Analysis of Factors Related to Cardiac Arrhythmia

Melissa Vasquez, Olivia Apuzzio, Joanne Nie, Ramiro Romero, Taylor Froomin

Abstract

This analysis focuses on the relationship between clinical and demographic data of patients from the UCLA Sleep Lab and their risk for cardiac arrhythmia. Specifically, this was our research question of interest : is there correlation between a patient's limb movements and the presence of abnormal sinus rhythm? The text columns in the data were mined to extract the response variable for the research as well as for further feature engineering. After performing statistical testing, it was found that there was no significant relationship between periodic limb movements and cardiac arrhythmia. Other factors such as the Apnea-Hypopnea Index and the patient's age were shown to be more important. A random forest model to classify patients with higher risk for arrhythmia had weak predictive power.

Statement of the Problem

According to the National Institute of Health, around one in twenty people have a type of arrhythmia, which is defined as any abnormal rhythm in the heart. Since arrhythmia may or may not have symptoms and comes in many forms, it is difficult to diagnose and estimate the proportion of the general population with arrhythmia. We were given sleep data on 401 patients from the UCLA Santa Monica Pulmonary & Sleep Medicine Clinic, which included demographic as well as clinical data. We centered on our analysis on the questions:

- Does the frequency and duration of periodic limb movements of sleep correlate with the presence of sinus or cardiac arrhythmia?
- What is the general profile/distribution of patients in the lab?
- What are the most important factors associated with an individual at-risk of arrhythmia?

- Can we use these findings to predict which patients have arrhythmia?

Our goals were to both understand the patients in the Sleep Clinic and apply statistics and machine learning techniques to draw insights about cardiac arrhythmia. We examined a subset of the variables and created a few new ones:

Variable Name	**Type**	**Description**
Age	Numeric	Patient's age in years
Gender	Categorical	Male or female
BMI	Numeric	Patient's body mass index
ESS	Numeric	Epworth Sleepiness Scale
AHI	Numeric	Apnea-Hypopnea Index
AHI.REM	Numeric	Apnea-Hypopnea Index during REM
Apnea.Counts	Numeric	Number of apneas
Apnea.Counts.REM	Numeric	Number of apneas during REM
Latency.to.Sleep.Onset	Numeric	Time it takes for patient to fall asleep after turning lights out
Latency.to.REM	Numeric	Time it takes for a patient to reach REM after turning lights out
Desats.LT.90	Numeric	Number of times that oxygen saturation < 90
Desats.LT.80	Numeric	Number of times that oxygen saturation < 80
Desats.LT.70	Numeric	Number of times that oxygen saturation < 70
PLM.Total	Numeric	Total number of periodic limb movements
Sleep.Eff.Index	Numeric	Percentage of time spent sleeping in relation to time spent in bed
LEG1.Index	Numeric	Number of movements of leg 1
LEG2.Index	Numeric	Number of movements of leg 2
Sleep_notes	Categorical	One of 10 groups describing how the patient slept in the lab
Arrhythmia	Categorical	1 or 0 indicating the presence or absence of arrhythmia

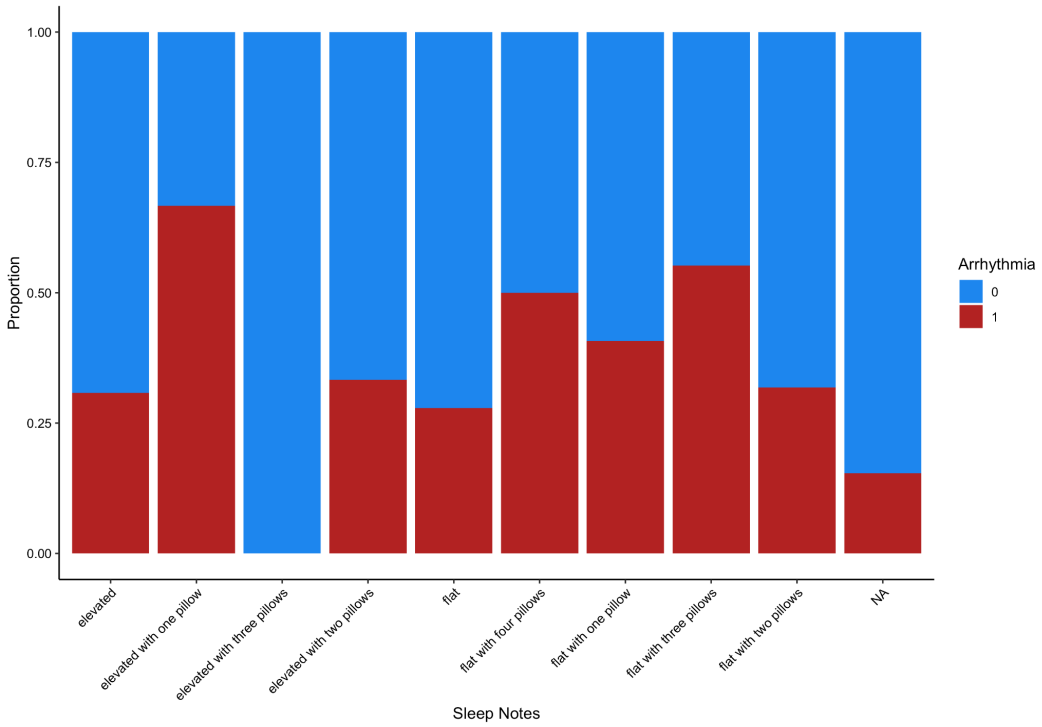
We also performed some data cleaning to prepare our data for analysis. First we standardized the height and weight columns so that everything was expressed in the same unit. We also removed 16 rows from the data that correspond to infants, as recommended by the client. The response variable 'Arrhythmia' was extracted from the original text column 'SleepStudy.EKG.Analysis.' Any patients with mentions of abnormal sinus rhythm, including PACs, PVCs, Tachycardia, and possible sinus arrhythmia were labeled as '1' in the 'Arrhythmia' column. Conversely, patients that had the notes 'NSR,' 'normal sinus rhythm,' or 'unremarkable' were denoted with 0 in the column. The breakdown of our response variable is detailed in the chart below.

