Group name: Data Dragons

Team members: Josh Garman, Katherine Min, Heleyna Tucker

STATS 503: Predicting Overconsumption of Caffeine and Inadequate Sleep

1. Introduction

The National Health and Nutrition Examination Survey (NHANES) provides various data about people's mental state, nutrition, demographics, among other categories. Based on this provided data, the group decided on two questions. The first question seeks to answer how well can we predict caffeine consumption (in mg) will be above the daily recommended limit (400mg) with demographic, nutritional, and sleep information. The second question is how well can we predict sleep hours, do they get the recommended amount (more than 7 hours) a night with demographic, nutritional, and sleep information. These questions interested the group because sleep and caffeine are essential parts to most people's everyday lives, majorly affecting students. Getting the proper amount of sleep can be difficult, but are there ways we can predict if someone will get good or bad sleep based on other factors in their life. Caffeine intake can vary with different people, but are there key factors that contribute to a person going over the recommended amount? These questions will be addressed and explored in this report.

2. Data

We decided to use 3 datasets from the NHANES: Total Nutrient Intakes, First Day (P_DR1TOT), Sleep Disorders (S_SLQ), and Demographic data (P_DEMO). All these datasets were taken from 2017-2020 survey format questionnaires, initially having 20,000 participants. After data cleaning, there were approximately 8,000 data points. Since each dataset has a large amount of data, we needed to select variables that we thought would be relevant to our two questions. In other words, we picked variables that would affect sleep and caffeine intake.

From each dataset, we needed to take the SEQN number in order to merge the three datasets together. From the demographic, nutrition, and sleep data we chose the following variables shown in Tables 1, 2, 3, and 4:

Demographic Data	**Race/Hispanic origin**	**Education Level**	**Marital status**
1	Mexican American	Less than 9th grade	Married/living with partner
2	Other Hispanic	9-11th grade	Widowed/Divorced/ Separated
3	Non-Hispanic White	High school grad/GED or equivalent	Never married
4	Non-Hispanic Black	Some college or AA degree	-
5	Other Race	College graduate or above	-

Table 1: Demographic data information.

Nutrition Data	**Energy (kcal)**	**Vitamin B12 (mcg)**	**Total sugars (gm)**	**Dietary fiber (gm)**
Range of values	0 - 12,501	0 - 249.71	0 - 931.16	0 - 107.8

	Sodium (mg)	Caffeine (mg)	Alcohol (gm)	
Range of values	0 - 25949	0 - 4320	0 - 1152.8	

Table 2: Nutrition Data Information.

Sleep Data (Categorical)	Ever told doctor had trouble sleeping?	How often do you snore?	How often do you snort or stop breathing?	How often feel overly sleepy during day?
0	-	Never	Never	Never
1	Yes	Rarely (1-2 nights/week)	Rarely (1-2 nights/week)	Rarely (1 time/month)
2	No	Occasionally (3-4 nights/week)	Occasionally (3-4 nights/week)	Sometimes (2-4 times/month)
3	-	Frequently (5+ nights/week)	Frequently (5+ nights/week)	Often (5-15 times/month)
4	-	-	-	Almost always (16-30 times/month)

Sleep Data (Quantitative)	Usual sleep time on weekends or workdays	Sleep hours
Range of values	HH:MM - 00:00 to 23:30	3 - 13.5

Tables 3&4: Sleep data information.

After merging the data together, it is important to get a good idea of the relationship and initial trends. We can do this via the graphs below. There are more graphics that can be made between various variables, below are a few that had important information to note.

