MACHINE LEARNING MODEL FOR PREDICTING INSOMNIA LEVELS AMONG UNIVERSITY STUDENTS

BY

JESSICA HASNEEN
ID: 201-15-3156

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

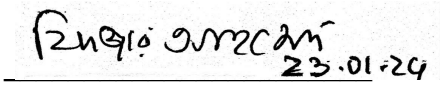
APPROVAL

This Project titled "**Machine Learning Model for Predicting Insomnia Levels Among University Students**", submitted by **Jessica Hasneen**, **Student ID: 201-15-3156** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 23 January, 2024.

BOARD OF EXAMINERS

Chairman

Narayan Ranjan Chakraborty (NRC)
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

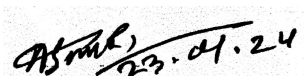


Internal Examiner

Dr. Fizar Ahmed (FZA)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Sharmin Akter (SNA)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR)
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that this project has been done by me under the supervision of **Ms. Subhenur Latif, Assistant Professor, Department of CSE, Daffodil International University.** I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ms. Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Jessica Hasneen
ID: 201-15-3156
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to **Ms. Subhenur Latif, Assistant Professor, Department of CSE Daffodil International University, Dhaka.** I'd like to thank my supervisor's deep knowledge and strong interest in Machine Learning for helping me to complete this project. This project could not have been done without his endless patience, expert advice, constant encouragement, constant and energetic supervision, constructive criticism, good advice, reading many bad draughts, and fixing them at every stage.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor and Head, of the Department of CSE, and Dr. S. M. Aminul Haque, Professor and Associate Head, of the Department of CSE** for their kind help to finish my project and also, to all of the faculty members and the staff of the CSE department of Daffodil International University for kind support and help on the technical and the administrative aspect of the study.

I would like to extend my gratitude to all of my classmates from Daffodil International University who participated in this conversation while they were working on their course
assignments.

Finally, I would like to express my gratitude to my parents for the everlasting love and patience. They have shown me throughout my life. They have been there for me throughout my whole life to support me and encage me in all I do.

ABSTRACT

Insomnia is a common and concerning issue among university students that can have a significant impact on their academic performance, physical and mental health. Early detection and treatment of insomnia is important to reduce its harmful effects. This study uses machine learning techniques to provide a novel approach of predicting the severity of insomnia among university students. Our research uses a dataset that was collected from a wide range of university students and includes demographic information, lifestyle variables and mental health indicators. In order to develop predictive models for insomnia levels, we utilize some of the popular machine learning algorithms, such as K-Nearest Neighbors, Support Vector Machine (SVM), Naïve Bayes, Linear Discriminant Analysis, Stochastic Gradient Descent, Extra Trees Classifier, AdaBoost Classifier, Ridge Classifier. All of the classifiers predict with high accuracy and Support Vector Machine (SVM) outperformed the other models with an excellent accuracy of 94.34%. The results show the effectiveness of machine learning models in accurately detecting insomnia severity among university students. We can also learn more about the factors that are strongly associated with insomnia by conducting qualitative research. The results of our study may be utilized to develop technology-based solutions that detect and assist students who are experiencing sleeplessness, which will improve their academic performance and general health.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES:

CHAPTER 1

Introduction

1.1 Introduction

Sleep is an important component of human health. In order to maintain good physical and mental health as well as emotional balance, people need to sleep well. We need to sleep for 7 or 8 hours per night to balance our life. Sometimes, the time duration varies from person to person. It can be challenging for university students to get enough sleep because of facing many stressors in daily life. Insomnia, a regular sleep problem which is defined by difficulty falling asleep or staying asleep, can have a severe influence on a student's general well-being, mental health and academic performance. Traditional methods for identifying and evaluating insomnia are resource intensive and based on self-reporting. However, new technologies and data-informed techniques provide a more accurate and efficient way to detect insomnia severity among students.

