Sleep Apnea Prediction Based on Reported Health Metrics

Calvin Knowles
calvin.knowles@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Abstract

This project aimed to evaluate and compare the performance of multiple machine learning models to determine the best predictor of sleep apnea based on reported health metrics. A comprehensive comparison was done over 9 models which were assessed through key performance metrics such as accuracy, precision, recall, F1 score, and cross validation. The results demonstrated that the Random Forrest and XGBoost models clearly outperformed the others. We then further compared these two models to determine an overall best model for our goal. By running a combined cross validation F1 score over 5 folds, we determined that XGBoost was a better model than Random Forrest in this scenario. XGBoost consistently outperformed the other models and when put to the test against Random Forrest it proved superior in predictive accuracy, precision balance, and consistency.

CCS Concepts

• **Mathematics of computing** → **Probability and statistics**; • **Applied computing** → **Health informatics**.

Keywords

Sleep Apnea, Machine Learning, Random Forrest, XGBoost

ACM Reference Format:
Calvin Knowles. 2024. Sleep Apnea Prediction Based on Reported Health Metrics. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Sleep apnea is a common sleep disorder that can significantly impact physical health, mental well-being, and overall quality of life. Despite the importance of treating sleep apnea, diagnostic methods, such as nocturnal polysomnography (doctor-monitored sleep study), are extensive and expensive. More recently, at-home polysomnography tests have been developed, but they can also become costly due to the volume of tests needed. With recent advances in computer science, data science, and machine learning, we have the tools to derive cost-efficient alternatives readily available to everyone. Our goal is not to replace polysomnography but to supplement it as a prerequisite before committing to expensive and extensive testing.

This project aims to predict the likelihood of sleep apnea using a dataset of health metrics, lifestyle habits, and current diagnosis. Some health predictors include, but are not limited to, the following metrics: blood pressure, daily steps, and overall sleep quality. By leveraging such predictors we can train a machine learning model to identify key patterns which drive a sleep apnea diagnosis. We plan to manipulate our dataset to quantify categorical predictors into a digestible scale to improve the accuracy of the machine learning model. This model will provide a more accessible and personalized tool for people to use before committing to further testing.

2 Related Work

There are various existing studies that aim to achieve the same goal as this project. One notable study is published to ScienceDirect.com: OSApredictor: A tool for prediction of moderate to severe obstructive sleep apnea-hypopnea using readily available patient characteristics (Talukder et al., 2024).

This study used patient predictor data such as age, sex, weight, height, pulse oxygen saturation, heart rate, and respiratory rate. The researchers then applied various machine learning models and chose the most accurate. Once the best machine learning model was chosen, they found their model accurately identified 72.5 percent of those with moderate to severe sleep apnea and 62.8 percent of those without moderate to severe sleep apnea. With an overall accuracy of 66 percent, I believe there is a significant amount of room for improvement in accuracy.

2.1 How does this project build upon previous work?

Based on this existing research, I can build my model off of the methods they outlined and improve upon it to output for more accuracy in diagnosing sleep apnea.

3 Proposed Work

3.1 Datasets and Tools

I will utilize a dataset from Kaggle that holds 5000 rows of records containing health metrics that can be utilized in our model. I will download this dataset to my computer and open it in a Jupyter Notebook running Python. With the Python packages OS, Kagglehub, and Pandas I can read the downloaded file into a data frame for cleansing and manipulation.

Once the dataset is clean, I will manipulate the categorical data using a Python function applied to the respective Pandas dataframe column. For example, there is a column of blood pressure that will be remapped to a scale from 1-8 through a Python function.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

This mapping will increase the usability within a machine learning model. Thus improving the accuracy of our predictions.

4 Tables

Category	Scale	Systolic (mm Hg)
Very Low	1	< 89
Low	2	90-99
Normal	3	100-129
Pre-Hypertension	4	130-139
Hypertension Stage 1	5	140-159
Hypertension Stage 2	6	160-179
Hypertension Stage 3	7	180-210
Hypertension Stage 4	8	> 210

Table 1: Blood Pressure Scale to Remap for Model Use

4.1 Main Tasks

The main tasks include cleaning the data, transforming the data, breaking the dataset into a training set and a testing set, and then evaluating the results to identify features to improve the model.

