The Effect of Mental Health and Lifestyle Habits on Happiness Level and Depression Risk
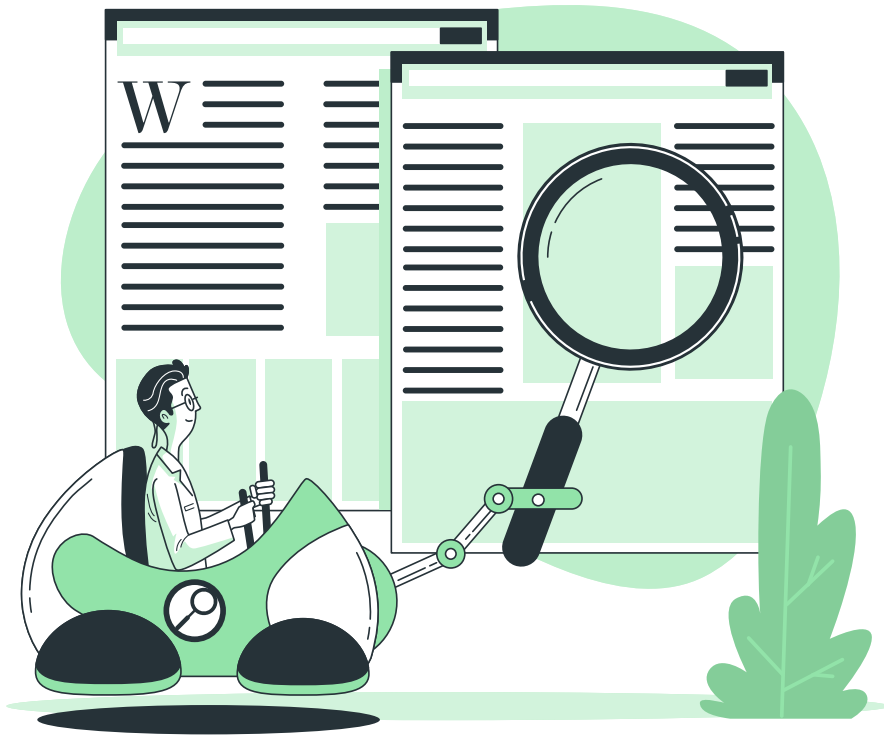
A REGRESSION AND CLASSIFICATION PROJECT

Enes Çalışkan
Şeymanur Şaşmaz
Nuran Kamışlı

Are your exercise level, diet type or screen time per day determining happiness or depression risk?

We tried to estimate Depression Risk and Happiness Score by looking only at daily habits **instead of traditional surveys.**

Dataset Overview

- Dataset from Kaggle, **3000 instances** and **12 feature** as orginally.

- The dataset is from 2019-2024(**COVID Period**), This is the biggest handicap of our data. It was hard times so it may not reflect nowadays.

- Some features from data;

Working Hours, Screen Time per Day, Social Interaction Score, Mental Health Condition, Exercise Level etc..