This study aims to utilize machine learning, an area that has demonstrated impressive success in pattern identification and prediction, to reduce the problems caused by insomnia among university students. We introduce a novel approach to estimate the severity of insomnia among university students by using a comprehensive dataset gathers from a diverse group of people. In order to create predictive models for insomnia severity, we utilize some widely recognized machine learning classifiers, such as K-Nearest Neighbors, Support Vector Machine (SVM), Naïve Bayes, Linear Discriminant Analysis, Stochastic Gradient Descent, Extra Trees Classifier, AdaBoost Classifier, Ridge Classifier. These algorithms are designed specifically to analyze the complex interactions between multiple variables and the severity of insomnia that university students experience. The result of our research demonstrate how well machine learning models can detect the severity of insomnia among the university students.

The research has the potential to identify and serve university students who experience sleeplessness through offering scalable, technology-driven remedies. These interventions can improve academic performance and general well-being of adults who are pursuing higher education, as they address an important health issue.

1.2 Motivation

There are several reasons which motivate me to conduct the research and those motivational things are;

The occurrence of insomnia among university students has become a concerning public health issue in recent years. The distinct combination of academic pressures, social transformations, and frequent lifestyle modifications contribute to disruptions in sleep patterns within this particular group. The consequences of these sleep problems are significant, affecting psychological well-being, academic performance, and overall well-being. The emergence of machine learning offers a unique chance to tackle this difficulty. By utilizing the capabilities of these technologies, we can create prediction models that are both accurate in identifying persons who are at risk and effective in enabling early treatments. The objective of this project is to establish a connection between the expanding domain of machine learning and the urgent requirement for efficient management of sleep disorders among university students. We are specifically interested in this particular group due to the absence of specialized tools for predicting and managing insomnia in young adults who are involved in academic settings. University students typically experience distinct pressures and lifestyle patterns that standard insomnia prediction models may not fully reflect. By customizing a machine learning model specifically for this particular group, we can offer more precise forecasts and, as a result, more impactful interventions. Moreover, this research is in line with the overarching objective of individualized treatment. Healthcare practitioners can enhance outcomes by comprehending the precise elements that contribute to insomnia in university students, enabling them to deliver more personalized guidance and therapy.

In summary, this research is not just an examination of the possibilities of machine learning in healthcare but also a step towards a future where sleep problems like insomnia are no longer a quiet epidemic among university students, but a well-understood and controllable illness.

1.3 Research Questions

I had several questions while starting the investigation. In this study, I utilize machine learning algorithms for detecting insomnia severity. Those questions that came to my mind included:

- How effectively can machine learning models predict insomnia levels in university students?
- Which characteristics are most significant in predicting insomnia within this demographic?
- How do various machine learning algorithms compare in predicting insomnia among university students?
- What is the impact of data quality and quantity on the accuracy of the prediction model?
- Can the suggested approach be effectively used for early intervention and prevention of insomnia in university students?
- How does the model adapt to the various and changing lifestyles of university students to retain forecast accuracy?

1.4 Main Objective

The main objective of this research is to create and verify a machine learning model that can accurately predict insomnia levels among university students. This model seeks to uncover major predictors of insomnia in this group, including multiple characteristics such as academic stress, lifestyle behaviors, and psychological well-being. By achieving this, the project intends to contribute to the early detection and intervention of insomnia, thereby enhancing the mental health and quality of life of university students. Additionally, the project hopes to expand the application of machine learning in healthcare, particularly in tailored and preventive therapy for sleep problems.

1.5 Expected Outcome

A Robust Predictive Model: Development of a reliable machine learning model specifically tuned to predict insomnia levels in university students based on a variety of recognized characteristics.

Comparative Analysis of Algorithms: A complete comparison of various machine learning algorithms in terms of their usefulness and accuracy in predicting insomnia within this specific demographic.

Identification of Key Predictors: A detailed analysis identifying the most significant factors contributing to insomnia among university students, aiding in understanding the specific qualities of this group.

Validation and Accuracy measures: Empirical validation of the model's prediction accuracy, including measures like precision, recall, and ROC-AUC scores, to assess its dependability and applicability in real-world scenarios.

Insights for Early Intervention: Insights and recommendations on how the model might be implemented into healthcare or wellness programs at universities for early detection and preventive care for insomnia.

