Group 32 Progress Report:
The Effects and Correlation Between Coffee Intake and Health

Anh Thieu, Swetha Anantha Krishnan
{thieup1,anants10}@mcmaster.ca

1 Introduction

This project aims to explore and model the relationships between coffee consumption, sleep patterns, and health outcomes using the Global Coffee Health Dataset from Kaggle. The dataset contains 10k records capturing demographics, lifestyle habits, and health metrics across 20 countries.

Coffee consumption is one of the most common daily habits worldwide. Almost every person today starts their day with a cup of coffee and has another cup within 3-4 hours. Although this is helpful as it keeps us awake and energetic, it leaves a serious effect on how our body functions that many ignore. These effects may not be immediately obvious, and might often show up once the individual is over 50 years old. This is a challenging problem because the effect of caffeine on sleep, stress, and long-term health remain complex and not yet fully understood. Therefore, this project is significant as understanding these relationships can help us decide how to adopt a better lifestyle and how much coffee is actually okay for an individual to consume, without affecting our sleep, stress, or overall health.

2 Related Work

There are existing solutions to our problems with some closely tied with machine learning, and others take more of a research and analysis approach. The first site was the closest match of data set and coffee machine learning prediction. The last 4 sites focused more on using machine learning to the correlations between caffeine consumption and certain health effects rather than machine model training on predicting features.

The most similar work we've encountered as a solution for predicting "Health Issues" and "Stress Level" with the same data set and different approach by *Rekhi*. This reference used RandomForestClassifier to train the model, whereas we took the XGBoost approach. There was also a difference in the approach to feature selection, the related documentation dropped both Stress_Level and Health_Issues from the features while we kept Stress_Level when predicting Health_Issues and vice versa. They are not closely related in terms of a duplicate feature or an unimportant feature so we see no reason to not keep it as a feature. Another reason we kept the feature is we are not sure of the importance the feature is to predicting the target, in this case, Stress_Level is found to be one of the higher importance to predicting Health_Issues.

Another site with related work on association between coffee intake and incident heart failure risk with machine learning by *Stevenset al.* (2021). They also used feature selection based on random forest analysis. The difference between their analysis is they identified potential risk factors associated with coronary heart disease and strokes while our dataset is more generalized with classification on health and stress issues.

The third related work site is a research article focused on the responsiveness pattern to caffeine using mathematical models and non-linear analysis by *Domingueset al.* (2024). Included in the report are classification models they used (Decision Trees, Discriminant Analysis, Logistic Regression, Naive Bayes, SVM, KNN, Ensembles) with classifiers and default parameters. This documentation's feature selection is based on the mathematical models and the database used were state to be collected in the "Human Neurobehavioral Laboratory (HNL) facilities at the Universidade Católica Portuguesa in Porto, Portugal".

The fourth site is also a research article with a different focus on caffeine's effect on cardiovascular properties with wearable biosensors and machine learning analysis by *Chowdhuryet al.* (2025). This experiment monitored cardiovascular behaviour under biosensor devices with 2 groups: pre-caffeine and post-caffeine participants. For machine learning, they used unsupervised k-clustering

to assess the separability of the two groups. The conclusion to their experiment states, "moderate caffeine intake induces disinhibition and coordinated neural-cardiovascular changes that are prominent and measurable by machine-learning methods," meaning there is some correlation between the intake of caffeine and health fluxuations.

Lastly, there was a research done correlating the negative side effects of caffeine consumption on children by $RichardsandSmith$ (2015), in which there was a positive relationship between these effects and children's caffeine intake. From this we can infer the relationship between the effects of caffeine consumption and heatlh.

3 Dataset

The dataset used in this project is synthetic_coffee_health_10000.csv, containing 10,000 samples and multiple health-related and lifestyle features. Each row represents an individual and includes features such as sleep habits, coffee consumption, and health issues. The target variable that we've chosen for now is Stress_Level. Later we also plan to incorporate another target, Sleep_Quality.