EDA Stage

The Problems Encountered

1) Null values in **Mental health Condition** (categoric data, about %20).
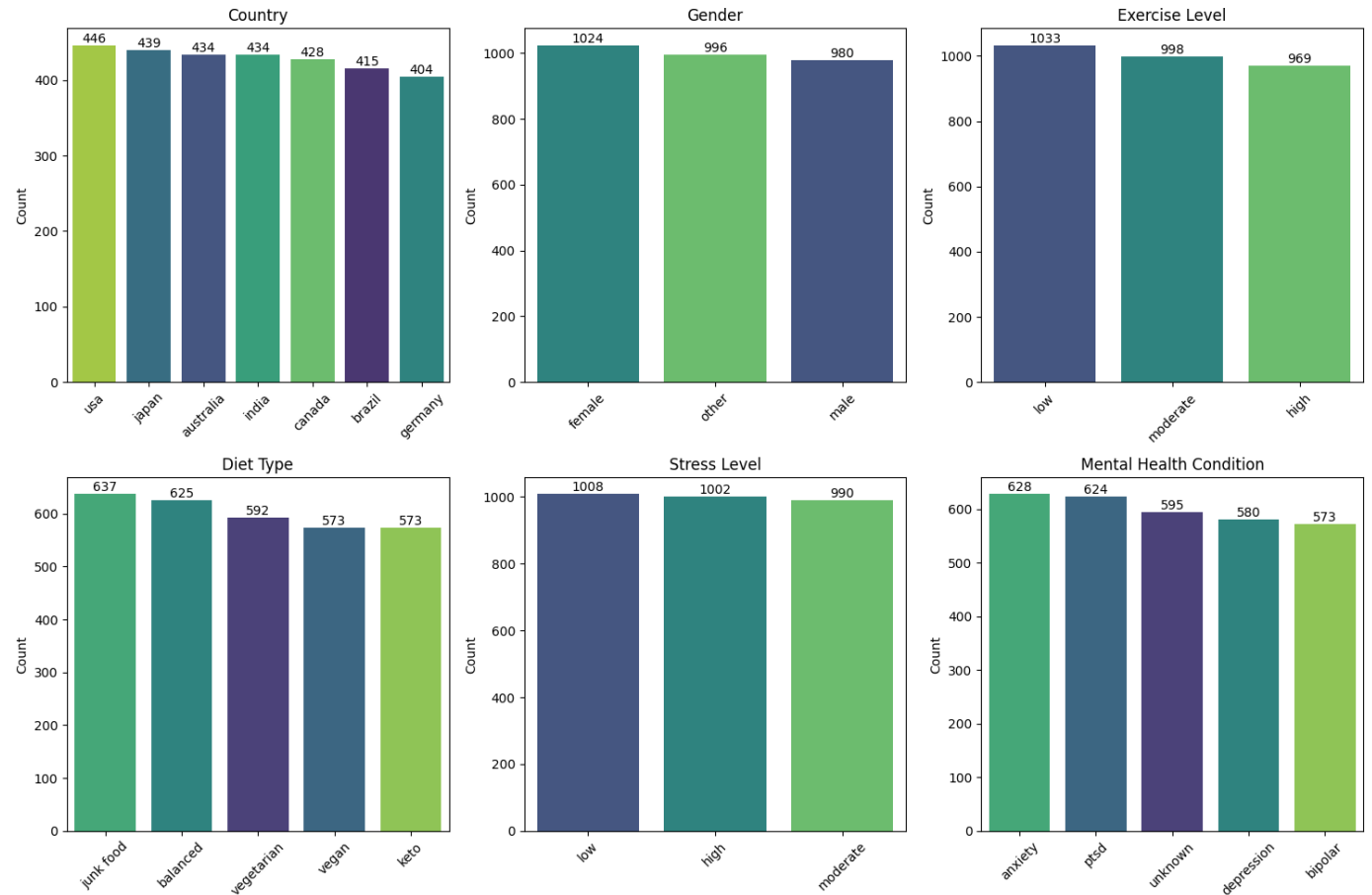
2) Outliers of **Sleep Hours**. (numeric data, 16 value)

How we solve it?

o We assign nulls as **'Unknown'**.If we delete all null values, the 20% of data has gone.If we asssign max or min counted value of it the distribution it would become unstable.

