Predicting the Risk Of Sleep Disorder Using Machine Learning

Vinayak Arora
vinayak21112@iiitd.ac.in

Sridhar Sodhi
sridhar21104@iiitd.ac.in

Saksham Pandey
saksham21486@iiitd.ac.in

Vinayak Goel
vinayak21113@iiitd.ac.in

Abstract

The aim of this study is to utilize the power of Machine Learning to create a prediction system, which can utilise the data from a patient's sleep record, and predict whether the patient is suffering from a sleep disorder or not.We utilized different Classification Algorithms to create a model which can predict sleep disorders .By leveraging data-driven insights, this project showcases the potential of machine learning to address critical healthcare challenges and improve quality of life on a larger scale. The code can be found at `https://github.com/VinayakG311/ML_Pro`

1. Introduction

Sleep disorders are conditions that affect the quality, amount and timing of sleep .Disorders such as insomnia and sleep apnea impact millions, often remaining undiagnosed and leading to serious health risks like heart disease and cognitive decline. However, with the advent of AI and advancement in the fields of Machine Learning, the field of medicine has evolved rapidly.One such example is training Machine Learning models on data of past patients,then making future predictions. These predictions, although not completely accurate, provide insights into the different symptoms, and sometimes succeed in providing an early diagnosis. Different studies have appeared in support of utilising the power of Machine Learning in detecting these disorders. Some of these studies talk about algorithms like KNN, Random Forest, Naive Bayes,SVM. We aim to create a prediction model by comparing different Machine Algorithms against one another, and enhance their efficiency by fine tuning the models.

2. Literature Survey

Unveiling Sleep Disorders presents a research paper discussing the different repercussions of sleep based disorders. It utilises RandomSearchCV to find the hyperparameters, and then goes in depth regarding the different ML Algorithms such as Decision Tree, Naive Bayes,Random Forest, KNN which have been employed , and can be utilised using class label prediction for predicting whether an individual is susceptible to insomnia or sleep apnea. They utilised techniques such as Precision, Recall and confusion matrix to properly compare the Machine Learning algorithms. RandomForest was found to be the one with highest accuracy, and believe that ANN can be utilised in future for increased accuracy.

Sleep Health Prediction and Algorithms Based on Big Data the study uses Logistic Regression and Random Forest to predict sleep health, It focusses on Logistic Regression ,and how it has a few disadvantages as it only works with discrete numbers and cannot find out complex relationships. The paper then creates a Machine Learning Prediction system piting these two algorithms against one another to see what creates a better prediction system.The study also goes in depth for the feature importance criteria, how it can be used to study and analyse features which were more relevant to the Random Forest.

Potential of Machine Learning for Predicting Sleep Disorders goes through twenty three Regression Models and twenty nine Classification Models to conduct the research. It also finds the best learning strategy by choosing amongst six learning strategies for higher order datasets. First they completed the pre-processing by filtering out the relevant features, and used correlation matrix for feature engineering. The study goes in depth about the mathematical formulas for each regression and then compares them on attributes such as MAE, MSE, RMSE and accuracy. They then perform the same steps for classification algorithms such as Random Forest, Decision Tree etc, and find out precision, recall and F1 scores for accurate measurements. The algorithms are then compared and the best one for predicting the sleep disorders is then find out that RandomForest was one of the most accurate models.

3. Dataset

3.1. Dataset Overview

The dataset consists of 3000 rows and 12 columns.The features are as follows :-

1. Gender - Gender of the person
2. Age - Age of the person
3. Occupation - Occupation of the person
4. Sleep Duration - Average hours of sleep per day
5. Quality of Sleep - Rating of the quality of sleep, 1 being the lowest and 10 being the highest
6. Physical Activity Level - Average minutes of physical activity per day
7. Stress Level - Rating of stress levels, 1 being the lowest and 10 being the highest
8. BMI Category - BMI classification i.e. Overweight / Normal / Obese
9. Blood Pressure - Blood pressure of the person measured as systolic/diastolic values
10. Heart Rate - Heart rate of the person in bpm
11. Daily Steps - Average number of steps per day
12. Sleep Disorder - Which disorder is the person suffering from. "None" in case of no disorder