After loading the dataset, we first examined the data set shape (number of rows and columns), data types (categorical vs numerical), preview of the first rows (head), distribution of the target labels, and inspected for any missing values.

We found out that we had features with mixed data types, numerical and categorical. The categorical features were Gender, Country, Sleep_Quality, Stress_Level, Health_Issues, Occupation. The numerical features are ID, Age, Coffee_Intake, Caffeine_mg, Sleep_Hours, BMI, Heart_Rate, Physical_Activity_Hours, Smoking and Alcohol_Consumption.

3.1 Preprocessing Performed

Handling Missing Values
We identified missing values and filled them where appropriate. For example, Health_Issues had 5941 missing entries, but upon closer inspection, we found out that it was because one of the categories in the column was named None. So we renamed that categories to "No_issues" to avoid problems.

Encoding categorical features
We converted object-type columns into numeric values by mapping the categories to integers.

After this, we splitted our dataset into training and testing sets (80-20 split) before checking correlations to avoid test leakage.

4 Features

Our first approach for feature selection was using XGBoost's get_importance() function to evaluate the feature importance and gather the features the model relied on the most. Then, we took the 8 most important features as a feature selection step.

For classifying Stress_Levels, the get_importance() function showed Sleep_Hours and Sleep_Quality has the highest importance with both having an importance score ≥ 100 while the others were ≥ 0.17 so the top 2 features were included in the classification. See figure 1 for the Stress_Levels importance scores.

For classifying Sleep_Quality, the top 8 was taken even though the top 2 scores were overwhelming, the other scores had somewhat significant compared to the Stress_Level classification as these scores were ≥ 0.3 on the importance score meter. See figure 2 for the Sleep_Quality importance scores.

For classifying Health_Issues the top 8 scores were taken for the same reason as Sleep_Quality. The threshold for the feature selection is having an importance score ≥ 0.17 as selected by the overwhelming difference in the Stress_Level classification. See figure 3 for the Health_Issues importance scores.

We also tried Random Forest to double-check which features mattered the most. It gave us a similar ranking to XGBoost, which helped confirm that the top features we picked were actually strong predictors and not just random noise. Using Random Forest alongside XGBoost basically gave us more confidence that our feature selection was solid.

Continuing the improvements of the model, we plan to incorporate neural networks into the system.

5 Implementation

So far, the current models use XGBoost and Random Forest models in feature evaluation and training. The implementation of feature selection was discussed in the Features section so we will focus on the other parts of the model implementation in this section. After preprocessing and feature selection, we start training the model on the training data (80% of dataset).

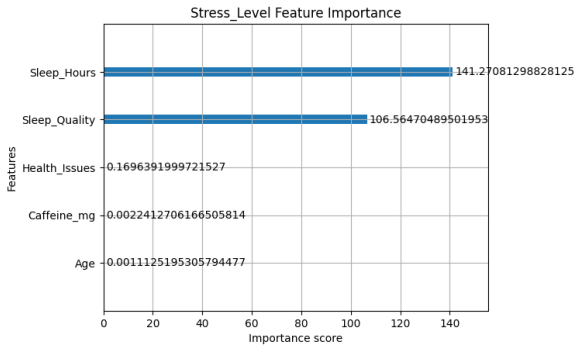

Figure 1: Stress_Level-Importance Chart

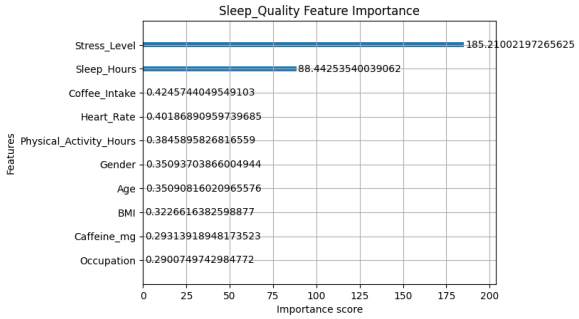

Figure 2: Sleep_Quality-Importance Chart

For hyper parameter tuning, we used RandomizedSearchCV with 20 random combinations and 3-fold cross-validation. After tuning, we used the tuned models to predict the fitted values for evaluation.