o Outliers aren't problem. Max is 11.3 hours, min is 1.4 hours and only 16 outliers are there. Extreme points is better for model learning of depression risk.
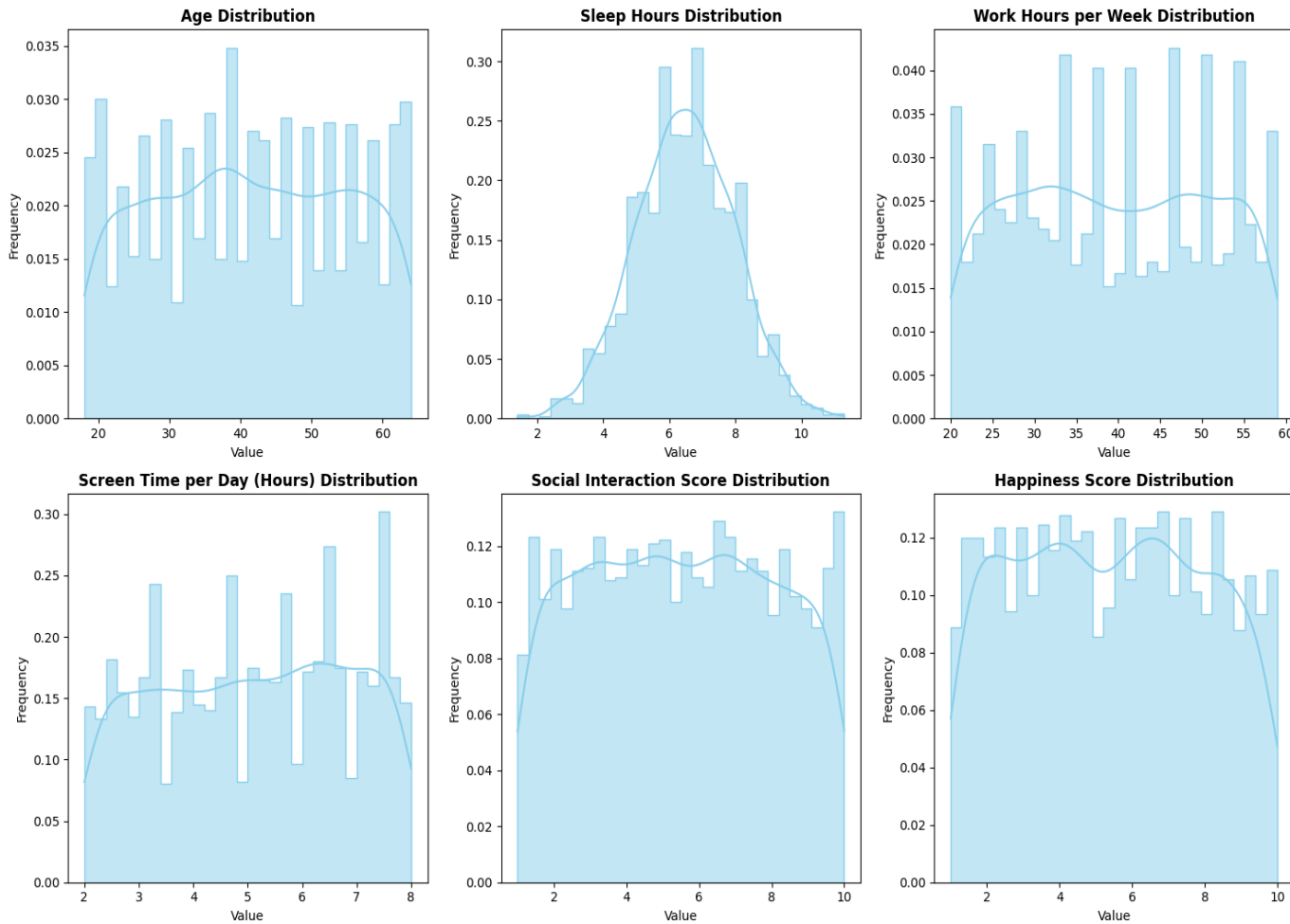
Preprocess Stage

This is our categoric variables distribution table

o The datas actually distributed very well. Ready for encoding

o As you see, 'Stress Level' and 'Exercise Level' have ordered values like 'Low','Moderate','High'. We apply ordinal encoding for it.