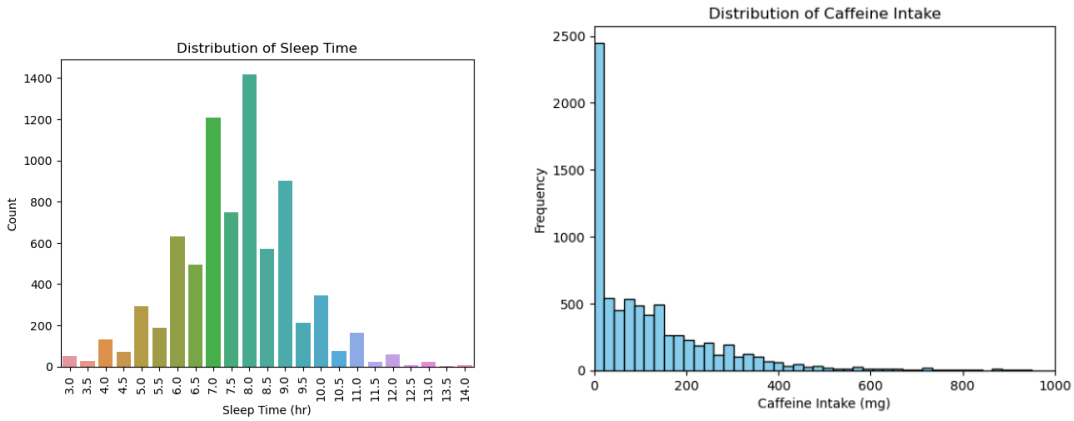


Figure 1: Distribution of Sleep and Caffeine Intake (excluding data points that exceeded 1000mg, or extreme outliers).

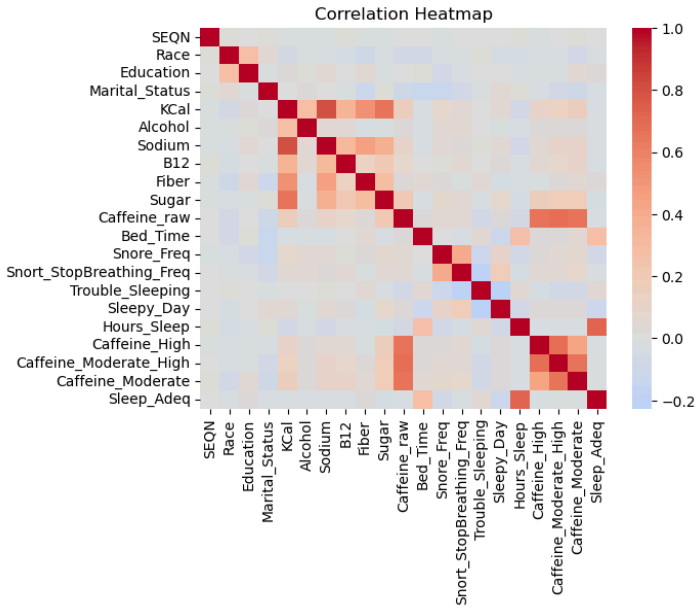


Figure 2: correlation heatmap that shows correlation values between variables within the merged dataset, ranging from -0.2-1.

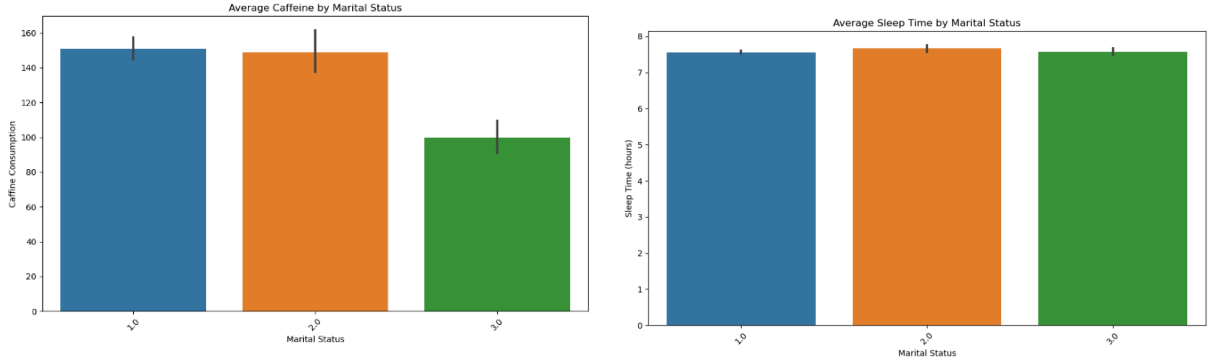


Figure 3: Bar plots displaying the relationship between marital status and both caffeine consumption and sleep time.