The loss model used in the XGBoost training is the Multiclass Logarithmic Loss, also known as the Cross-Entropy Loss for multi-class classification. This loss function is used in optimization method of the model, known as the second-order gradient boosting, or Newton Boosting. This method computes the gradient (first derivative of the loss / slope) and then the hessian (second derivative of the loss / curvature) and builds a decision tree with the leaf calculated as the the summation of the gradients over the summation of the hessians. In simple terms, the model makes a prediction tree, then the next prediction tree is a correction to the first tree's prediction errors, and so on.

The baseline from the most closely related work by *Rekhi*, where RandomForestClassifier was used with an accuracy score of 99.45% for predicting Health_Issues. Our model outperforms the related work's model with an accuracy score of 100% for training and 99.75% score for validation which is highly suspicious for overfitting. Another reason our model performed better than the related

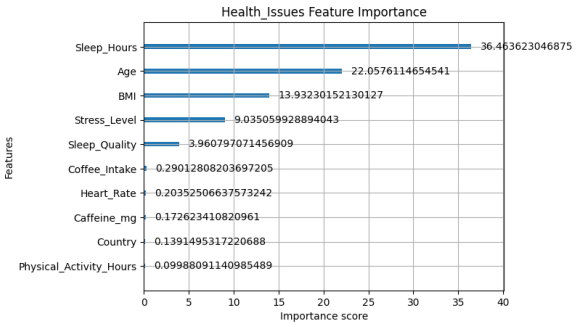

Figure 3: Health_Issues-Importance Chart

work's is because they dropped the Stress_Level feature from the feature vector even when predicting Health_Issues and through our feature evaluation, we found that Stress_Level had the highest importance score in terms of model gain.

We also added a Random Forest model next to compare its performance with XGBoost on the same dataset. Just like with XGBoost, we used feature importance from the fitted Random Forest model to pick the same top features for each target. We then ran RandomizedSearchCV on the Random Forest classifier using the same tuning setup (20 random combinations, 3-fold CV) to keep things consistent.

6 Results and Evaluation

To evaluate the model, we used the metric functions accuracy_score, confusion_matrix, ConfusionMatrixDisplay, and classification_report from sklearn. The accuracy_score and classification_report functions allow us to calculate the accuracy, precision, f1, recall, and support score, while the confusion_matrix function and ConfusionMatrixDisplay import allow us to visualize the correctness of predictions. We calculated and displayed these scores for both the training set and validation set. For XGB, the results are 100% accuracy score for both the training and validation dataset for predicting Stress_Level; 99.81% accuracy score for training and 99% for validation for predicting Sleep_Quality; and finally, 100% accuracy score for training and 99.75% for predicting Health_Issues.

For the Random Forest model, the results are 100% accuracy score for both the training and validation dataset for predicting Stress_Level; 100% accuracy score for training and 99% for validation for predicting Sleep_Quality; and finally, 100% accuracy score for both training and predicting

Health_Issues. This tells us that the Random forest model slightly outperformed the XGB model. See figures 4, 5, 6 for Stress_Level CMs, Sleep_Quality CMs, and Health_Issues CMs in that order.

Comparing to our baseline (the most closely related work) by *Rekhi*, our model performs a little better than theirs on predicting Health_Issues. Their model had an accuracy score of 99.45% while ours had an accuracy score of 99.75%. We suspect there is overfitting of features due to the near perfect results, and to counter this, we plan to incorporate neural networks with a dropout functionality.