o Others has unordered values like 'Country', 'Age' etc. We use one-hot encoding for it.
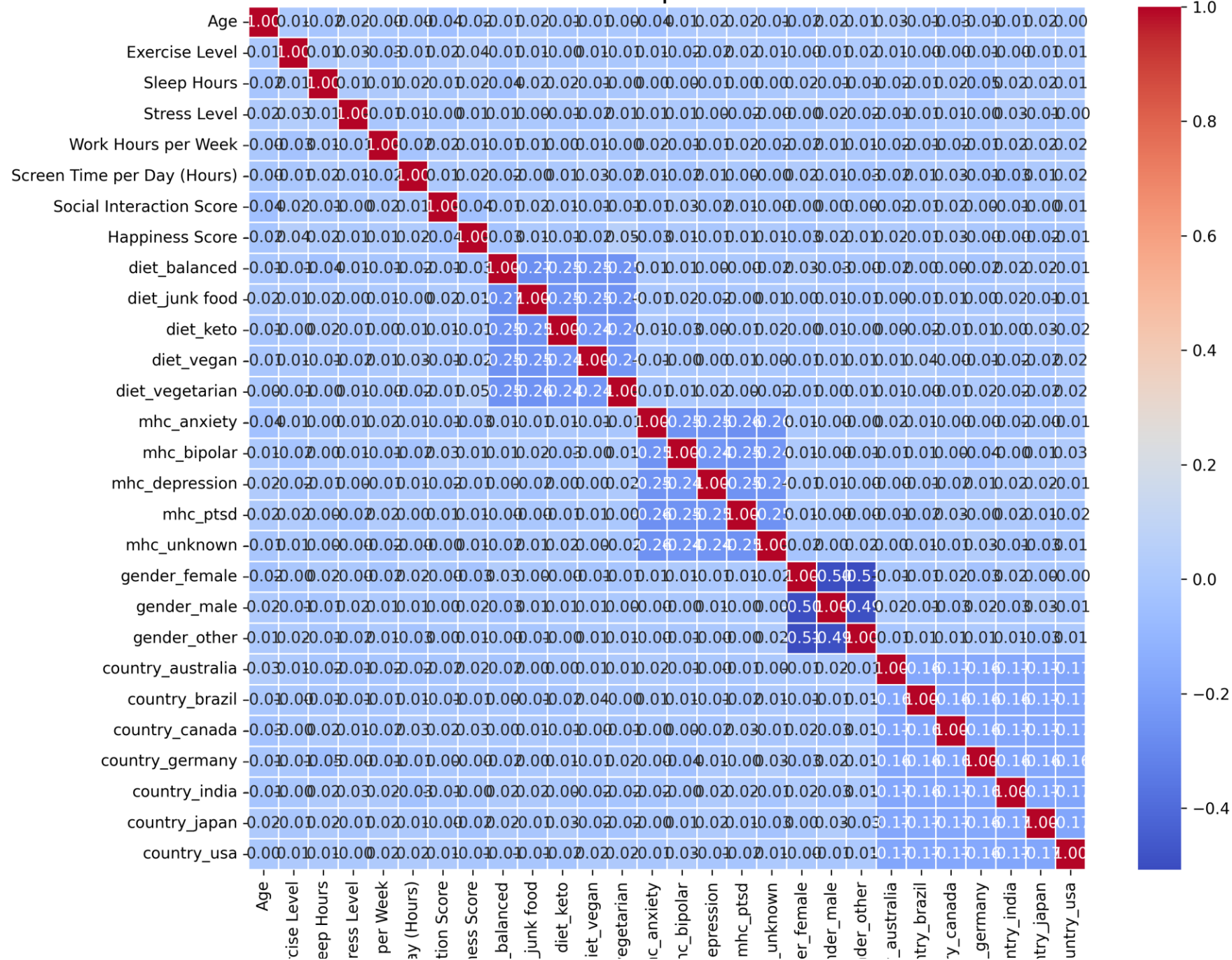
Preprocess Stage

o Except 'Sleep Hours', all numeric datas looks uniform distributed.

o 'Sleep Hours' is normal distributed.