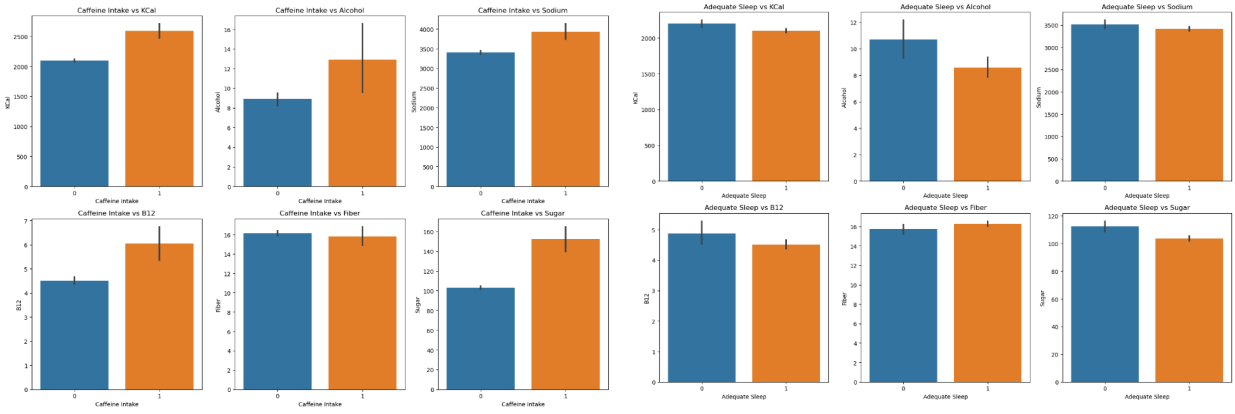


Figure 4: Bar plots displaying the relationship between low/high(0/1) caffeine consumption (high is when caffeine is above 400 mg) and bad/good(0/1) sleep (good sleep is when sleep hours is above 7) with Energy (KCal), Alcohol, Sodium, Vitamin B12, Fiber, and Sugar.

The distributions of the variables we are interested in predicting can be found in Figure 1. This tells us that Sleep time has a bell-shaped (normal) distribution and caffeine intake has a right skewed distribution. In the correlation heatmap from Figure 2, we can notice some specific relationships between our variables. For example, Caffeine consumption looks to be highly correlated with energy and sleep hours is correlated with the time one went to sleep. Looking further into these relationships, Figure 3 shows that caffeine consumption goes down when the marital status is 3 (never married). Figure 4 helps us get a better understanding of how nutrients can affect both caffeine and sleep. For example, those who have higher caffeine intake (above 400mg) have more energy, alcohol, sodium, vitamin B12, and sugar in their diet and those who have better sleep have less alcohol in their diet. After data visualization, the quantitative variables were scaled and categorical variables were one-hot-encoded.

3. Methods

For both questions, we decided to use five models: K-Nearest Neighbors (KNN), Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Random Forest. These models were used because of their various complexities. KNN, Logistic Regression are not as complex, while LDA and QDA have a moderate complexity and Random Forest is more complex. Each model was tested with and without interaction terms, with and without up-sampling and tuned. We utilized up-sampling, because only 5% of the observations have caffeine consumption above the 400mg threshold, and roughly 25% of the observations have inadequate sleep.

For the first question, the 400mg high/low level caffeine classification was modeled without interaction, with interaction, and then with bootstrap up-sampling to address class imbalance. The performance of these models at the different levels were analyzed. For the second question, adequate sleep was classified using all five models with no interaction, then no interaction with up-sampling The models were also tested using interaction terms with and without up-sampling. These models were compared to find which combination performs the best. The equations for each model can be found below. We utilized the following models (where X is our predictor vector and x_i are the elements of the vector) :

Model	Equation Form
KNN	$f(X) = Pr(Y = j \mid X = x_0) = argmax_{y_i} \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$
Logistic Regression Model (* subject to L_1 penalty for fitting)	$f(X) = log(\frac{p(x)}{1-p(x)}) = \beta X$ (* fitted to minimize with respect to also minimizing $\frac{1}{c}\sum_{i}^{p} \lvert\beta_i\rvert$)

LDA	$f(X) = log(\frac{Pr(Y=k \mid X=x)}{Pr(Y=K \mid X=x)}) = a_k + \sum_{j=1}^{p} b_{kj} x_j$
QDA	$f(X) = log(\frac{Pr(Y=k \mid X=x)}{Pr(Y=K \mid X=x)}) = a_k + \sum_{j=1}^{p} g_{kj}(x_j)$
Random Forest (RF) (* utilizing balance class weight for fitting)	$f(X) = \sum_{m=1}^{M} c_m * 1_{(X \in R_m)}$ where R_m corresponds to one of trees, and c_m is the respective output of the tree

We utilize the fields defined above as predictors for both models, with a few exceptions. Firstly, in the caffeine model we utilize the hours sleeping as a quantitative variable (not to be confused with the adequate sleep variable), and we do not utilize the caffeine quantitative variable. Secondly, in the sleep model, we utilize the caffeine consumed quantitative variable (not to be confused with the caffeine categorical variable), and we do not utilize the quantitative sleep variable.