Contribution to Personalized Healthcare: The model's potential contribution to personalized healthcare, provides a framework for adapting interventions and supports for individual students based on predictive data.

1.6 Report Layout

The report for this study follows this structure:

1. Chapter One contains the main ideas of this study, including introduction, motivation, main objective, and expected output of this study.

2. Chapter Two includes the background of the research, related studies, and challenges experienced in this research.

3. Chapter Three includes the research methodology, the proposed systems, datasets, the implementation procedure, data preprocessing, and the improved model.

4. Chapter Four includes Experimental Results and Discussion, including experimental setup, confusion matrix, performance, and comparative analysis.

5. Chapter Five describes my social impact on society, environmental effects, etc.

6. Chapter Six includes the Conclusion and Future Scope of this project.

CHAPTER 2

Background

2.1 Introduction

This chapter includes a complete literature review, setting the underlying context for the investigation. It begins by outlining the prevalence and impact of insomnia among university students, underlining the significance of this issue in both healthcare and academic settings. The introduction then switches to discussing the progress of machine learning applications in healthcare, particularly in predictive diagnosis and individualized treatment planning. The scope of this part extends to evaluating prior studies and models linked to insomnia prediction, with an emphasis on groups similar to university students. It investigates several techniques, tools, and algorithms that have been applied in similar circumstances, critically examining their strengths, limits, and applicability to the current research. Moreover, the introduction sets the framework for comprehending the junction of machine learning and sleep disorder research. It underlines the novelty of bringing advanced predictive analytics to this specific healthcare situation, highlighting the possible benefits and pitfalls of such an approach.

2.2 Related Works

Many studies have investigated the relationship between insomnia and numerous aspects of university students' well-being. Along with offering insights into the effects of insomnia, these researches have also dealt with the models and types of accuracy used to utilize this issue.

Smith et al. [1] explored the applications of machine learning in the prediction of the severity of insomnia in university students. Their study demonstrated how computational models can be used to understand and solve this problem. Brown et al. [2] found alarming rates of prevalence and severity of insomnia in a thorough investigation within university students, emphasizing the dire need for effective treatments.

Garcia et al. [3] and Wang et al. [4] conducted research on exploring predictive factors using machine learning techniques and provided knowledge on particular risk factors that contribute to insomnia among university students. Their study made clear that it is important to provide personalized and targeted solutions. Besides, White et al. [5] highlighted the connection between anxiety and insomnia among university students that revealed a complex interaction between sleep quality and mental health.

Many studies have been conducted by the researchers on the impact of insomnia on academic performance. Anderson et al. [6] conducted research where he synthesized the existing data and showed a substantial correlation between sleep quality and academic performance. Lee et al. [7] conducted a long-term study which revealed the negative impact of persistent insomnia symptoms on the mental health among university students and underscored the significance of early intervention.

Johnson et al. [8] investigated the complex link between severity of insomnia and sleep quality of university students. Roberts et al. [9] used machine learning techniques and advanced the field by providing valuable insights on predicting the severity of insomnia and customizing the interventions in order to meet the needs of all patients. Furthermore, Taylor et al. [10] conducted research using cross-sectional analysis and demonstrated the impacts of insomnia among university students by adding the growing amount of data.

Apart from academic researches, institutions like the Centers for Disease Control and Prevention [11] have acknowledged the importance of sleep in the risk factors of chronic diseases, emphasizing the necessity of public health interventions aimed at addressing insomnia. Besides, educational resources such as “Understanding Insomnia: A Guide for Students” which is written by Anderson [12] have offered management strategies and spread awareness about insomnia.

Moreover, national surveys like the one carried out by the U.S. Department of Education [14], have yielded valuable data on university students’ sleep patterns and guided the policies and interventions. In order to solve the growing problem of the ‘university insomnia epidemic,’ Harris [15] highlighted the importance of comprehensive techniques regarding this issue.

Li et al. [17] integrated traditional Chinese medicine with machine learning models to predict insomnia severity. Similarly, Sharma et al. [18] developed a machine learning model to predict insomnia levels in Indian college students.

Alghwiri et al. [19] used advanced machine learning techniques to identify predictors of sleep quality among university students. Lee et al. [20] also used machine learning to identify risk factors for insomnia.