o We used to scaling Min Max Scaler for all of numeric data.
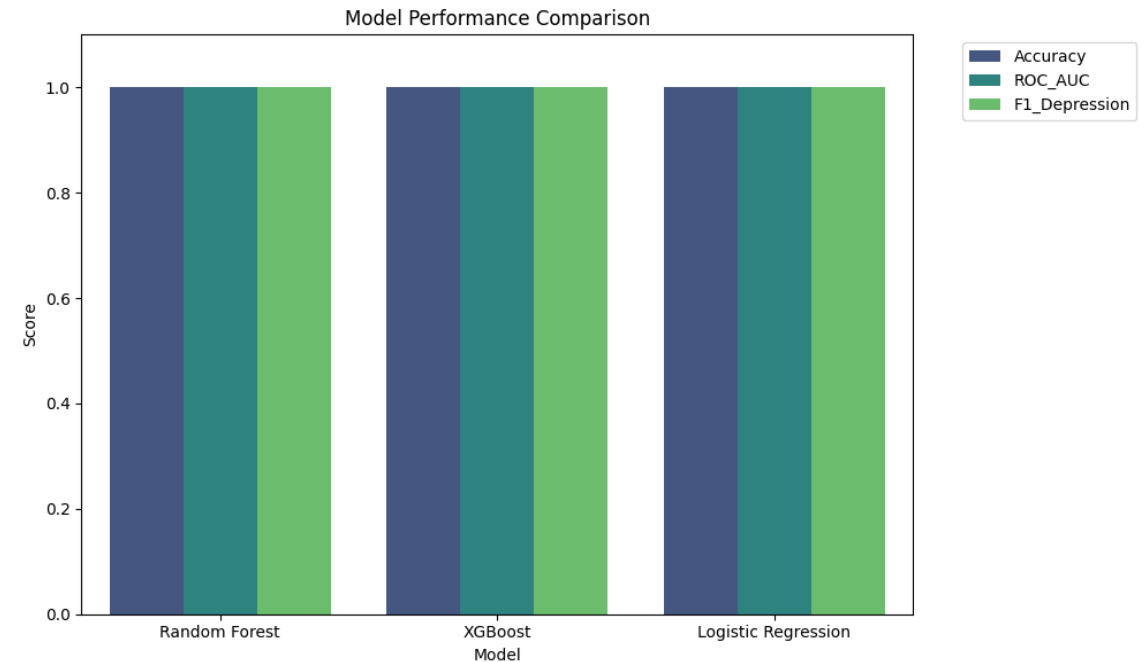
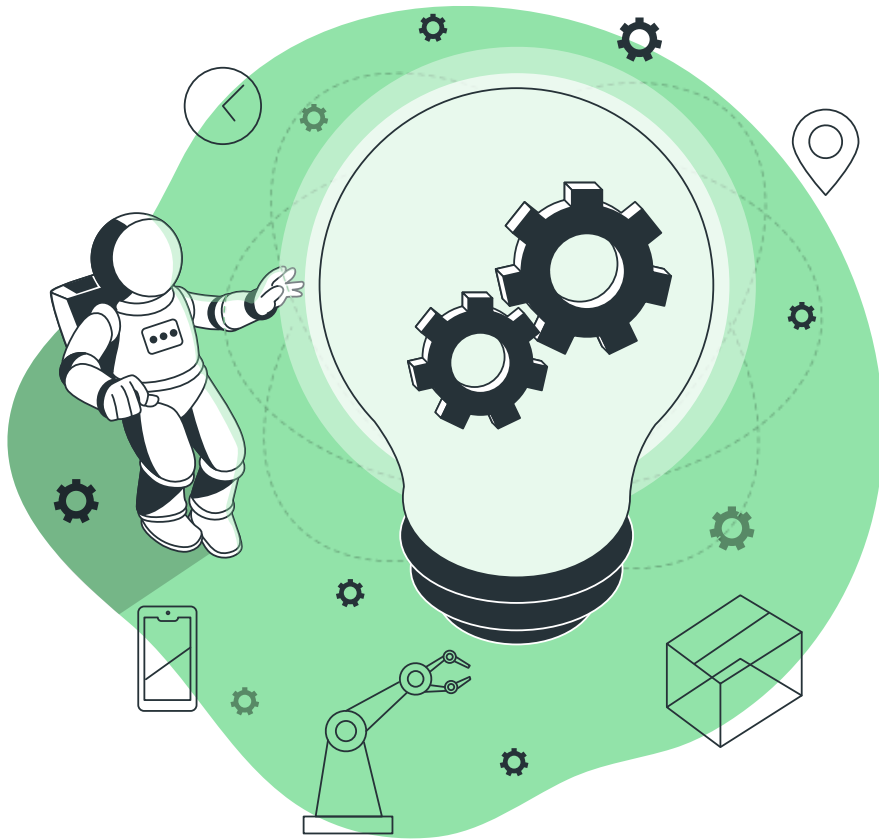Correlation Heatmap Between Variables

Weak Correlation

Depression Prediction

Data Leakage Problem

Data leakage occurs when a model has access to information during training that it would not have in a real prediction scenario, leading to overly optimistic performance.

