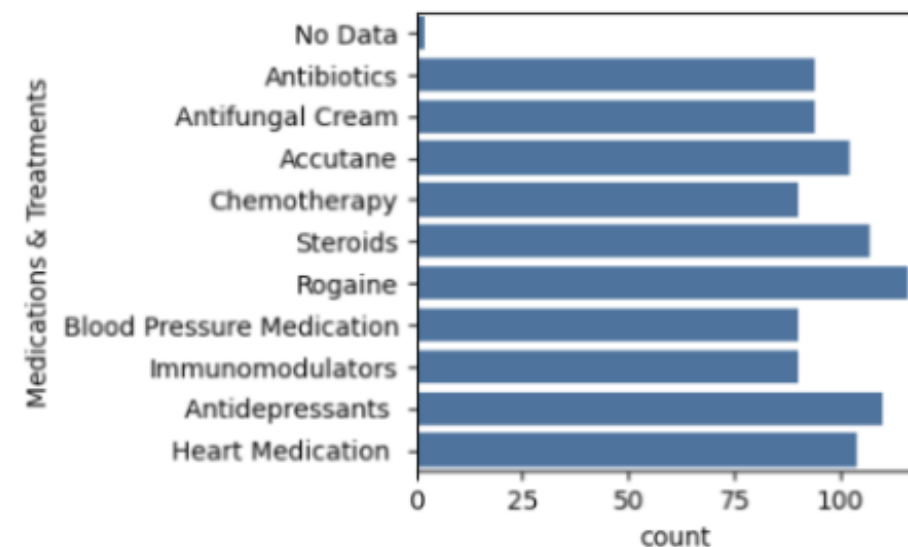
- separate train & test data
- target label is distributed by half
- there was no duplicated values
- data visualization using barplot, heatmap and histogram
- All other columns are also uniformly distributed
- Correct the typographical errors in the column names
- No outliers, deleted column 'id'

```
# the distribution of the categorical variable grouped by the target class

target_counts = df['Hair Loss'].value_counts()
target_ratio = df['Hair Loss'].value_counts(normalize=True)

print(f"target_counts ['0'] = {target_counts[0]} target_counts ['1'] = {target_counts[1]}")
print(f"target_ratio ['0'] = {round(target_ratio[0]*100,2)} % target_ratio ['1'] = {round(target_ratio[1]*100)} %")

target_counts ['0'] = 502 target_counts ['1'] = 497
target_ratio ['0'] = 50.25 % target_ratio ['1'] = 50 %
```



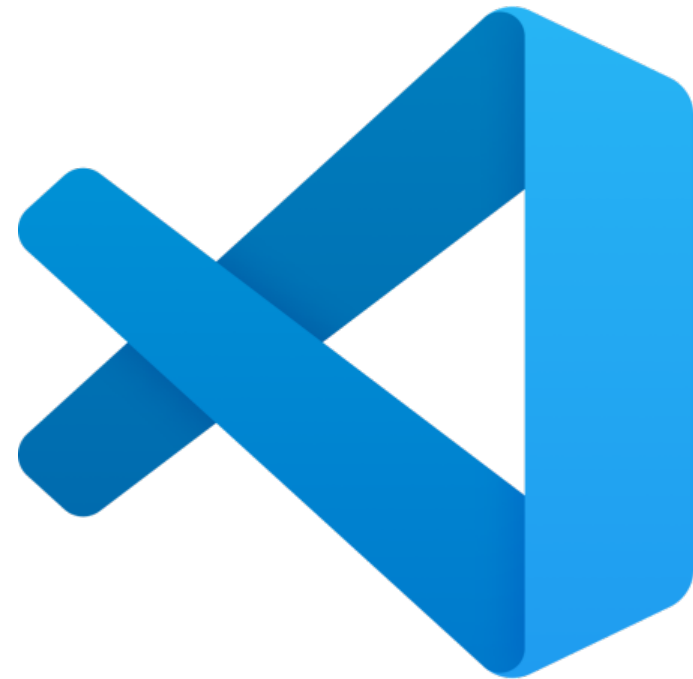


# | Formulation

Column Descriptions :