3.2. Exploratory Data Analysis

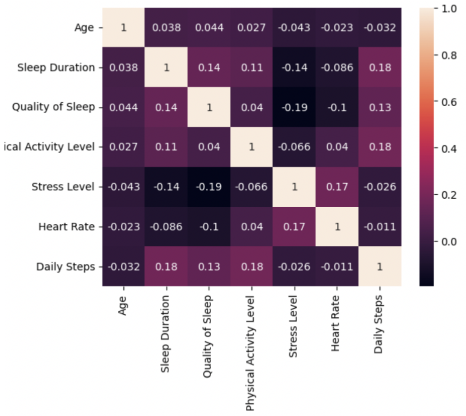
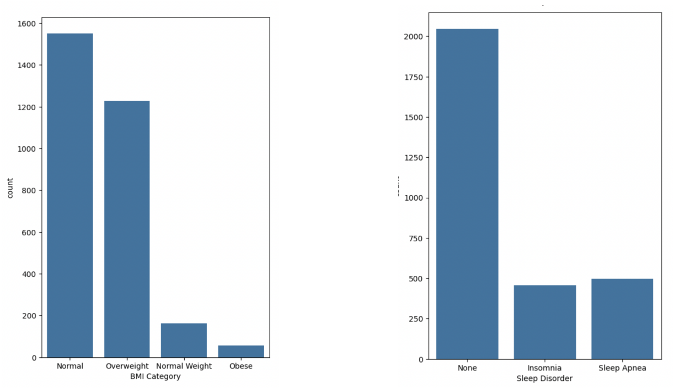


Figure 1. Correlation Matrix

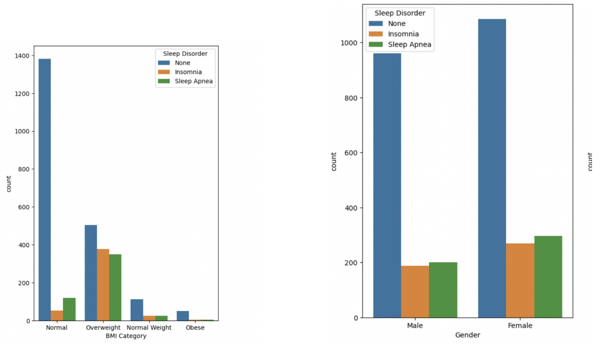
From Figure 1, there is a strong positive correlation between (Daily Steps - Sleep Duration), (Daily Steps - Physical Activity levels), (Heart Rate - Stress Levels), (Quality of Sleep - Sleep Duration), there is a strong negative correlation between (Stress Level - Quality of Sleep), (Stress Level - Sleep Duration).



(a) BMI Category (b) Sleep Disorder

Figure 2. Count Plot

Figure 2a shows the presence of class imbalance in the BMI Category column. There are very few samples with BMI Category label of Obese and Normal Weight.Figure 2b shows the presence of class imbalance in the Sleep Disorder column. There are very few samples with Sleep Disorder label of Insomia and Sleep Apnea.



(a) BMI Category (b) Gender

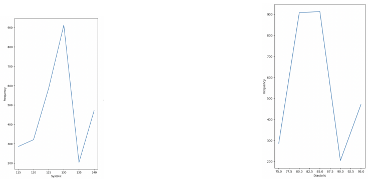
Figure 3. Relation between target and categorical featues

Figure 3a shows the direct positive relation between a person being overweight and the presence of a sleep disorder. If the person is normal, then high proportion of samples show no sleep disorder. Figure 3b shows the that sleep disorders and no disorder occur in equal proprtion in both male and female. This shows that gender is largely non-detrimental to the presence of sleep disorder.