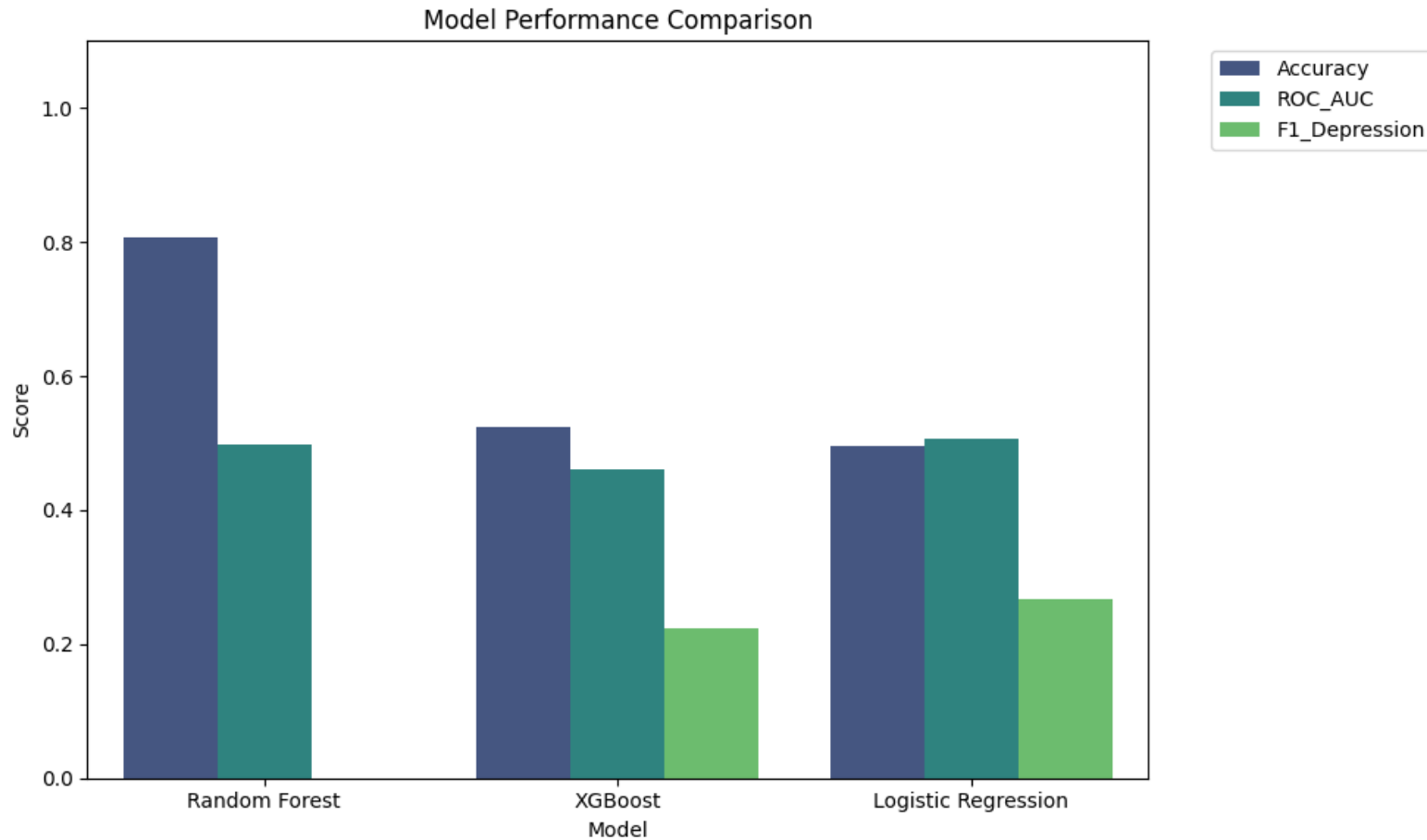If we hadn't removed the other mental health conditions:

Models used

- Logistic Regression

- Random Forest

- XGBoost Classifier

Results of Depression Prediction Models

Model Performance Comparison

The Random Forest model shows a relatively high accuracy of around 80%, but looking at the confusion matrix:

[[484 0]
[116 0]]

It fails to correctly predict any patients with depression. As a result, the F1 score equals 0.

Conclusion

The problem was not the models,it was the data .