4. Results

For the following models, we fit them on a training set of data, corresponding to 80% of the total data. We then calculated the accuracy, true positive rate (TPR), and the true negative rate (TNR) with the held out test set of data.

I. Caffeine Consumption

We had 4 different modeling approaches to the prescribed models. The approaches are as follows: (1) a model predicting caffeine consumption over 400mg with uneven class distributions, (2) an approach utilizing up-sampling in order to create equal class proportions for the predicted class, repeating of (1) and (2) utilizing interaction features.

The 2nd approach produced the best models, and is displayed below. The other approaches have been placed in the appendix for completeness.

Class-Balanced Model with 400mg Threshold:

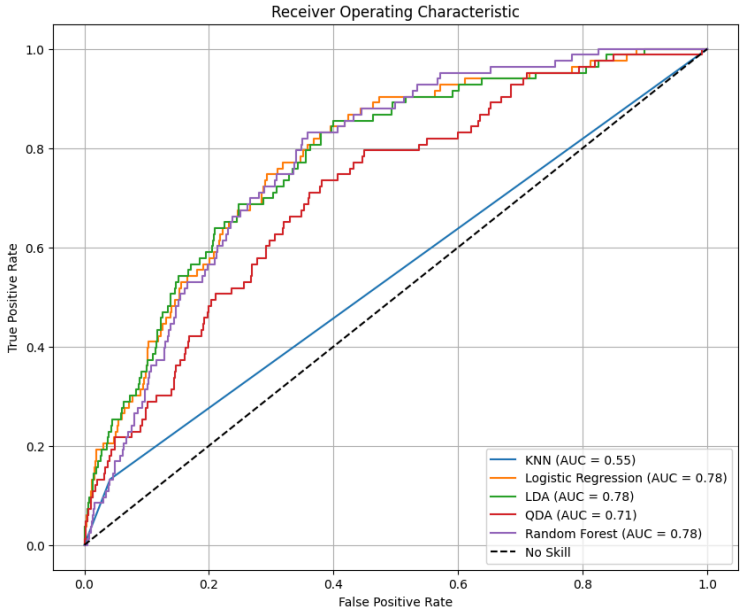
Model	**Accuracy**	**TPR**	**TNR**
KNN(K=1)	0.91	0.13	0.96
Logistic Regression(α=1/0.1)	0.71	0.72	0.71
LDA	0.70	0.70	0.70
QDA	0.86	0.29	0.89

Random Forest(Trees=600, Max Depth=10)	0.80	0.53	0.82

Notably, most of these models are not ideal. LDA and random forest do stick out from the others in their cohort. LDA is our preferred model, because it does achieve the highest TPR (our main metric of concern) and it does so with a reasonable expense of accuracy. Random forest also does somewhat well, however its TPR is considerably worse.

Utilizing up-sampling via bootstrap, we were able to achieve much better results than without. It appeared that the interaction terms introduced far too much noise to the model than they provided in predictive power.

Below is the ROC curve for these models:



As we see in the ROC curves, the models do reasonably well at predicting with AUCs greater than 0.70 for all but KNN. However, they do not strike the desired balance between predictive power and TPR.

II. Sleep

Similar to the caffeine problem, we had 4 different modeling approaches. The approaches are as follows: (1) a model predicting sufficient sleep (< 7 hours) with uneven class distributions, (2) an approach utilizing up-sampling in order to create equal class proportions for the predicted class, repeating of (1) and (2) utilizing interaction features.

The 4th approach produced the best models, and is displayed below. The other approaches have been placed in the appendix for completeness.