5 Evaluation

I will statistically evaluate the accuracy of the machine learning model by calculating the accuracy, precision, specificity, and balanced accuracy, and create various graphics of these metrics to find a conclusion.

5.1 Evaluation Metrics

Evaluation metrics will include accuracy, precision, recall, F1 score, and cross validation (1-5 folds). Based on these outcomes we will evaluate the results and modify the model to improve accuracy.

After initial rounds of machine learning modeling, the greatest evaluation metric has been the F1 score. The F1 score of a model is built from the precision and recall of a model.

Formulas for Precision, Recall, and F1 Score

The formulas for Precision, Recall, and F1 Score are as follows:

Precision

Precision is defined as the ratio of true positives (TP) to the total number of predicted positive observations (TP + FP):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where:

- TP: True Positives (correctly predicted as positive)
- FP: False Positives (incorrectly predicted as positive)

Recall

Recall (also called Sensitivity) is defined as the ratio of true positives (TP) to the actual number of positive observations (TP + FN):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where:

- TP: True Positives
- FN: False Negatives (actual positives missed by the model)

F1 Score

The F1 Score is the harmonic mean of Precision and Recall. It balances the two metrics into a single number:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.2 Experimental Setup

The experimental setup involved training and evaluating machine learning models to predict sleep apnea based on physiological data. Data pre-processing steps included scaling features, such as blood pressure, using standardization techniques, and splitting the dataset into a training subset and testing subset. We utilized several machine learning models to compare against one another to find the best model for our data. The models we used were Logistical Regression, Decision Tree, Random Forrest, Support Vector Machine, K-Nearest Neighbor, MLP Classifier, Naive Bayes, and XGBoost.

After comparing the various models, we determined Random Forrest and XGBoost to be the best performing. We then

5.3 Results

I ran 9 machine-learning models on the dataset: Logistic Regression, Decision Tree, Random Forrest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), MLP Classifier, Naïve Bayes, XGBoost, and an Ensemble of models. I ran each model 5 times and averaged their results to mitigate randomness.

Model	Accuracy	Precision	F1 Score
Logistic Regression	0.92	0.85	0.80
Decision Tree	0.93	0.87	0.84
Random Forest	0.91	0.85	0.77
Support Vector Machine	0.91	0.85	0.76
K-Nearest Neighbors	0.92	0.86	0.80
MLP Classifier	0.92	0.92	0.79
Naive Bayes	0.93	0.87	0.84
XGBoost	0.93	0.87	0.84

Table 2: Comparison of Various Machine Learning Approaches Diagnosis Ability for Sleep Apnea Based on Health Metrics

Based on these metrics, many of the models are performing on par with one another. We will need to extract more metrics from these models to find the best one.

I need to explore another evaluation method to supplement the averaged analysis. The next method I used was cross-validation folds. At a high level, cross validation breaks the dataset into various subsets (folds) to train and test a model. I ran cross-validation on each model over 1-5 folds. I then plotted the results into a box plot.

This boxplot (Figure 1) shows the distribution of model performance during cross validation. The higher the median of a box plot,