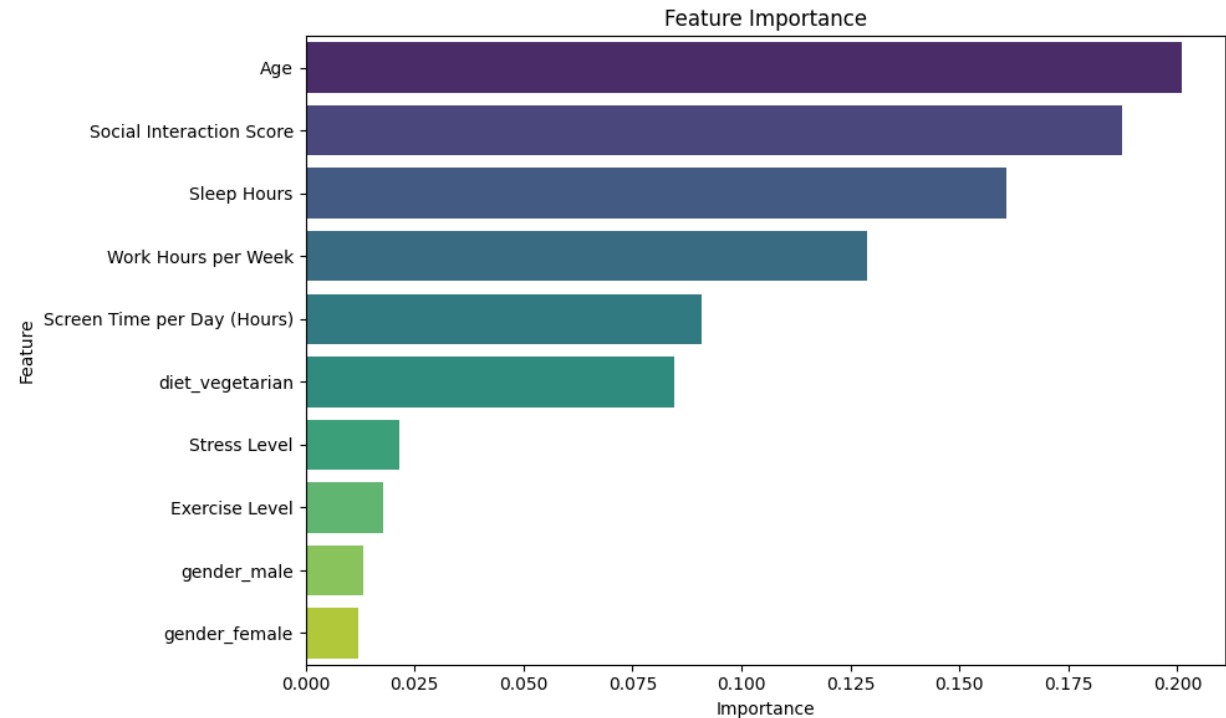The features don't give enough information to predict depression

Regression Problem

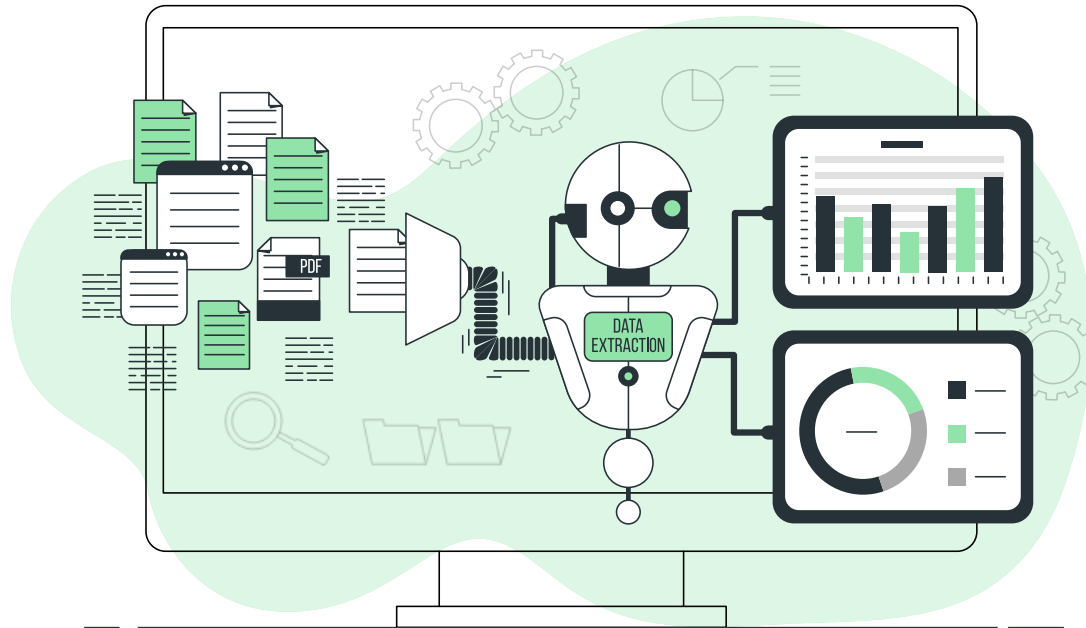We train regression models to predict happiness score.