Arrhythmia	Count
0	255
1	129

Additionally, we parsed the text column 'Sleep.Generic11.' for information about the way the patient slept in the lab. The column was originally very messy, but by using regular expressions, we were able to group the patients into ten categories which we will go into more detail in exploratory data analysis.

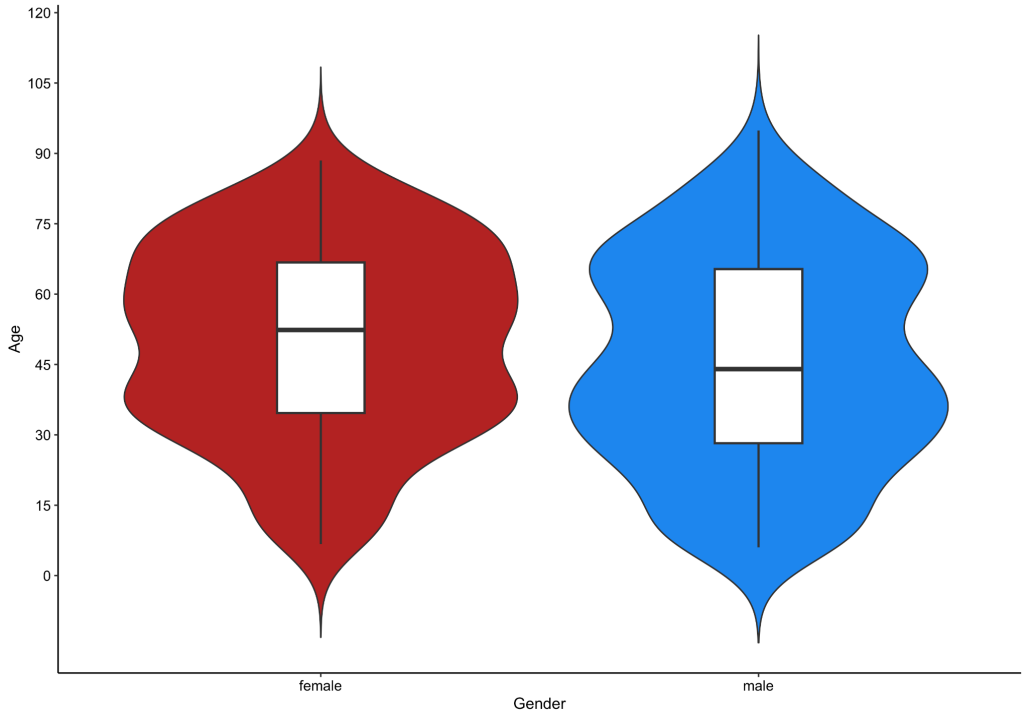
Exploratory Data Analysis

We first performed EDA on our data to extract some initial trends about our data and to understand the demographic distributions of the patients.

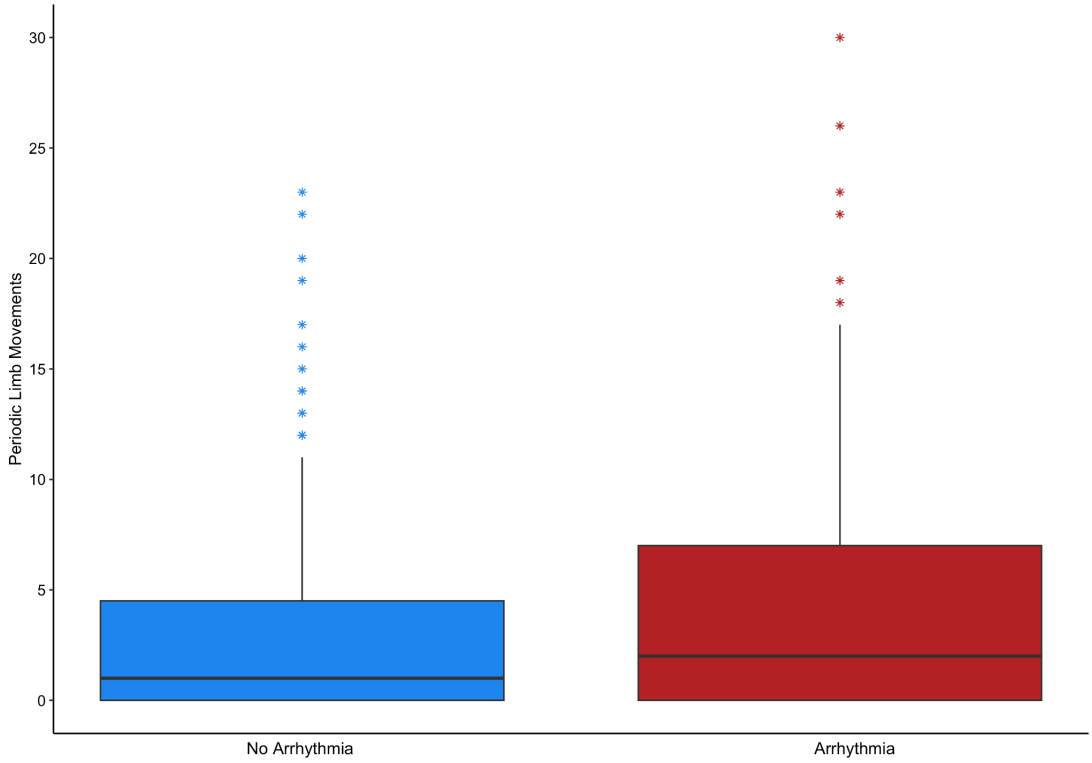


	Count
flat with two pillows	176
flat	79
flat with one pillow	54
flat with three pillows	29
elevated	13
NA's	13
elevated with two pillows	12
elevated with one pillow	3
elevated with three pillows	3
flat with four pillows	2