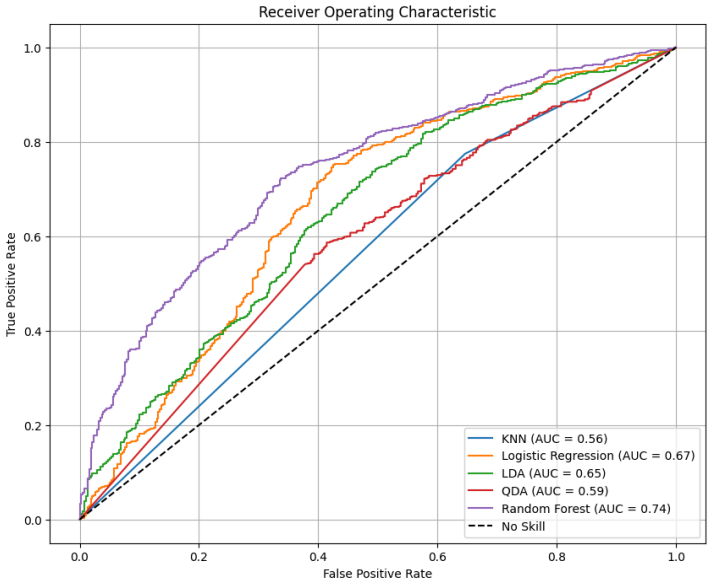
<u>Class-Balanced Model with Interaction Terms:</u>

Model	Accuracy	TPR	TNR
KNN(K=1)	0.67	0.77	0.35
Logistic Regression(α=1/0.01)	0.71	0.77	0.54
LDA	0.66	0.69	0.54
QDA	0.59	0.59	0.58
Random Forest(Trees=300, Max Depth=10)	0.73	0.79	0.53

Similar to the previous question, none of these models are perfect. The most important for this modeling endeavor is the TNR (negative in this context is not having adequate sleep). Logistic regression and random forest do similarly well in this context. While LDA and QDA achieve similar results in terms of TNR, it comes at the expense of accuracy.

Interestingly, in this approach there was a benefit to utilizing interaction terms. To truly try to understand why this is the case would require more modeling. It is unsurprising that up-sampling performs better for our desired metric, due to there being a class imbalance in the sleep adequacy target variable.

Our obvious preference in models is random forest and logistic regression. We do give slight preference to logistic regression due to it being a more simplistic model than random forest, and having a marginally better capacity for inferential analysis. Below is the ROC curve for these models:



We note here that the AUC for these models is considerably worse than caffeine models. However, random forest and logistic regression perform reasonably well. As stated above we do wish that the TPR was higher for these models, however they don't have incredible predictive power to begin with.

5. Conclusions

For the first question regarding caffeine intake, the up-sampling model demonstrated the best overall performance in terms of accuracy, sensitivity, and specificity. The random forest model with 600 number of trees, and maximum depth of 10 showed a strong performance of minimizing false negatives, while showing a significant improvement in sensitivity. For the second question regarding adequate sleep, the up-sampling model with interaction terms demonstrated the best model performance. Similarly, the logistic regression with interaction terms showed a balanced and strong performance in all metrics.

We treated the modeling of this like we would for a disease in a medical problem by giving a very high preference for predicting the incidence of a negative outcome. Our models did reasonably well, but left room for improvement. Something we keep in mind is that unlike a disease, these targets are largely driven by human behavior, which makes incidence much harder to truly predict. Additionally, the immediate severity of misclassification in our problem is not as dire as it is in most medical settings.

This analysis may be misleading for various reasons. First, we did not explore all the given variables from each dataset to model, we selected what variables we thought would be relevant. This brings the question of feature selection, could there be variables that have more importance when predicting caffeine consumption and sleep and are they included in this analysis. Exploring more features could give us a better prediction. In addition, this data is only based in the United States population. Therefore, these findings could only be related to the American population, not other countries. There could also be some overfitting in the model, making our test accuracy decrease compared to the training accuracy. There was also a lot of missing data within the datasets which made our sample size smaller. This could lead to a worse prediction.

6. Contributions

This project was a group effort, initial data cleaning was done by each team member and the initial data visualization was done by Heleyna Tucker. Josh Garman and Katherine Min did a lot of the modeling code and visualization for the models performance. The report was done collaboratively, with all team members writing and editing parts.

7. Reproducibility

The submitted code includes all the steps outlined in this report. If it is wished to reproduce this data, download the given datasets and redirect to where they were downloaded in your personal directory, then run the code. There are seeds set throughout in order to obtain the same results outlined in this report.