Why did we use regression?

o Continuous Variable
o Understanding to reliable

Random Forest Feature Importance Table

Model Used

- Linear Regression

- Decision Tree

- Random Forest Regressor

- XGBoost

Results of Regression models

```
==================================================
            Regression Model Performance Summary
==================================================
                 Model  MSE (Hata)   R2 Score
       XGBoost (Tuned)    0.083133   0.005257
 Random Forest (Tuned)    0.083146   0.005099
     Linear Regression    0.083652  -0.000951
 Decision Tree (Tuned)    0.083933  -0.004315
==================================================
```
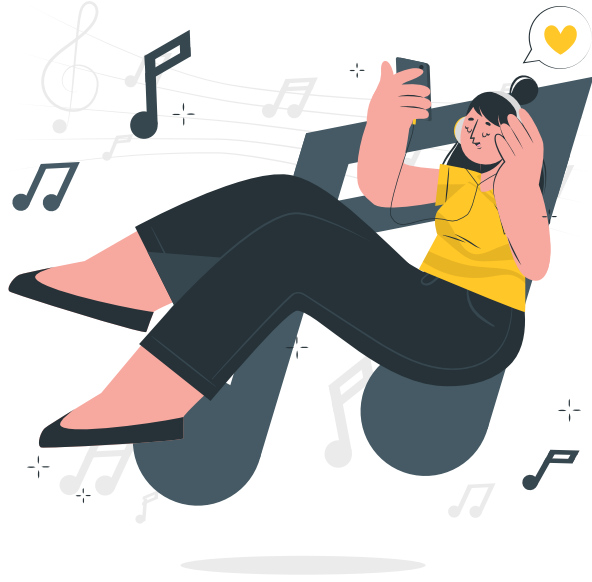
o There is no Convenient or Best Model. Results are very close each other.

o Models didn't train because r2 score shows that. (too low)

Conclusion

We understand that the poor performance not due to the models themselves.

Real reason is that the dataset doesn't contain strong predictor of happiness.

Real Life Prediction

She is **Ceren, female and from USA**

Age of **22**

8 Hours Sleep

20 Hours of Work per Week

2 Hours Screen Time per Day

Social Interaction Score is **9**

Happiness Score is **9**

Exercise **Level 3**

Stress **Level 1**

Diet Type is **Vegeterian**

Mental Health Condition is **Unknown**

Real Life Prediction

He is **Murat, male and from India**

Age of **58**

3 Hours Sleep

55 Hours of Work per Week

7 Hours Screen Time per Day

Social Interaction Score is **2**

Happiness Score is **1**

Exercise **Level 1**

Stress **Level 3**

Diet Type is **Junk Food**

Mental Health Condition is **Anxiety**

Real Life Prediction

We expect Ceren should be very happy in life according to her parameters.

Also the opposite side Murat should be very unhappy in life.

Result says Ceren is sligthly happy from Murat and there is no depression risk.

Even though there are two extreme human examples, the model is wrong. It has not been learned.

```
--- Predictions for User 1 ---
Happiness Score Prediction: 5.67
Depression Risk Prediction: 0

--- Predictions for User 2 ---
Happiness Score Prediction: 5.17
Depression Risk Prediction: 0
```