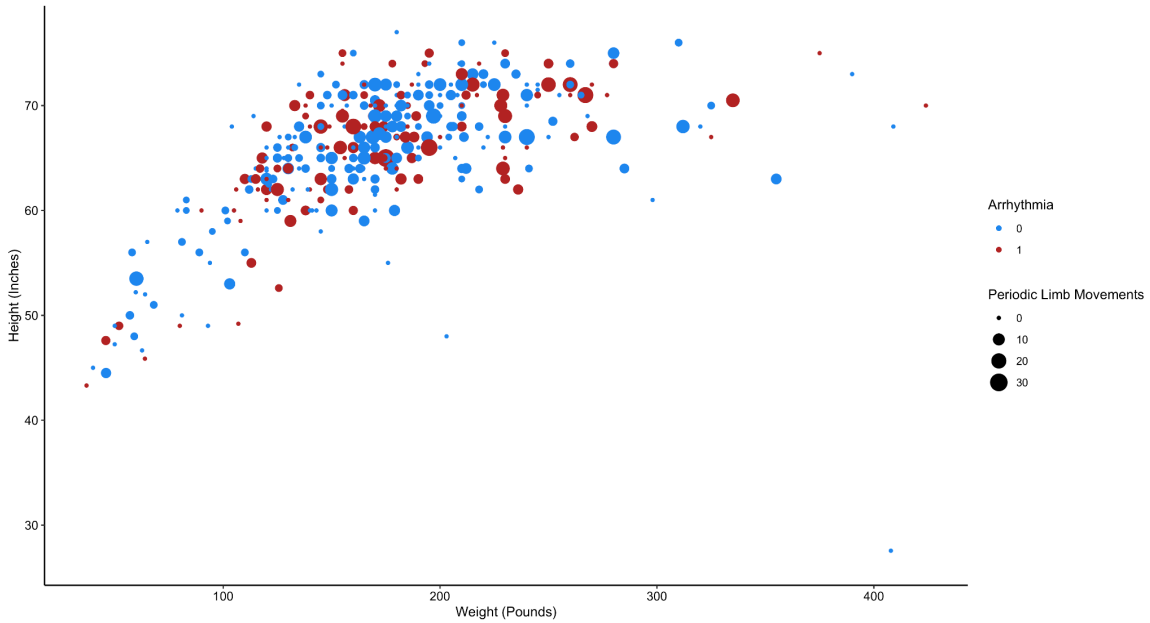
This proportional stacked barplot shows the proportion of patients with and without arrhythmia grouped by their sleeping position. The proportion of patients with arrhythmia is shown in red and the proportion of patients without arrhythmia is shown in blue. However, the number of patients in each category must be taken into consideration. As shown in the chart, some categories have vastly larger or fewer data points, so the proportions shown in the plot may be skewed due to small amounts of data.



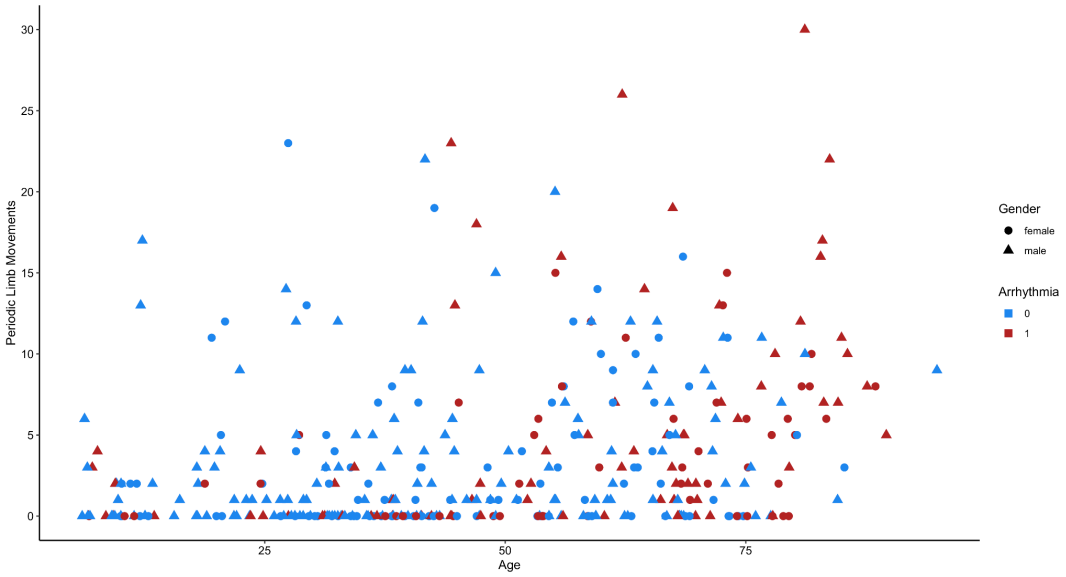
This violin plot shows the age distributions by gender of patients in the data. Data for female patients is shown in red and data for male patients is shown in blue. It shows that most female patients fall between ages 30-75 and most male patients fall between ages 30-45 and 60-75, with a decrease between ages 45-60. As noted by Dr. Aysola during the presentation, these age distributions align with the sleep center's expectations. Explanations for these distributions as given by Dr. Aysola are changes in life circumstances, such as marriage (leading to discovery of sleep apnea at a relatively young age) and retirement (leading to discovery of sleep apnea after realizing that tiredness and bad sleep are not caused by lifestyle).



These boxplots show the distribution of the number of periodic limb movements (PLMs) by arrhythmia status. The distribution of PLMs for patients without arrhythmia is shown in blue and the distribution of PLMs for patients with arrhythmia is shown in red. The range in the number of PLMs for patients with arrhythmia is larger than that of patients without arrhythmia, and the median number of PLMs for patients with arrhythmia is higher than the median number of PLMs for patients without arrhythmia.



This scatterplot shows the height and weight of patients with the color of the points representing arrhythmia status and the size representing how many PLMs they had. There seem to be more arrhythmias as weight increases, but this is not a hard and fast rule since there is not an especially clear relationship between arrhythmia and PLMs.