- 1. **Genetics:** Indicates a familial history of baldness (yes or no).
- 2. **Hormonal Changes:** Reflects whether the individual has undergone hormonal shifts (yes or no).
- 3. **Medical Conditions:** Enumerates specific ailments linked to baldness like Alopecia Areata, Thyroid Problems, Scalp Infection, Psoriasis, Dermatitis, etc.
- 4. **Medications & Treatments:** Lists drugs or therapies potentially causing hair loss, such as Chemotherapy, Heart Medication, Antidepressants, Steroids, etc.
- 5. **Nutritional Deficiencies:** Details deficiencies like Iron, Vitamin D, Biotin, Omega-3 fatty acids, etc. are associated with hair loss.
- 6. **Stress:** Indicates stress level (low, moderate, or high).
- 7. **Age:** Denotes individual age.
- 8. **Poor Hair Care Habits:** Indicates negligent hair care practices (yes or no).
- 9. **Environmental Factors:** Notes exposure to environmental elements linked to hair loss (yes or no).
- 10. **Smoking:** Specifies a smoking habit (yes or no).
- 11. **Weight Loss:** Indicates significant weight reduction (yes or no).
- 12. **Baldness (Target):** binary variable indicating baldness presence (1) or absence (0).

# | Execution Environment



```
import sys  
sys.version
```

✓ 0.0s

'3.10.10 (tags/v3.10.10:aad5f6a, Feb 7 2023,

We performed this code on our **VS Code on personal labtop**,  
not on Colab Notebooks.  
+ with python version 3.10.10



# | Formulation

## Input & Output Variables

Data columns (total 13 columns):

#	Column	Non-Null Count	Dtype
0	Id	999 non-null	int64
1	Genetics	999 non-null	object
2	Hormonal Changes	999 non-null	object
3	Medical Conditions	999 non-null	object
4	Medications & Treatments	999 non-null	object
5	Nutritional Deficiencies	999 non-null	object
6	Stress	999 non-null	object
7	Age	999 non-null	int64
8	Poor Hair Care Habits	999 non-null	object
9	Environmental Factors	999 non-null	object
10	Smoking	999 non-null	object
11	Weight Loss	999 non-null	object
12	Hair Loss	999 non-null	int64

## Model Selection

We blended following 5 methods :  
(these are types of ensemble models)

- Decision Tree Classifier
- K-Nearest Neighbor
- AdaBoost
- Logistic Regression
- Linear Discriminant Analysis

(type of Linear Dimensionality Reduction)

⇒ We opted for these models because they perform better than other models when most of the variables are categorical

# | Formulation

## Model Validation

We chose to do K-Fold Cross Validation

**K-Fold Cross Validation :**

A validation method that systematically changes the subsets to measure the model's performance across all data

This method is time-consuming compared to other validation methods. However, we still have opted for it due to the relatively small size of our dataset.

## Model Evaluation

- We selected F2 Score to test our model

$$F2\ Score = 5 \times \frac{Precision \times Recall}{(4 \times Precision) + Recall}$$

- Why?

In the given context, the mislabeling of a user as having a low likelihood of baldness (False Negative) carries a greater magnitude of loss compared to mislabeling them as having a high likelihood of baldness (False Positive).

# | Formulation

## Model Validation

We chose to do K-Fold Cross Validation

**K-Fold Cross Validation :**

A validation method that systematically changes the subsets to measure the model's performance across all data

This method is time-consuming compared to other validation methods. However, we still have opted for it due to the relatively small size of our dataset.