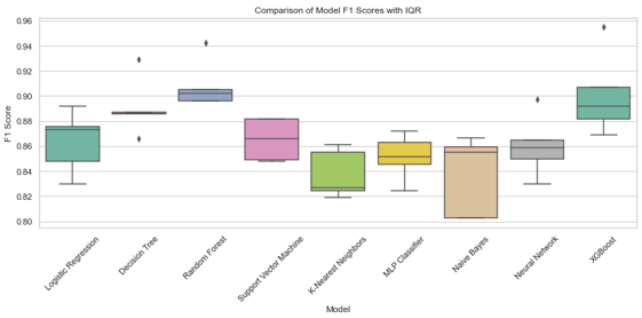

Figure 1: Box Plot of Model F1 score distribution over 1-5 folds

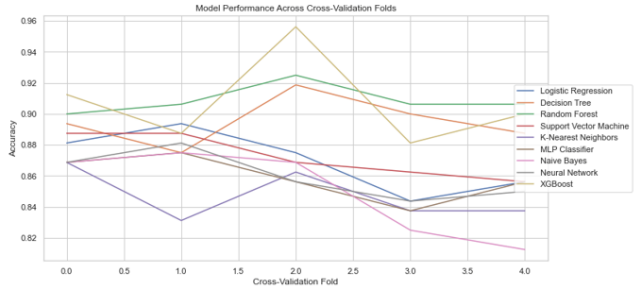

Figure 2: Model performance across validation folds to further determine the best model

the better that model performs. Therefore, we can break down the top two performing models are Random Forrest and XGBoost.

We can also plot a line of the cross validation performance for each model to further deduce the best two for further testing.

In this line graph (Figure 2) it is clear to see that for every cross-validation fold, except for one, the best performing model was either Random Forrest or XGBoost. Therefore, we will conduct one final test strictly between these two models to determine the best predictor for our problem.

Based on the previous results, Random Forrest, and XGBoost are the most accurate in identifying Sleep Apnea. This is not surprising because the XGBoost model is derived from Random Forrest. I decided to further test these two models before determining which one fit best. I compared the two model's by calculating their respective Cross Validation F1 score for 5 folds with a margin of error. Here is a breakdown of the two model's performances.

Model	CV F1 Score	Margin of Error
XGBoost	0.9126	± 0.0215
Random Forrest	0.9068	± 0.0137

Table 3: Comparison of Random Forrest and XGBoost for Cross Validation F1 Scores with Margin of Error

XGBoost has the edge with a CV F1 Score of 91.26% ± 2.15% compared to Random Forest's CV F1 Score of 90.68%±1.37%. We can conclude based on this measurement that XGBoost is the best machine learning model for this sleep apnea prediction based on reported health metrics.

6 Discussion

The goal of this project was to determine the best performing machine learning model for predicting sleep apnea based on health metrics. This project remained true to the expected timeline. One potential challenge for our results is that XGBoost is computationally expensive compared to the other models. Depending on the scenario, another model may be better suited to meet computational limits. A key lesson learned from this project is that machine learning problems are not a one size fits all. There are applications for each machine learning model and it is on us, the researcher, to determine the best model for our problem.

6.1 Timeline

The expected timeline for this project is 2 weeks broken down into finding data (1-2 days), cleaning and transforming the data (1 day), evaluating the results from the model (1 day), improving the model for accuracy (4 days), and creating presentations and reports (7 days). The timeline remained true for the entirety of the project.

6.2 Potential Challenges

Although our model provides a relatively high accuracy, there are potential challenges. One challenge is the size of the dataset we have access to. It is difficult to find and get access to patient data which has limited us to the amount of records we can use in our model. A way to address this challenge would be to pay for access to larger datasets or conduct a research survey of self-reported data (though self-reported data is another challenge in and of itself).

7 Conclusion

In this project, the machine learning models were rigorously tested and compared amongst each other based on various performance metrics such as: accuracy, precision, recall, F1 score, and cross-validation (1-5 folds). The comparison highlighted the power of XGBoost when it comes to a classification task. The key findings of our project were: XGBoost proved to be the strongest model for our problem, Random Forrest was a close second and may be better suited of these two models were better under these conditions. under computational limitations, and Cross Validation analysis was crucial for determining which Based on these finding, XGBoost is our recommendation for a real-world implementation of providing the public a quick and efficient way to determine sleep apnea. Although this method does not replace a true diagnosis, it can serve as an easy step for patients to test their health metrics before seeking medical treatment that may be costly in both time and money

References

Talukder, A. et al. (2024) OSApredictor: A tool for prediction of moderate to severe obstructive sleep apnea-hypopnea using readily available patient characteristics, Computers in Biology and

Medicine. Available at: https://www.sciencedirect.com/science/article/pii/S001048252400862X (Accessed: 09 December 2024).