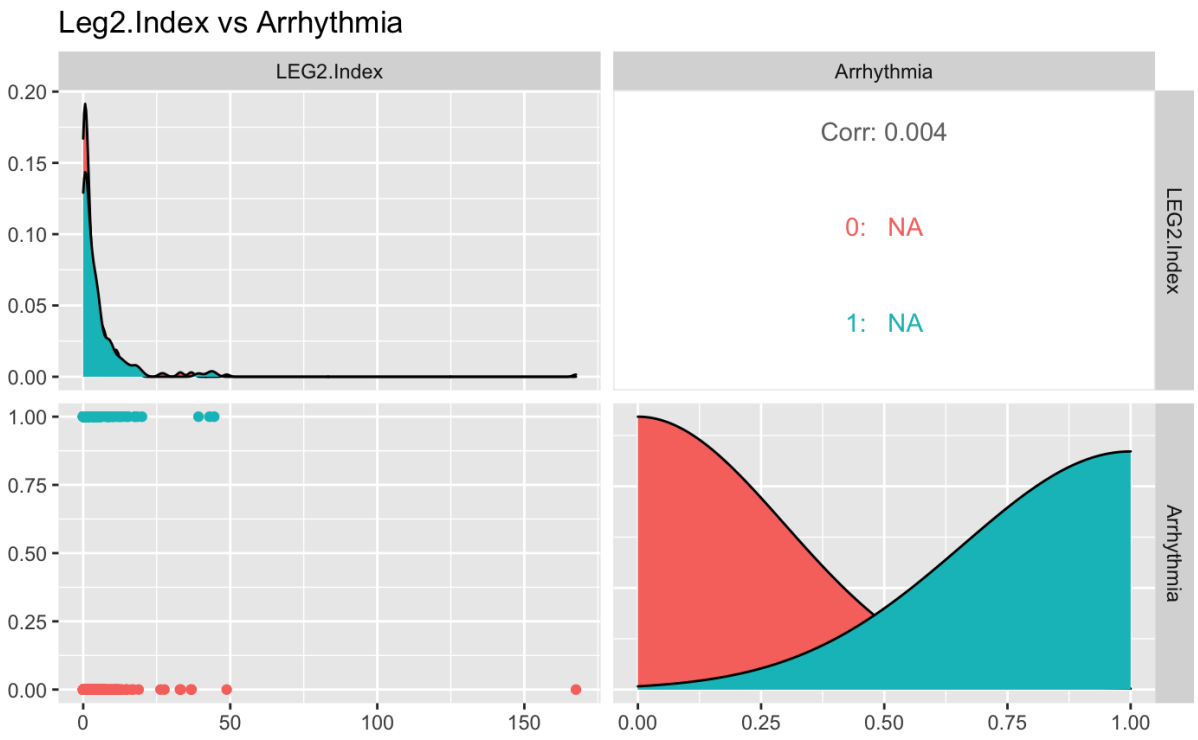


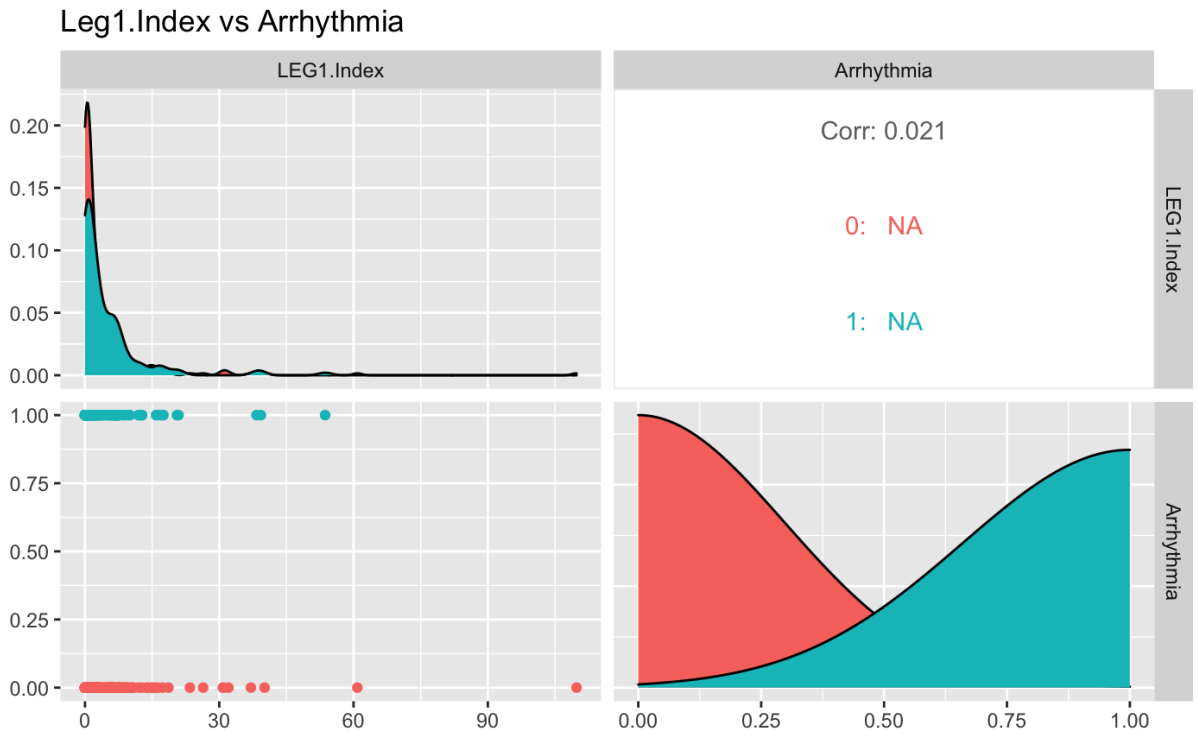
This scatterplot shows the number of PLMs and age with shape indicating gender and color indicating arrhythmia status. There seem to be more arrhythmias in older patients and there is a very minimal increase in PLMs as age increases, but it is important to keep in mind the fact that the vast majority of patients don't experience any PLMs.