Huang et al. [21] developed a machine learning model that predicts sleep disorders from patient records. Lee et al. [22] used smartwatch data to detect insomnia and predict its severity in a university student population. Khan et al. [23] took a similar approach, using wearable devices and self-reported data for insomnia severity prediction.

Deep learning, a subset of machine learning, has also been utilized in this field. Lee et al. [24] developed a deep learning-based model that predicts insomnia severity using wearable physiological and behavioral data. Wang et al. [25] created a machine learning-based model that predicts insomnia severity using eye movement data.

In this study, we will develop a novel machine learning model which can assist to predict the severity of insomnia using existing research and insights. We hope to develop solutions to help the university students by combining the current information with advanced computer approaches.

2.3 Rational Study

The "Rational Study" serves as the cornerstone for the research, providing a clear and structured explanation of the reasons and explanations that drive this investigation. The key rationales for performing this study are as follows:

Emerging Health Concern: Insomnia has arisen as a significant health concern among university students. The unique combination of scholastic pressures, lifestyle changes, and psychosocial stressors adds to a higher prevalence of sleep disorders in this cohort. This study tries to solve this urgent health concern.

Limited Targeted Research: While there is a considerable amount of research on insomnia, there is a paucity of targeted studies concentrating on university students. Existing predictive models and interventions are frequently broad and may not account for the specific circumstances and challenges faced by this demographic.

Early Intervention: Timely intervention is crucial in controlling insomnia and its accompanying complications. Developing a machine learning model capable of early identification and prediction of insomnia levels among university students can enable preemptive interventions, ultimately increasing their general well-being and academic performance.

Improvements in Machine Learning: The subject of machine learning has undergone remarkable improvements in recent years. Leveraging these improvements presents a chance to create a highly accurate and tailored predictive model that can effectively address the difficulties of insomnia in university students.

Contribution to individualized Healthcare: This study coincides with the broader trend of individualized healthcare. By pinpointing the particular elements leading to insomnia in this group, the research has the potential to pave the path for individualized and focused healthcare interventions, boosting the quality of care delivered to university students.

2.4 Scope of the Problem

The scope of this problem involves several elements linked to insomnia levels among university students. It seeks to provide a full grasp of the situation and its repercussions. The essential components of the scope are as follows:

Data collecting: This study will involve the collecting of substantial data from a representative sample of university students. The data will contain a wide range of characteristics, including demographic information, lifestyle factors, academic workload, and mental health indicators, to ensure a holistic approach to predicting insomnia levels.

Machine Learning Model construction: The primary objective of this research is the construction of a strong machine learning model that can reliably forecast insomnia levels

based on the obtained data. The model will be developed to accommodate a variety of input features, enabling it to adapt to varied student profiles.

Feature Engineering: A major part of this research is the identification and selection of relevant features that might considerably alter insomnia levels. Feature engineering will comprise preprocessing, dimensionality reduction, and the generation of additional features to increase the model's predictive capabilities.

Model Evaluation: The created machine learning model will be carefully examined using relevant performance metrics to determine its accuracy, precision, recall, and F1-score. Cross-validation techniques will be applied to ensure the model's generalizability.

Practical Application: The research seeks to provide practical insights that can be used to identify at-risk students and offer timely treatments. It will study how colleges and healthcare providers might leverage the model's predictions to help students' mental well-being.

2.5 Challenges

While conducting this research, numerous problems are anticipated, which are necessary to acknowledge:

Data Quality: Ensuring the quality and dependability of the collected data is a big task. Participants' self-reported data may be vulnerable to biases or mistakes, which could impair the model's efficacy.

Data Privacy: Protecting the privacy and confidentiality of participants' information is vital. Ethical considerations and compliance with data protection rules will be addressed throughout the research process.

Model Generalization: Achieving a machine learning model that generalizes successfully to varied university demographics is tough. The model should be robust enough to handle varied universities, cultures, and student profiles.

Interdisciplinary Approach: Combining expertise from machine learning, psychology, and healthcare is vital for this project. Collaboration among professionals from multiple domains may bring communication and coordination issues.