7 Feedback and Plans

For the remainder of the project, we plan on incorporating a neural network model as recommended by the TA to improve the complexity of the project. With neural networks, we can experiment with dropouts and hyperparameter tuning in hopes of adjusting the model to deal with issues such as overfitting and over selection of features. Additionally, we plan to switch to one-hot encoding rather than label encoding to introduce no implicit ordering when working with neural networks. Label encoding is compatible with XGBoost since the trees care about the order for splitting and not arithmetic, but for linear or layered models (neural networks), we don't want to mislead the model with the meaning of the encoded numerals.

Team Contributions

Both team members contributed evenly to the progress of this report. We created the outline and planned the objective as a whole before initializing the code. For particular contributions, Swetha started the project with preprocessing and encoding of the vectors. Anh took over with initialization of XGBoost and hyper parameter tuning. Then, Swetha created the evaluation models and functions, and also added the Random Forest model for comparison. In the end, both members took turn cleanning up and finalizing the code.

For contributions to writing the report, we divided the report based on the part of the code the members worked on individually. Swetha arranged and wrote the sections as follows: Introduction, Dataset, Feedback and Plans, and Results and Evaluation. On the other hand, Anh wrote the following sections: Related Work, Features, Team Contributions, and Implementation.

References

Shabbir Chowdhury, Ahmed Munis Alanazi, and Eyad Talal Attar. 2025. Caffeine on the mind: Eeg and cardiovascular signatures of cortical arousal revealed by wearable sensors and machine learning—a pilot study on a male group. *Frontiers in Systems Neuroscience*, Volume 19 - 2025.

Rita Domingues, Patrícia Batista, Manuela Pintado, Patrícia Oliveira-Silva, and Pedro Miguel Rodrigues. 2024. Evaluation of the responsiveness pattern to caffeine through a smart data-driven ecg non-linear multi-band analysis. *Heliyon*, 10(11).

Rekhi. Global coffee health — kaggle.com. https://www.kaggle.com/code/seki32/global-coffee-health.

Gareth Richards and Andrew Smith. 2015. Caffeine consumption and self-assessed stress, anxiety, and depression in secondary school children. *Journal of Psychopharmacology*, 29(12):1236–1247. PMID: 26508718.

Laura M. Stevens, Erik Linstead, Jennifer L. Hall, and David P. Kao. 2021. Association between coffee intake and incident heart failure risk. *Circulation: Heart Failure*, 14(2):e006799.