Analysis

In our investigation regarding the potential correlation between the frequency and duration of periodic limb movements during sleep and the presence of sinus or cardiac arrhythmia, we employed a Pearson correlation test. The variables pertaining to body movements were LEG1.Index and LEG2.Index, both represented by numerical values. A higher numerical value indicated a greater frequency of leg movement. On the other hand, Arrhythmia was a binary variable that informed us whether the patient had or was suspected to have any type of arrhythmia. Upon computing the Pearson correlation coefficient between LEG1.Index and Arrhythmia, we obtained a correlation of .021, indicating the absence of a significant association between these two variables. Similarly, when calculating the Pearson correlation coefficient between LEG2.Index and Arrhythmia, we obtained a correlation of .004, which also signified the lack of a significant association between these two variables. The correlation test enabled us to conclude that there was no noteworthy correlation between the frequency of body movements and the presence of sinus or cardiac arrhythmia.

To further support our findings, we conducted independent t-tests using Pearson's correlation coefficient to examine if there were any differences in group means between individuals with arrhythmias and those without. For LEG1.Index, we obtained a t-value of 0.43473 and a corresponding p-value of 0.664. Similarly, for LEG2.Index, the t-value was 0.099043, and the p-value was 0.9212. Both p-values indicate that there were no significant differences observed between the two groups. Considering the results from both the correlation test and t-tests, we fail to reject the null hypothesis. Consequently, we recommend against allocating excessive time to further investigate the relationship between these two variables.





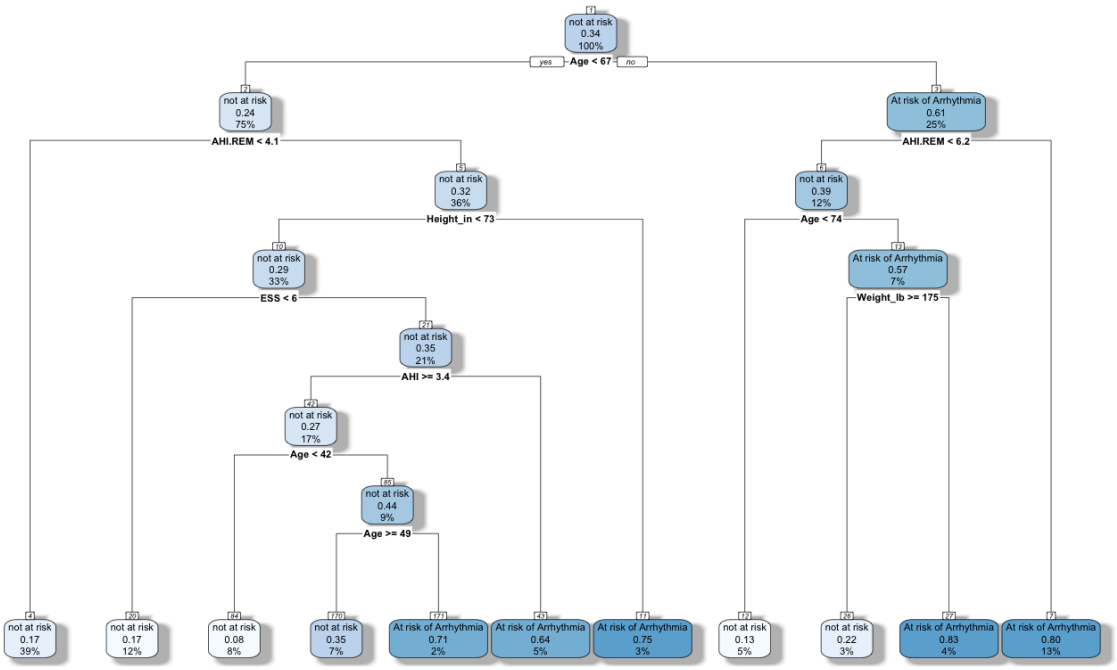
A correlation matrix was used to obtain an overview of possible significant relationships between other variables and examine the correlation between periodic limb movements and

presence of arrhythmia. A linear correlation matrix is essentially a table that displays the correlations between different variables. It shows how each random variable in our dataset is correlated with each of the other values in the table so we can see which pairs have the highest correlation at a glance. The correlation for periodic limb movements and arrhythmia is 0.14, which is not significant enough that we recommend investigating further. However, as mentioned in earlier analysis, there are other variables that display a relationship with arrhythmia, such as age, apnea hypopnea index, and AHI.REM (0.28, 0.15, and 0.15, respectively). Other strong correlations include but are not limited to: PLM.total with age, weight and height with age, weight with BMI, apnea counts and AHI, AHI.Rem and AHI, weight and AHI, apnea counts and AHI.Rem, height and weight with AHI.Rem, and LEG1 and LEG2 index with Sleep.Eff.Index. While some of these correlations are obvious, such as height and weight, the relationships between other variables such as LEG1 and LEG2 index with sleep efficiency may be worth investigating in future studies.

	Age	BMI	ESS	AHI	AHI.REM	Apnea.Counts	PLM.Total	Sleep.Eff.Index	LEG1.Index	LEG2.Index	Arrhythmia	Weight_lb	Height_in
Age	1.00	0.11	0.09	0.12	0.08	0.11	0.29	-0.25	0.15	0.15	0.28	0.23	0.25
BMI	0.11	1.00	-0.03	0.02	0.00	-0.00	0.01	-0.02	-0.00	-0.00	-0.00	0.52	-0.33
ESS	0.09	-0.03	1.00	0.02	0.03	-0.05	0.02	-0.03	0.01	0.01	0.08	0.09	0.10
AHI	0.12	0.02	0.02	1.00	0.58	0.70	0.04	-0.11	0.03	0.02	0.15	0.23	0.05
AHI.REM	0.08	0.00	0.03	0.58	1.00	0.35	-0.04	-0.09	0.02	-0.01	0.15	0.15	0.01
Apnea.Counts	0.11	-0.00	-0.05	0.70	0.35	1.00	0.05	-0.01	-0.03	-0.03	0.09	0.14	0.09
PLM.Total	0.29	0.01	0.02	0.04	-0.04	0.05	1.00	0.01	0.18	0.15	0.14	0.14	0.10
Sleep.Eff.Index	-0.25	-0.02	-0.03	-0.11	-0.09	-0.01	0.01	1.00	-0.30	-0.28	-0.05	-0.08	-0.07
LEG1.Index	0.15	-0.00	0.01	0.03	0.02	-0.03	0.18	-0.30	1.00	0.95	0.02	0.01	-0.02
LEG2.Index	0.15	-0.00	0.01	0.02	-0.01	-0.03	0.15	-0.28	0.95	1.00	0.00	-0.01	-0.01
Arrhythmia	0.28	-0.00	0.08	0.15	0.15	0.09	0.14	-0.05	0.02	0.00	1.00	0.02	-0.00
Weight_lb	0.23	0.52	0.09	0.23	0.15	0.14	0.14	-0.08	0.01	-0.01	0.02	1.00	0.53
Height_in	0.25	-0.33	0.10	0.05	0.01	0.09	0.10	-0.07	-0.02	-0.01	-0.00	0.53	1.00