3.3. Data Preprocessing and Feature Engineering

The dataset had no duplicate rows, but had 166 rows with Occupation being NaN and 218 rows with

Blood Pressure being NaN. The Occupation column was encoded using One-Hot Encoding to generate columns ’Occupation_Accountant’, ’Occupation_Doctor’, ’Occupation_Engineer’, ’Occupation_Lawyer’, ’Occupation_Nurse’, ’Occupation_Salesperson’, ’Occupation_Teacher’. The columns have boolean values indicating if the person has the specified profession as its occupation or not.The rows with NaN values for occupation were considered having none of the above professions and each of the columns were set to false. The Blood Pressure is represented as systolic / diastolic pressure. The column was split into 2 columns - ’Systolic’ and ’Diastolic’ and the Blood Pressure column was dropped. The distribution of ’Systolic’ and ’Diastolic’ are as follows :-

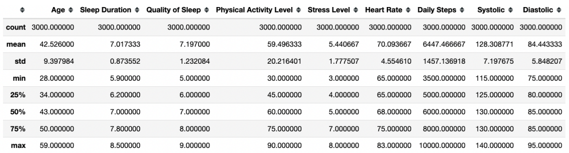


(a) Systolic

(b) Diastolic

Figure 4. Blood Pressure Distribution

From Figure 4 it is evident that Systolic has a more even distribution, so the NaN values were filled by meand and Diabloic has a more skewed distribution and hence the NaN values were filled by median.The description of the dataset is as follows :-



♦	Age ♦	Sleep Duration ♦	Quality of Sleep ♦	Physical Activity Level ♦	Stress Level ♦	Heart Rate ♦	Daily Steps ♦	Systolic ♦	Diastolic ♦
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
mean	42.526000	7.017333	7.197000	59.496333	5.440667	70.093667	6447.466667	128.308771	84.443333
std	9.397984	0.873552	1.232084	20.216401	1.777507	4.554610	1457.136918	7.197675	5.848207
min	28.000000	5.900000	5.000000	30.000000	3.000000	65.000000	3500.000000	115.000000	75.000000
25%	34.000000	6.200000	6.000000	45.000000	4.000000	65.000000	5000.000000	125.000000	80.000000
50%	43.000000	7.000000	7.000000	60.000000	5.000000	68.000000	6000.000000	130.000000	85.000000
75%	50.000000	7.800000	8.000000	75.000000	7.000000	75.000000	8000.000000	130.000000	85.000000
max	59.000000	8.500000	9.000000	90.000000	8.000000	83.000000	10000.000000	140.000000	95.000000

Figure 5. Description of Dataset

From the statistics, it is evident that there are no outlier in any of the columns hence standard scaling with the help of mean and variance would suffice. Thus the numeric columns were normalized using StandardScaler. The categorical columns (’Gender’,’BMI Category’) were transformed using Label Encoding.

4. Methodology and Model Details

4.1. Class Imbalance

The dataset was split into training : testing set in a 80:20 ratio. The training set contained a high proportion of No sleep disorder rows due to class imbalance in the original dataset.To overcome this, SMOTE (Synthetic Minority Over-sampling Technique) was used. SMOTE works by randomly selecting rows with minority class, selecting k of the nearest neighbours of the selected row, constructing a feature space between the 2 examples and generating synthetic data from the derived feature space. The training set before oversampling had the following distribution - ’None’: 1630, ’Sleep Apnea’: 404, ’Insomnia’: 366. After oversampling, the distribution was as follows - ’None’: 1630, ’Sleep Apnea’: 1630, ’Insomnia’: 1630.

4.2. Model Details

Support Vector Machine (SVM): SVM is a supervised learning algorithm that classifies data by finding the optimal hyperplane to separate classes. It works well in high-dimensional spaces and handles non-linear data using kernel functions. SVM achieved the highest accuracy of 94.5, highlighting its robustness and strong classification performance.