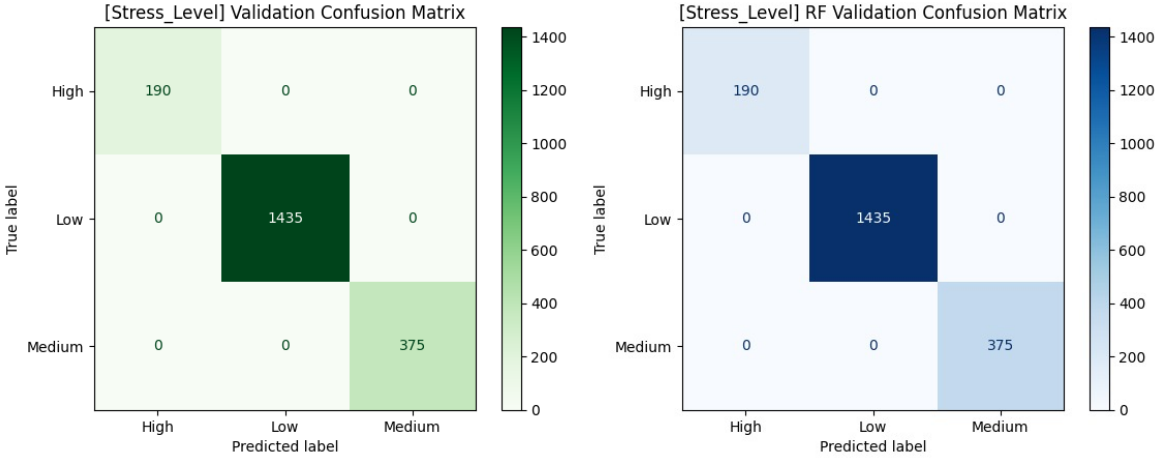

Figure 4: Stress_Level Validation Confusion Matrices for both models

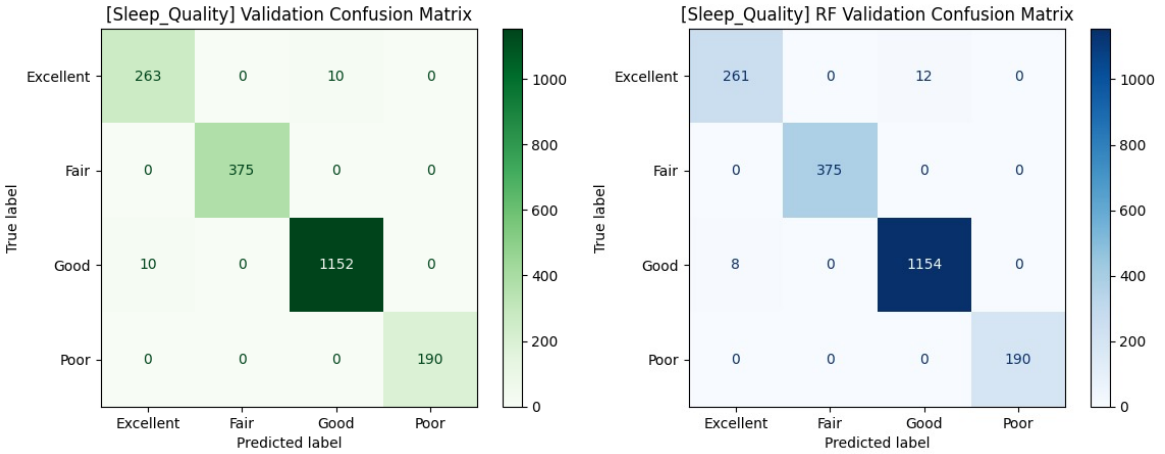

Figure 5: Sleep_Quality Validation Confusion Matrices for both models

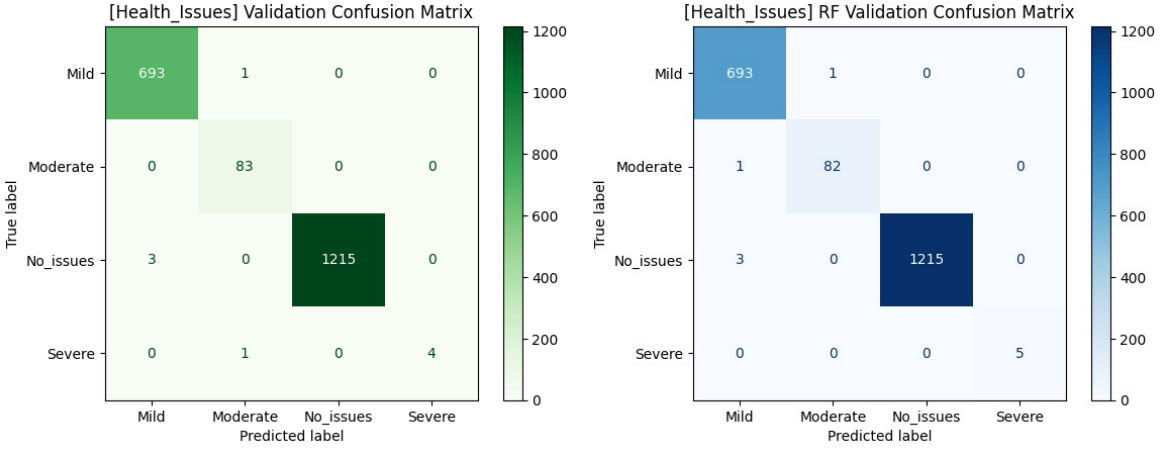

Figure 6: Health_Issues Validation Confusion Matrices for both models