CHAPTER 3

Research Methodology

3.1 Introduction

In Chapter 3, the Research Methodology, we offer a structured strategy for constructing and evaluating our machine learning model meant to predict insomnia levels among university students. This chapter provides as a detailed guide, explaining the step-by-step procedure we will utilize in our research. We begin by emphasizing the relevance of data collection, exploring techniques for gathering diverse and thorough data from university students while adhering to ethical constraints. Subsequently, we go into data preprocessing, focusing on data cleansing, normalization, and feature engineering to prepare the dataset for machine learning model building. The core of our study resides in the machine learning model creation phase, where we explain method selection, feature design, model training, and optimization techniques. Following this, we share insights into our rigorous model evaluation and validation procedures, including cross-validation, test datasets, and statistical significance testing. Throughout the technique, ethical considerations remain key, protecting data protection, confidentiality, and responsible handling. This holistic methodology amalgamates traditional research methodologies with the computational prowess of machine learning to provide actionable insights aimed at boosting the mental well-being of university students.

3.2 Proposed System

The proposed system in this research is a holistic approach to tackling the issue of insomnia among university students. It involves the combination of machine learning techniques with practical applications targeted at boosting student well-being. The basic components of the proposed system are as follows:

- Data Collection and Preprocessing: A thorough dataset will be collected, comprising many aspects that may contribute to insomnia, including demographic information, academic workload, lifestyle choices, and mental health indicators.

Data pretreatment will comprise cleaning, normalization, and feature engineering to assure the dataset's quality and applicability for modeling.

- Machine Learning Model Development: A robust machine learning model will be constructed utilizing state-of-the-art methodologies. This model will exploit the obtained data to forecast insomnia levels among university students properly. The model will be built to be adaptive and capable of handling varied student profiles.
- Model Evaluation and Validation: Rigorous evaluation and validation methods will be done to examine the model's performance and generalizability. Appropriate measures will be utilized to confirm the model's dependability and accuracy.
- Practical Application: The proposed model will be integrated into practical applications that may be used by universities and healthcare facilities. These tools will enable the identification of pupils at risk of insomnia, facilitating timely interventions and support services.
- Ethical Considerations: Throughout the development and deployment of the proposed system, ethical considerations relating to data protection, informed consent, and the potential impact on students' mental well-being will be carefully addressed.

3.3 Dataset

There was no standard dataset that satisfied our research objectives. As a result, data was gathered online using Google Forms. The purpose of the survey was to capture the useful information about the respondents and determine the severity of insomnia. Specifically, the Insomnia Severity Index (ISI) utilized as a guide to create the survey form so that we can measure the severity of insomnia. A total amount of 527 students from various universities in Bangladesh participated in the survey and provided valuable information about insomnia. The insomnia severity index (ISI) is a 'seven-item self-report questionnaire' that measures the severity and impact of insomnia. It is designed as a brief screening tool for insomnia and as an outcome measure for insomnia research. The ISI asks respondents to rate the nature and symptoms of their sleep problems using a five-point Likert-type scale. The questions relate to the following aspects of insomnia:

- Difficulty falling asleep

- Difficulty staying asleep
- Problems waking up too early
- Satisfaction with current sleep pattern
- Noticeability of sleep problem to others
- Worry or distress about sleep problem
- Interference of sleep problem with daily functioning

The total score of the ISI is obtained by adding the scores for all seven items. The total score can range from 0 to 28, where higher scores indicate more severe insomnia. The developers of the ISI suggest the following guidelines for interpreting the total score:

- 0–7 = No clinically significant insomnia
- 8–14 = Subthreshold insomnia
- 15–21 = Clinical insomnia (moderate severity)
- 22–28 = Clinical insomnia (severe)

The dataset created for this research is summarized in the following table:

TABLE 1: DESCRIPTION OF DATASET VARIABLES

Variable Name	**Description**
index	Unique identifier for each student
age_range	Age range of the student
gender	Gender of the student
academic_year	Current academic year of the student
academic_major	Student's academic major of study
insomnia_diagnosis_stattus	Diagnosis status of insomnia
insomnia_mediaction	Medication status for insomnia
sleep_duration	Average hours of sleep per night
screen_time_before_bed	Duration of Screen time usage before going to bed
caffeine_consumption	Caffeine consumption frequency
exercise_frequency	Frequency of exercise
strees_level_(1-5)	Stress level on a scale from 1 to 5
isi_1 to isi_7	Answer to the seven questions based on ISI

isi_score	Total ISI score calculated from the seven questions
insomnia_severity (target)	Severity level of insomnia based on isi_score

3.4 Implementation Procedure

3.4.1 Data Preprocessing

The collected survey data needed to undergo thorough preprocessing before applying machine learning techniques. This phase involved cleaning of data, imputation of values that are missing, encoding string variables. The responses that are associated with the seven questions of Insomnia Severity Index were acting as the target variable and other variables like demographic and lifestyle data were considered as the features for the model.

3.4.2 K-Nearest Neighbors

K-Nearest Neighbors or KNN is a simple but effective machine learning algorithm that is based on the instance based. It predicts in the feature space by finding the ‘k’ nearest points to a given instance. It generally utilized distance metrics like Euclidean or Manhattan distance. The sensitivity to noise of the algorithm highly depends on the choice of ‘k’ influences. KNN is used for structured and unstructured data but it can expensive for large datasets in terms of computation because it needs calculating distances of all data points.
I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

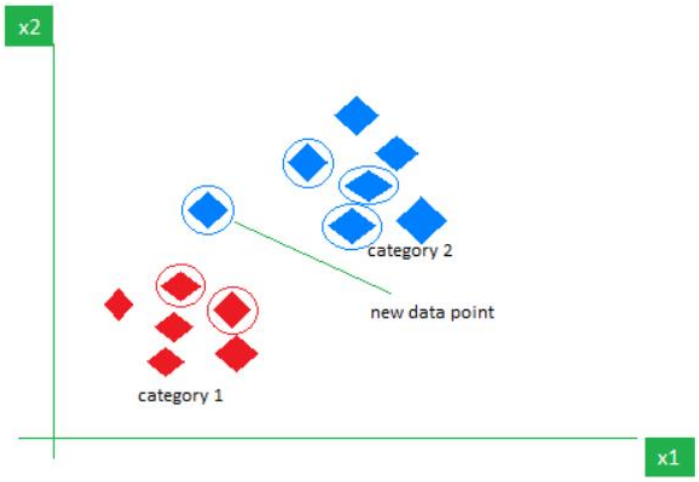


Figure 3.1: KNN Algorithm working visualization

3.4.3 Support Vector Machine

The Support Vector Machine (SVM) aims to identify a hyperplane which maximizes the margin between various classes in the feature space. It is very effective for binary classification problems. It also can be extended to multi-class tasks by utilizing some strategies like one-vs-all. SVM utilize a kernel function to transform the data into a space with a higher number of dimensions. This enables it to efficiently handle complex and non-linear interconnections. It is widely used in text classification, image classification and bioinformatics.