## Model Evaluation

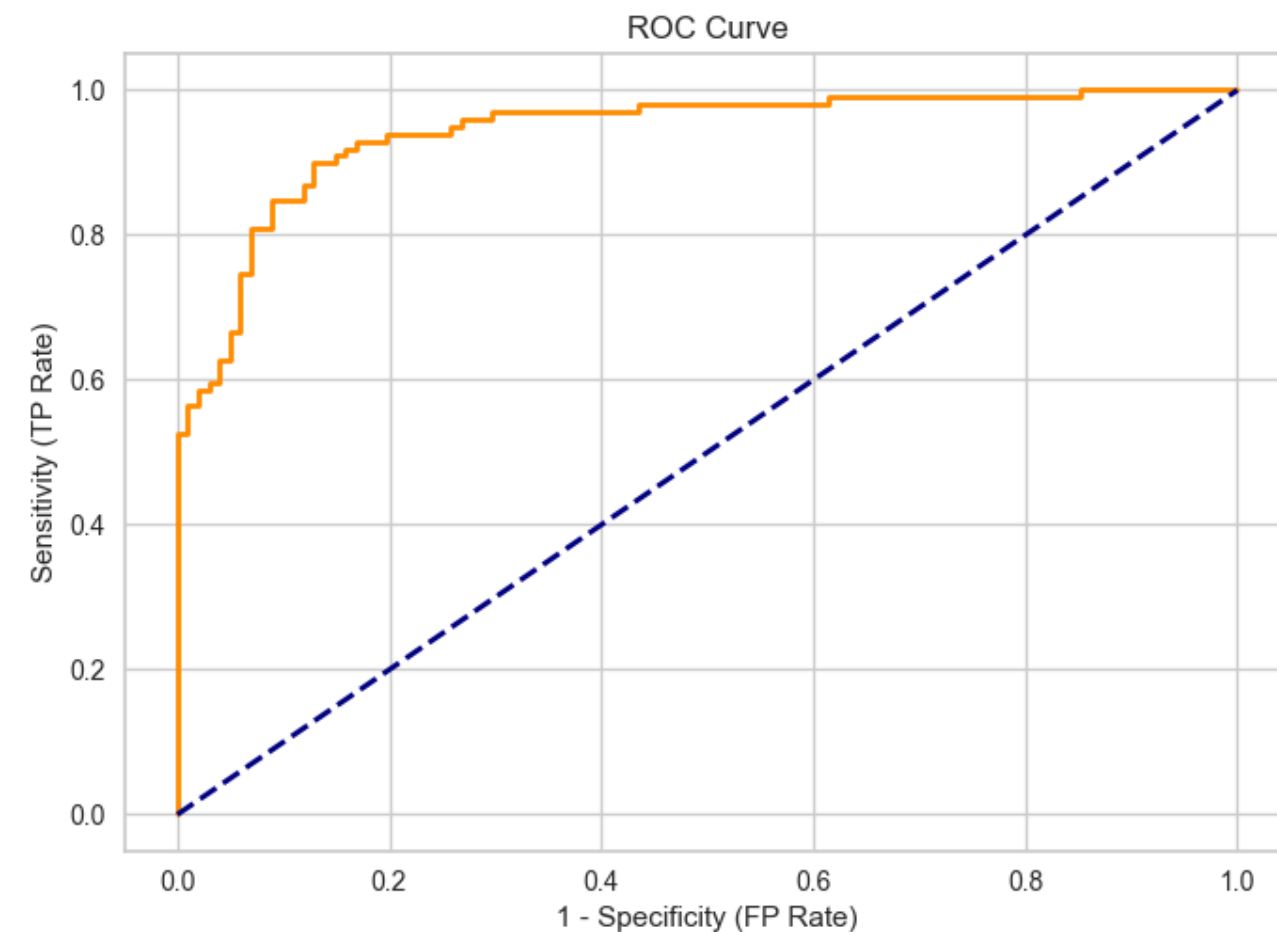
- We selected F2 Score to test our model

$$F2\ Score = 5 \times \frac{Precision \times Recall}{(4 \times Precision) + Recall}$$

		Predict	
		Positive (1)	Negative (0)
Actual	Positive (1)	C(1,1)	C(1,0)
	Negative (0)	C(0,1)	C(0,0)

Profit matrix

# | Model Performance



	Predicted yes	Predicted no
Actual yes	85	16
Actual no	9	90

By trained ensemble models,  
the F1, F2 score reaches around 0.9

Accuracy: 0.875  
F1: 0.8780487804878049  
F2: 0.896414342629482  
ROC AUC: 0.9424942494249424