Modeling

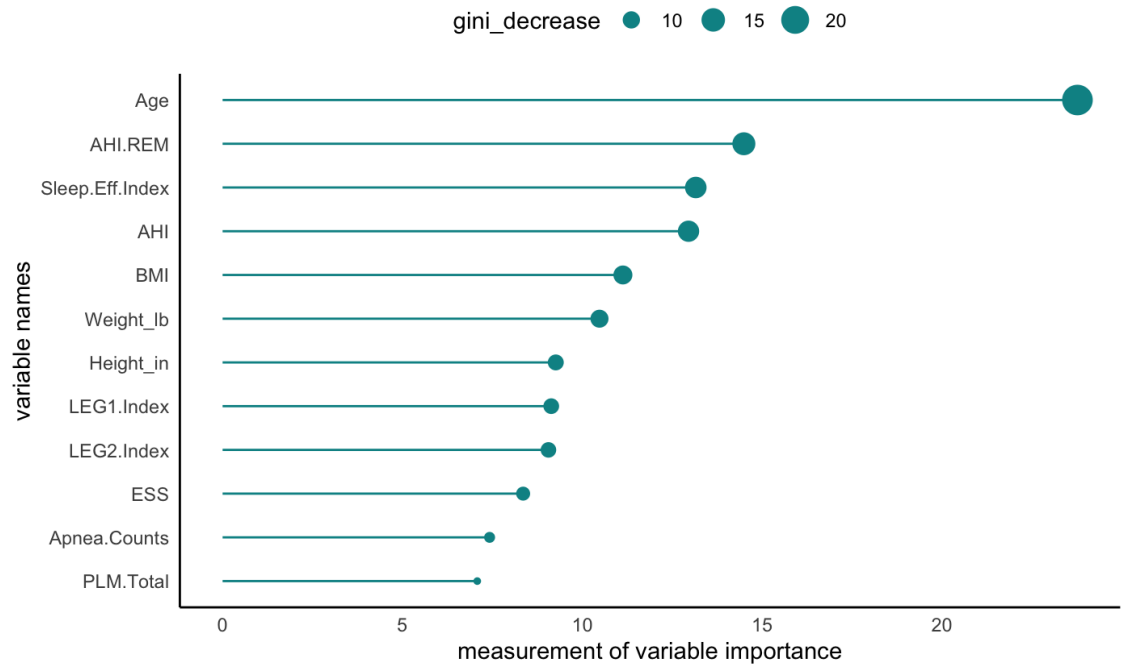
We decided to develop a random forest for 3 main reasons: the method is non parametric, the method performs exceptionally well when the data has missing values, and the method has clever manners of illustrating variable importance. Random forest is an ensemble algorithm that utilizes multiple decision trees to predict outcomes. It begins by creating bootstrap samples from the training data and constructing decision trees on each sample. At each node, a random subset of features is used for splitting. The final prediction is obtained by combining the predictions from all the trees in the forest. This technique addresses overfitting, handles high-dimensional data, and delivers reliable predictions by harnessing the collective knowledge of the tree ensemble.



The decision tree is a predictive model that partitions the dataset based on a series of decisions or criteria. In this case, the tree starts with 100% of all patients, with only 34% at risk

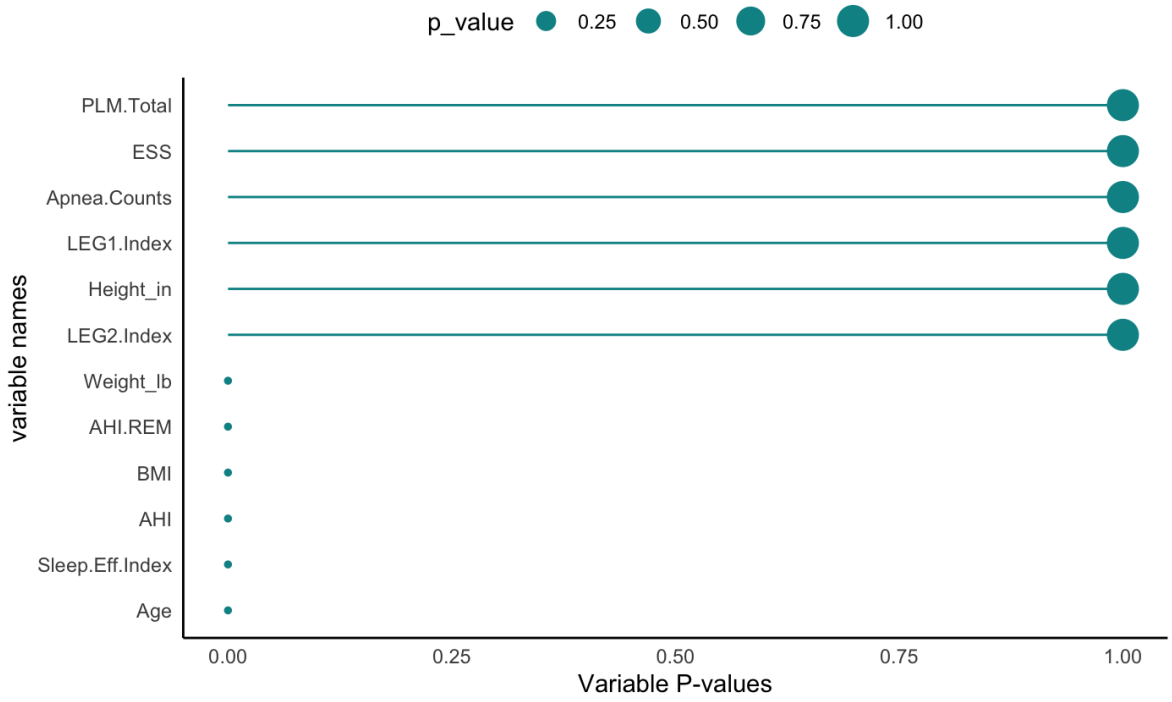
of arrhythmia. The most important variable is chosen, and a cutoff value is established to separate patients at risk from those not at risk. The tree then asks if the patient is under the age of 67, resulting in two subsets of patients. When a new patient is encountered, they follow a path along the decision tree based on their characteristics. The terminal nodes of the tree indicate the prediction outcome, with darker shades of blue representing higher confidence in the patient being at risk of arrhythmia and lighter shades indicating a higher confidence of no risk.