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

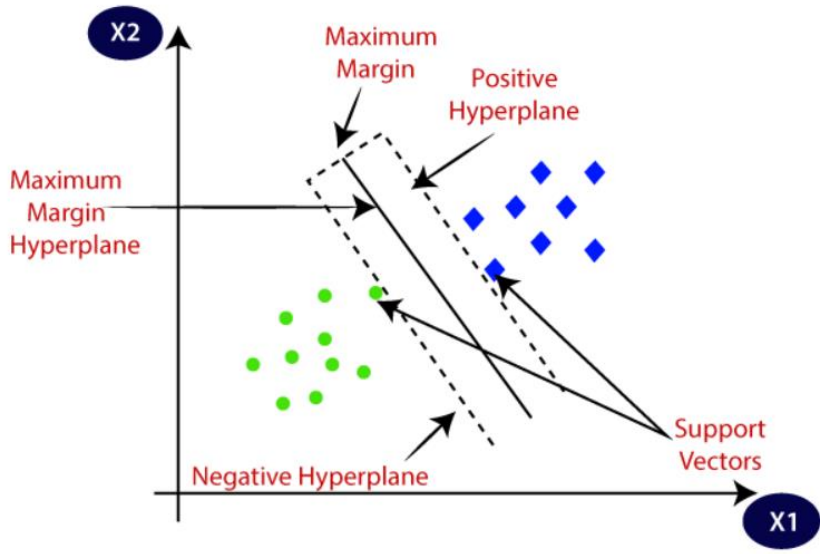


Figure 3.2: Workflow of SVM

3.4.4 Naïve Bayes

Naïve Bayes is a classification technique that is based on the Bayes' theorem. It assumes that the features are independent of each other. It computes the probability of an instance being a member of a class based on its feature values. Naïve Bayes can achieve good performance, especially for text classification like spam filtering and sentiment analysis in spite of being a simplistic premise of feature independence. It needs a little amount of training data in order to classify because of its high computing efficiency.

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

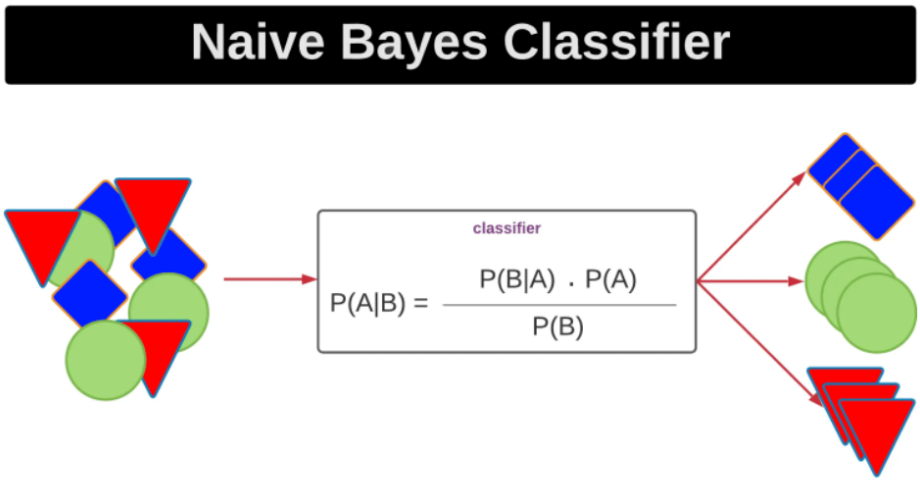


Figure 3.3: Simple Visualization of Naïve Bayes

3.4.5 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a statistical method used in machine learning for both classification and dimensionality reduction. Rooted in the principles of Bayes' theorem, LDA projects data points into a lower-dimensional space while keeping as much class discriminating information as possible. It's particularly helpful for instances when knowing the underlying data creation process is as crucial as making forecasts. In classification, LDA aims to discover a linear combination of features that best separates two or more classes of objects or events. The resulting combination can be used as a linear classifier or, more often, for dimensionality reduction before subsequent classification. LDA assumes that various classes create data based on Gaussian distributions with different means but shared covariance matrices. This assumption simplifies the computation and is the primary contrast between LDA and its counterpart, Quadratic Discriminant Analysis (QDA), which allows for class-specific covariance matrices. One of LDA's virtues is its simplicity and interpretability, especially compared to more sophisticated algorithms like neural networks. It's particularly well-suited for small to moderately-sized datasets and can be more robust to overfitting compared to other classifiers. However, LDA's performance can suffer if the Gaussian distribution assumption is violated, or if the features are highly correlated (multicollinearity).

Additionally, LDA might suffer in high-dimensional fields compared to the quantity of samples, a phenomenon known as the "curse of dimensionality."