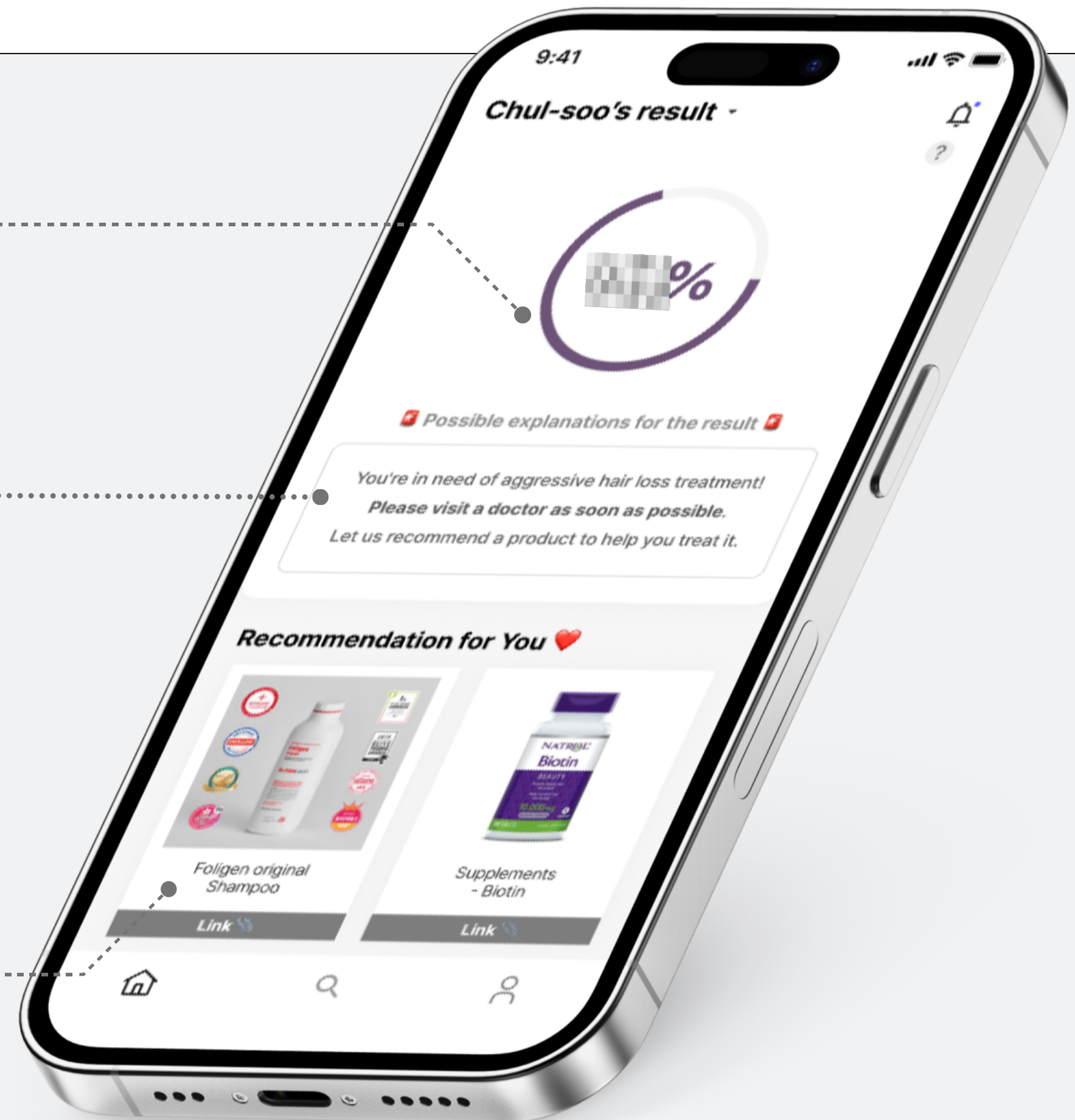
Compared with Former models,  
which had a figure of 0.6 or less  
→ improvement in performance

Display the user's probability of becoming bald

### Recommendation to visit a dermatology clinic

- 75~100% Assessed as severe hair loss
- 50~75% Assessed as hair loss at a certain stage of progression
- 25~50% Assessed as early-stage hair loss
- 0~25% - Assessed as a normal state

Personalized product recommendation



# | Conclusion

01

Business Revenue


- Personalized Product Recommendation
- Advertisements at our Service

02

Societal benefits

- Reducement of individual 's unnecessary consumption
- Improving Accessibility to Hair Loss Treatment
- Creating Added-Value on the Hair Loss Industry





Further Improvement

Introduce CNN Technique

Develop a service that utilizes CNN technique to accurately diagnose hair loss conditions by taking and uploading photos of user's scalp

# Remember Chulsoo?

01



**Chulsoo(31, male)**

Graduate Student

- hesitates to visit the hospital for treatment
- believes that it might not actually lead to baldness

## Input

- Genetics → NO
- Hormonal Changes → YES
- Medical Conditions → No data
- Medication → Accutane
- Nutritional Deficiencies → Magnesium
- Stress → High
- Poor Care Habits → YES
- Environmental Factors → YES
- Smoking → YES
- Weight Loss → YES



Probability of  
being bald

**57.3%**



**Thank you  
for listening**