Random forest is an ensemble learning algorithm that utilizes multiple decision trees to make predictions and assess variable importance. In the context of arrhythmia outcomes, Random forest can provide measures that highlight the significance of different variables. By analyzing these measures, we can gain insights into the factors that play a crucial role in determining arrhythmia outcomes. Furthermore, it allows us to understand how the leg indices align with those important factors, providing additional context and understanding of their relevance in predicting arrhythmia.



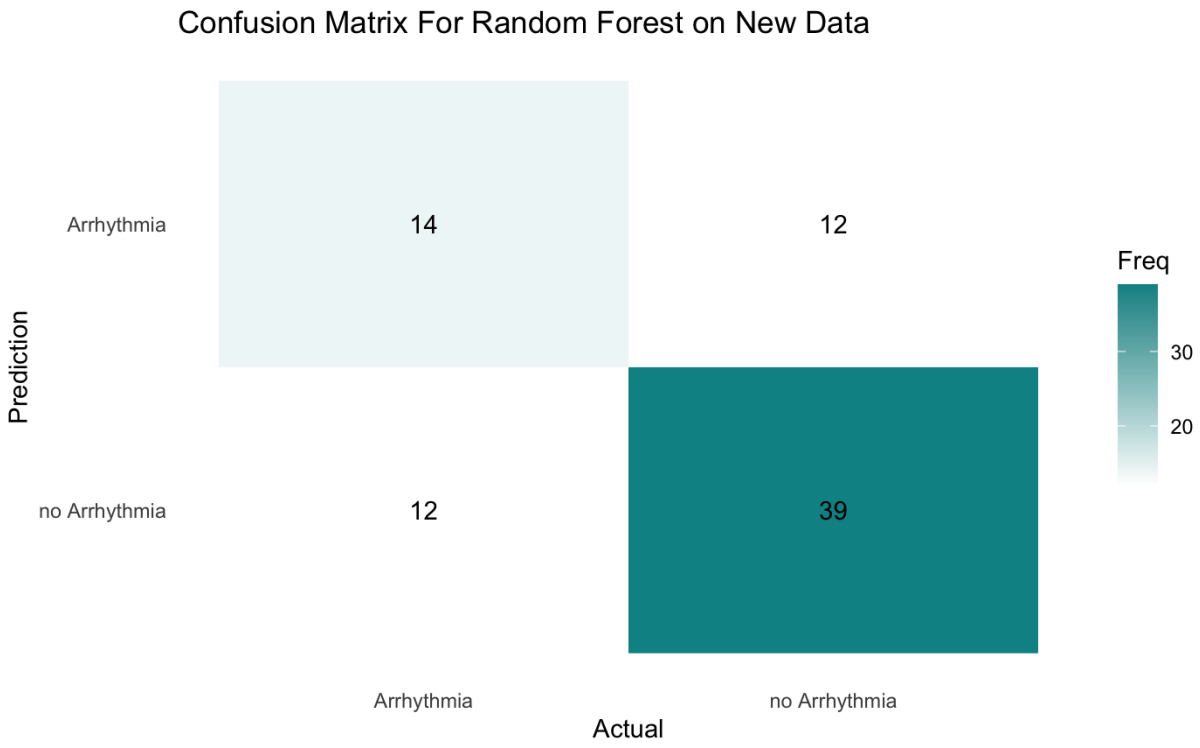
A Gini decrease variable importance plot is a visual representation of the importance of different variables in a random forest model based on the Gini impurity metric. The Gini impurity measures the degree of impurity or randomness in a node of a decision tree. In the plot, each variable is plotted on the y-axis, and the corresponding Gini decrease (or improvement) achieved by that variable is shown on the x-axis. A higher value on the x-axis indicates that the variable has a greater impact on reducing impurity and improving the accuracy of predictions. By examining the plot, we can identify the variables that contribute the most to the model's performance. Variables with higher Gini decrease values are considered more important in making accurate predictions. Conversely, variables with lower values have less influence on reducing impurity and may be less informative for the model. This plot helps us understand the relative importance of variables and their contribution to the overall model's performance.

Based on our findings, LEG1.Index and LEG2.Index are among the five least important predictors. This suggests that using these variables as criteria to split decision trees is not significantly beneficial or informative on average. In other words, these variables do not contribute as much to the predictive power of the model compared to other predictors.