Random Forest: Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) from individual trees. It helps to mitigate overfitting, a common issue with decision trees, and improves generalization. The model achieved an accuracy of 89.26 and an AUC-ROC score of 79.87, indicating strong classification performance and the ability to differentiate between classes effectively. Additionally, the version fine-tuned with GridSearchCV achieved a similar accuracy of 89.20, showing that the base model was already well-optimized.

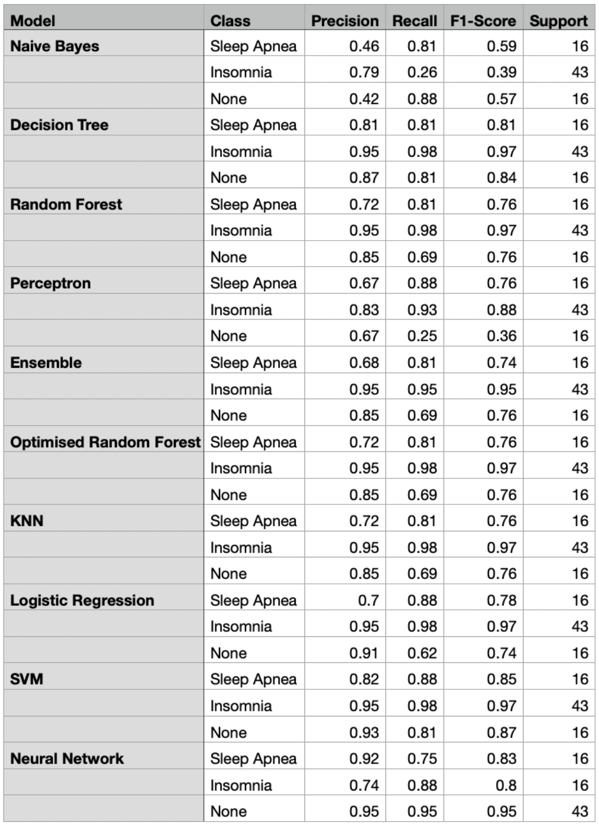
K Nearest Neighbors (KNN): KNN is a simple, instance-based learning algorithm that classifies data points based on the majority class of their nearest neighbors. It does not assume any underlying distribution for the data, making it versatile across different domains. KNN reached an accuracy of 87.44 with an AUC-ROC score of 73.86. Although KNN performed slightly lower than Random Forest, it remains a strong competitor, particularly when computational simplicity is a priority.

Other models explored include Naive Bayes, Decision Tree, Perceptron,neural networks and Logistic Regression. These models demonstrated varying degrees of performance but did not outperform the top three in terms of accuracy.

4.3. Principal Component Analysis

PCA was employed, first, the optimal number of components explaining the variance amont features was found out to be 9. The data was fit and transformed using PCA and the transformed data was also used in model training. Accuracy after using PCA was similar as before.