Appendix

Baseline Model with 400mg Threshold:

Model	Accuracy	TPR	TNR
KNN(K=1)	0.91	0.13	0.96
Logistic Regression(alpha=1/10)	0.95	0.06	1.00
LDA	0.94	0.08	1.00
QDA	0.65	0.70	0.64
Random Forest(Trees=600, Max Depth=10)	0.87	0.30	0.90

Baseline Model with 400mg Threshold with Interaction:

Model	Accuracy	TPR	TNR
KNN(K=1)	0.91	0.11	0.96
Logistic Regression(alpha=1/10)	0.93	0.07	0.98
LDA	0.92	0.13	0.97
QDA	0.95	0.00	1.00
Random Forest(Trees=600, Max Depth=10)	0.88	0.29	0.92

Class-Balanced Model with Interaction Terms

Model	Accuracy	TPR	TNR
KNN	0.91	0.11	0.96
Logistic Regression(alpha=1/100)	0.78	0.42	0.80
LDA	0.74	0.49	0.75
QDA	0.95	0.00	1.00
Random Forest(Trees=600, Max Depth=10)	0.82	0.47	0.84

Appendix 1: The above tables show TPR and TNR for the baseline model with 400mg threshold, baseline model with 400mg threshold with interaction, class-balanced model with interaction terms respectively.

Baseline Model

Model	Accuracy	TPR	TNR
KNN(K=29)	0.76	0.99	0.06
Logistic Regression(α=1/0.01)	0.75	0.96	0.11
LDA	0.76	0.93	0.22
QDA	0.73	0.95	0.08
Random Forest(Trees=300, Max Depth=10)	0.74	0.84	0.42

Baseline Model with Interaction Terms

Model	Accuracy	TPR	TNR

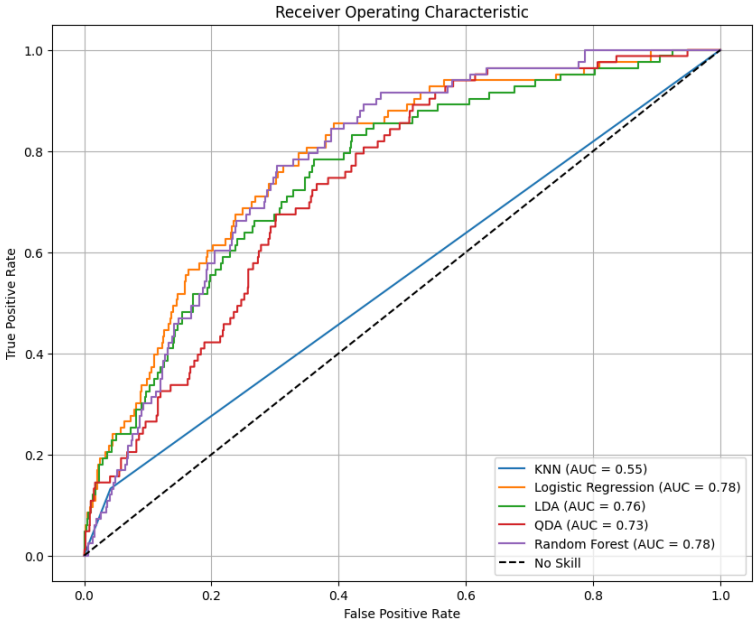
KNN(K=27)	0.75	0.96	0.12
Logistic Regression(α=1/0.001)	0.75	0.95	0.13
LDA	0.74	0.91	0.23
QDA	0.57	0.57	0.57
Random Forest(Trees=300, Max Depth=10)	0.74	0.82	0.49