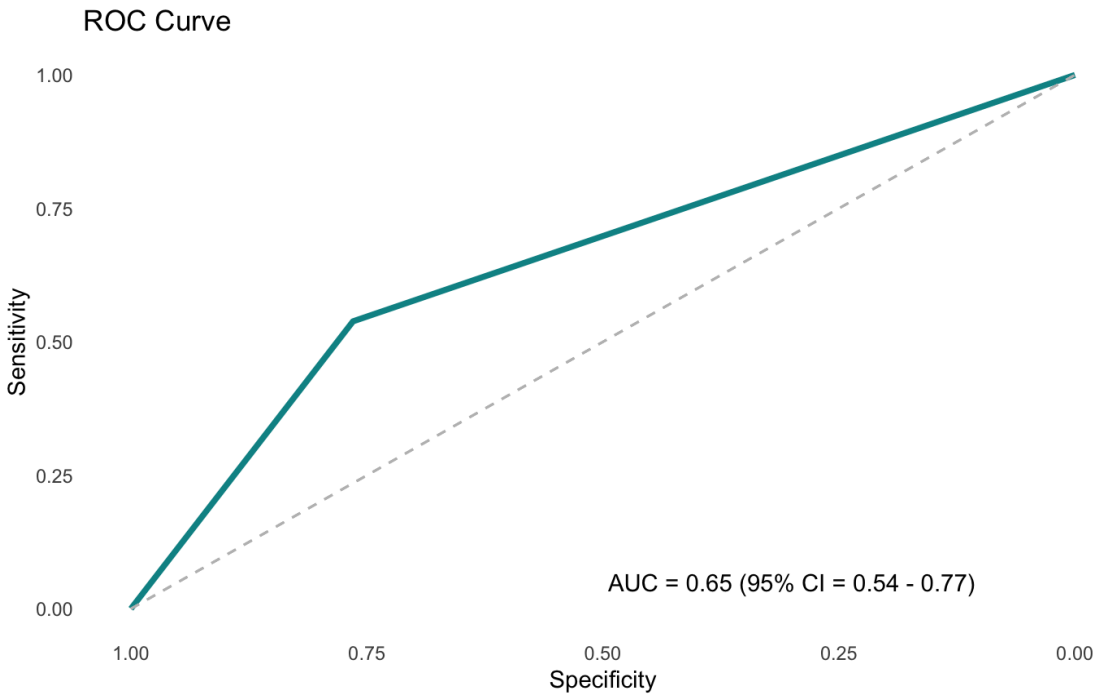


In the plot, each variable is plotted on the y-axis, and the corresponding p-value (a measure of statistical significance) is shown on the x-axis. The p-value indicates the probability of observing a relationship between the variable and the outcome by chance. A lower p-value suggests that the variable is more likely to have a significant impact on the outcome. Typically, a threshold (e.g., 0.05) is set to determine statistical significance. If the p-value for a variable falls below this threshold, it is considered statistically significant, indicating that the variable is important for the model's predictions. On the other hand, variables with higher p-values are considered less significant and may have less impact on the outcome. The p-value plot for variable importance provides a concise overview of the statistical significance of each variable, helping to identify which variables have a strong influence on the model's predictions and which ones may not be statistically significant.

In this case we found 6 variables to have a significant relationship with the response which have a low probability of being observed by chance alone: AHI, AHI.REM, WEIGHT, AGE, BMI, and Sleep Efficiency Index.



The confusion matrix provides an assessment of the performance of the Random Forest model. The model's performance is evaluated using the balanced accuracy metric, which measures the average accuracy across all classes. In this case, the balanced accuracy is reported as 65.16%, indicating the overall effectiveness of the model in making accurate predictions.



The ROC curve represents the performance of the Random Forest model in terms of sensitivity and specificity. Sensitivity: The model's sensitivity, also known as recall or true positive rate, indicates the proportion of actual positive cases (patients at risk of cardiac arrhythmia) that are correctly classified as being at risk. In this case, the sensitivity is reported as 53.9%, indicating that 53.9% of patients at risk are correctly identified by the model. Specificity: The model's specificity, or true negative rate, measures the proportion of actual negative cases (patients not at risk) that are correctly classified as not being at risk. The provided information states that the specificity is 76.5%, indicating that 76.5% of patients not at risk are correctly classified as not being at risk.

In summary, the ROC curve analysis of the random forest model shows that it demonstrates moderate sensitivity in correctly identifying patients at risk of cardiac arrhythmia (53.9%), while exhibiting a relatively higher specificity in correctly classifying patients not at risk (76.5%).

Results and Conclusions

Overall, we found no statistically significant relationship between the frequency and duration of periodic limb movements and the presence of cardiac arrhythmia. While there is a fairly weak positive correlation, it is not statistically significant enough to be investigated further. However, there are other factors associated with an individual at-risk of arrhythmia, such as age, BMI, AHI and AHI.Rem, and weight. To evaluate our model, it is fairly accurate at 65%, and confirmed earlier relationships we examined and found additional important variables such as sleep efficiency index.

Limitations and Challenges

We did run into some hurdles in our analysis. The first one was classifying our response variable through text mining. We decided on only two classes, but it was difficult to draw the line between the two since a lot of the patients had ‘possible sinus arrhythmia.’ An improvement in future work is to include an additional class specifically for the risk of arrhythmia. This addition could also enhance modeling capabilities by providing more granularity and capturing the specific risk level. Additionally, we were aware that our data was imbalanced, with more cases of normal sinus rhythm than arrhythmia cases. We opted not to balance the data since undersampling would require losing data while oversampling might lead to overfitting during the modeling process.

Next Steps

In the future, we would recommend adding both more clinical and demographic data specifically on cardiac arrhythmia to enrich the dataset. It may also be useful to establish more concrete methods of collecting and categorizing this data; while text columns provide a lot of

detail, creating bins and cutoffs from a medical perspective would be incredibly valuable for a statistician.

In terms of modeling, future work can be done to adjust the data before analysis by balancing it. Balancing the data can help prevent bias towards the majority class and improve the model's ability to predict all classes accurately. Another option is to augment the data, which involves generating synthetic samples or expanding the dataset to provide more diverse and representative examples. Data augmentation can help improve the model's generalization and robustness.

Additionally, there are multiple areas that future research groups could analyze. It would be interesting to test if certain sleep positionings are correlated to a patient’s AHI or ESS, since we found no such relationship to arrhythmia. Finally, there may also be some differences in measurements depending on who scored the patient. A future project could include blocked hypothesis testing to account for this possible source of variation.