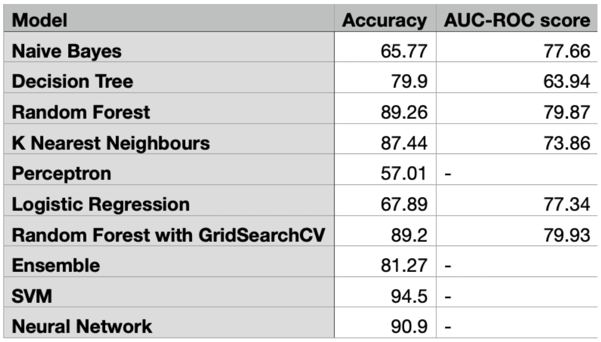
5. Results and Analysis



Model	Class	Precision	Recall	F1-Score	Support
Naive Bayes	Sleep Apnea	0.46	0.81	0.59	16
	Insomnia	0.79	0.26	0.39	43
	None	0.42	0.88	0.57	16
Decision Tree	Sleep Apnea	0.81	0.81	0.81	16
	Insomnia	0.95	0.98	0.97	43
	None	0.87	0.81	0.84	16
Random Forest	Sleep Apnea	0.72	0.81	0.76	16
	Insomnia	0.95	0.98	0.97	43
	None	0.85	0.69	0.76	16
Perceptron	Sleep Apnea	0.67	0.88	0.76	16
	Insomnia	0.83	0.93	0.88	43
	None	0.67	0.25	0.36	16
Ensemble	Sleep Apnea	0.68	0.81	0.74	16
	Insomnia	0.95	0.95	0.95	43
	None	0.85	0.69	0.76	16
Optimised Random Forest	Sleep Apnea	0.72	0.81	0.76	16
	Insomnia	0.95	0.98	0.97	43
	None	0.85	0.69	0.76	16
KNN	Sleep Apnea	0.72	0.81	0.76	16
	Insomnia	0.95	0.98	0.97	43
	None	0.85	0.69	0.76	16
Logistic Regression	Sleep Apnea	0.7	0.88	0.78	16
	Insomnia	0.95	0.98	0.97	43
	None	0.91	0.62	0.74	16
SVM	Sleep Apnea	0.82	0.88	0.85	16
	Insomnia	0.95	0.98	0.97	43
	None	0.93	0.81	0.87	16
Neural Network	Sleep Apnea	0.92	0.75	0.83	16
	Insomnia	0.74	0.88	0.8	16
	None	0.95	0.95	0.95	43

Figure 6. Precision, Recall, F1 score, Support

As the table shows, Random Forest and SVM best balance the precision and recall along the different classes. Naive Bayes and Logistic Regression perform well for 'Insomnia' but struggle for the other classes which are less prevelant, reducing the overall accuracy and F1 score. Ensemble models were good and consistent throughout, but faced the same issue of classifying the less prevalent classes properly.



Model	Accuracy	AUC-ROC score
Naive Bayes	65.77	77.66
Decision Tree	79.9	63.94
Random Forest	89.26	79.87
K Nearest Neighbours	87.44	73.86
Perceptron	57.01	-
Logistic Regression	67.89	77.34
Random Forest with GridSearchCV	89.2	79.93
Ensemble	81.27	-
SVM	94.5	-
Neural Network	90.9	-

Figure 7. Model Accuracy and AUC-ROC score

For this dataset, SVM is the best-performing model, offering the highest accuracy and the ability to differentiate between classes effectively, as reflected precision and F1-score. Random forest and Neural Network are other strong contenders, followed by the KNN, which might offer a good trade-off between performance and interpretability. Naive Bayes and Logistic Regression are moderately effective, while the Perceptron model is not suitable for this task.

6. Conclusion

Key learnings from the project include: Data Preparation and Feature Engineering: Handling missing values, balancing datasets with SMOTE, and normalizing features. Model Comparison: SVM, Neural Network and Random Forest outperformed other models like Naive Bayes, Decision Tree, and Perceptron in terms of both accuracy and overall performance metrics (e.g., precision, recall). The project highlighted the strength of a couple of different types of models, Vector Machines like SVM, Neural Network based models, and ensemble models, particularly Random Forest, in handling diverse features and complex decision boundaries. KNN and Neural Network, although simple, showed promising results but was computationally more expensive. In terms of Accuracy, SVM was the standout with the highest accuracy in predicting Sleep Disorders Evaluation Metrics: The importance of evaluating models using multiple metrics, such as accuracy, AUC-ROC, precision, and recall, became evident. Focusing solely on accuracy would have provided an incomplete picture of the model's true performance, especially given the class imbalance in the dataset. Challenges Faced: Class imbalance, particularly in underrepresented categories, posed a challenge and required attention during preprocessing.

6.1. Individual Tasks

Milestone	Members
1	Literature Review - All members
2	Dataset Exploration and visualization - All members
3	Data Preprocessing and Feature Engineering - Vinayak Arora and Vinayak Goel
4	Model training - All members
5	Model evaluation - Sridhar and Saksham
6	Hyperparameter Tuning - Sridhar and Saksham
7	Api and Frontend - Vinayak Goel and Vinayak Arora