Class-Balanced Model

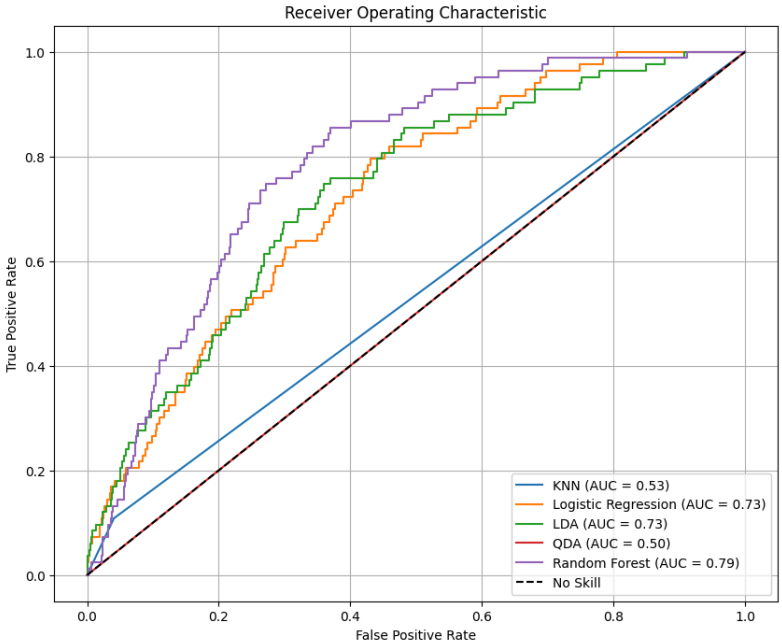
Model	**Accuracy**	**TPR**	**TNR**
KNN(K=1)	0.68	0.80	0.34
Logistic Regression(α=1/0.01)	0.72	0.78	0.53
LDA	0.68	0.72	0.58
QDA	0.71	0.88	0.21
Random Forest(Trees=900, Max Depth=10)	0.73	0.80	0.50

Appendix 2: The above tables show TPR and TNR for the adequate sleep classification problem for the baseline model, baseline model with interaction, class-balanced model without interaction terms respectively.

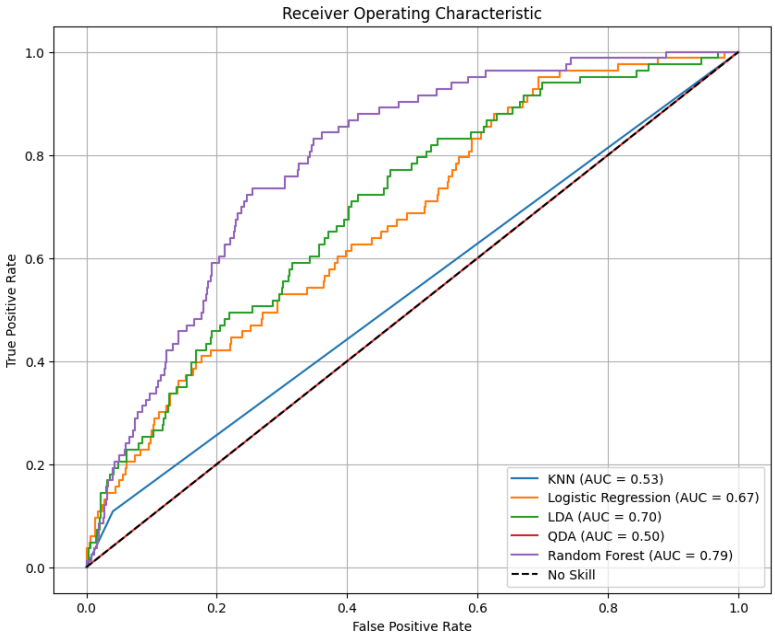
Caffeine Baseline



400mg with Interaction

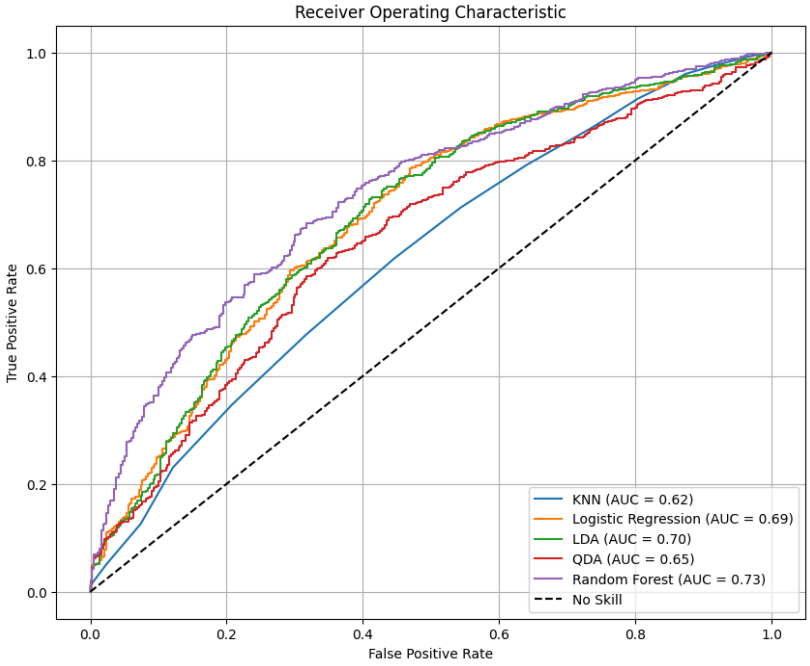


400mg Up-Sampling Interaction:

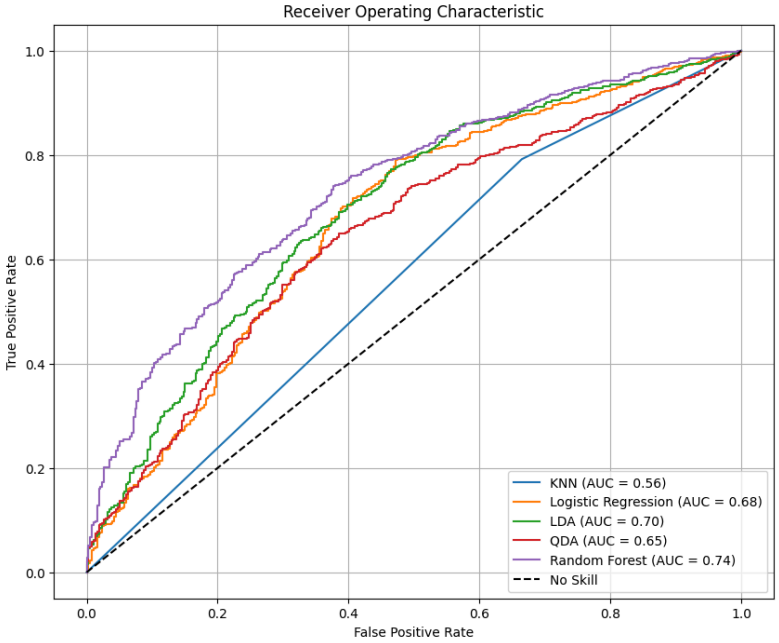


Appendix 3: The three ROC curves above display the model performance for Caffeine prediction with no up-sampling and no interaction, no up-sampling with interaction, and up-sampling with interaction respectively.

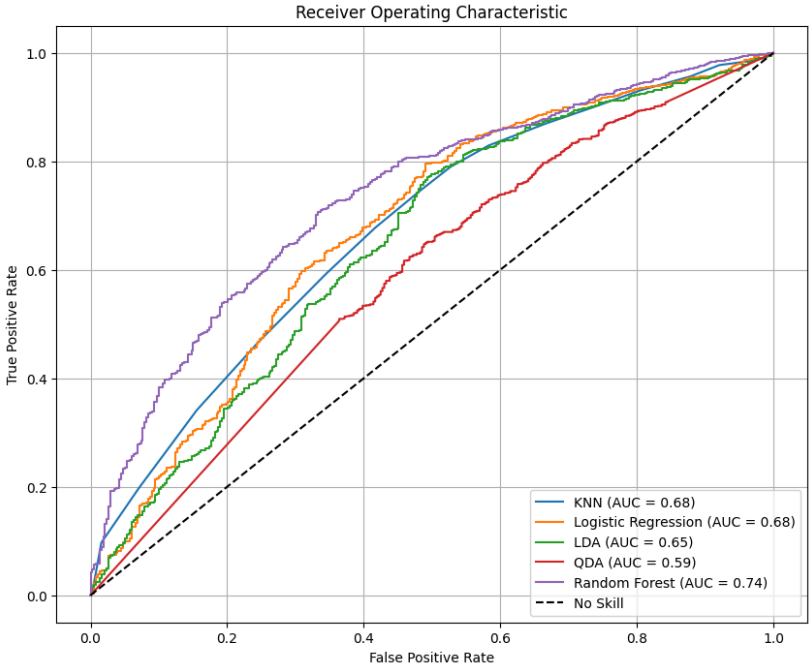
Sleep baseline:



Sleep Up-Sampling No Interaction



Sleep interaction:



Appendix 4: The three ROC curves above display the model performance for Sleep quality prediction with no up-sampling and no interaction, up-sampling no interaction, and no up-sampling with interaction respectively.