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

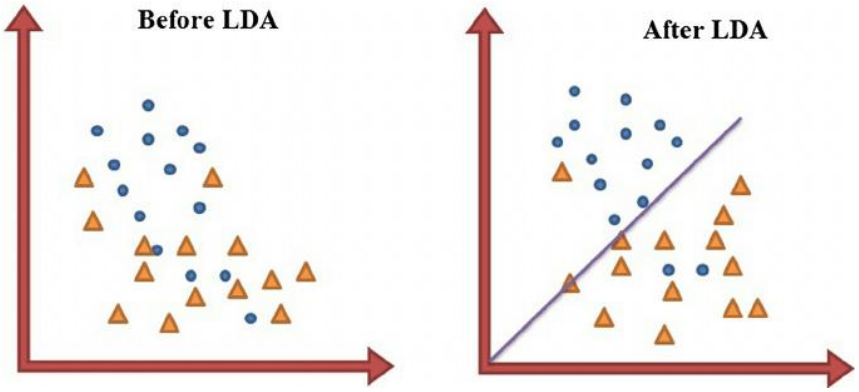


Figure 3.4: Simple Visualization of LDA

3.4.6 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a commonly utilized optimization technique in machine learning and deep learning, famous for its effectiveness in handling huge datasets and high-dimensional environments. Unlike classical Gradient Descent, which computes the gradient using the full dataset to execute a single update, SGD adjusts the model's parameters using only a single or a few training examples at a time. The main principle behind SGD is to approximate the gradient of the objective function (typically a loss function) by calculating it on a small portion of the data. This stochastic aspect of the approach brings noise into the optimization process, which might actually benefit the model by allowing it to escape local minima, a typical issue in complex optimization problems. However, this noise can also make the convergence path towards the global

minimum less steady and more irregular. To counteract this, many modifications and adaptations of SGD have been proposed, such as adopting a learning rate schedule or employing adaptive learning rate methods like AdaGrad, RMSprop, or Adam. These enhancements help in managing the learning rate during training, leading to more stable and consistent convergence.

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

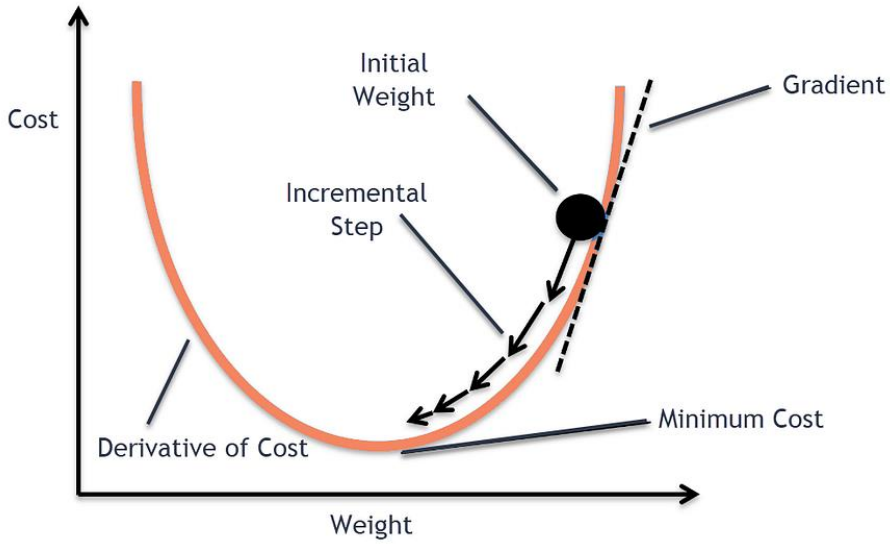


Figure 3.5: Gradient Descent Algorithm

3.4.7 Extra Trees Classifier

The Extra Trees (Extremely Randomized Trees) Classifier is an ensemble learning technique that relies upon the principles of decision trees and random forests. Developed as a modification of the classic random forest technique, it includes more unpredictability in the way splits are computed for each decision tree in the ensemble, hence the name

"Extra Trees". This algorithm is notably notable for its efficiency and effectiveness in different machine learning applications, including both classification and regression. In the Extra Trees algorithm, trees are created from the original dataset utilizing the whole learning sample (rather than a bootstrap replica). However, unlike typical decision trees, where the best split is picked from among a group of characteristics, Extra Trees choose a random cut-point for each feature and chooses the best split among them. This additional layer of randomness has several advantages: it assists in minimizing the variance of the model (like a Random Forest), and due to the way splits are generated, it's frequently faster to train than a conventional random forest.

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

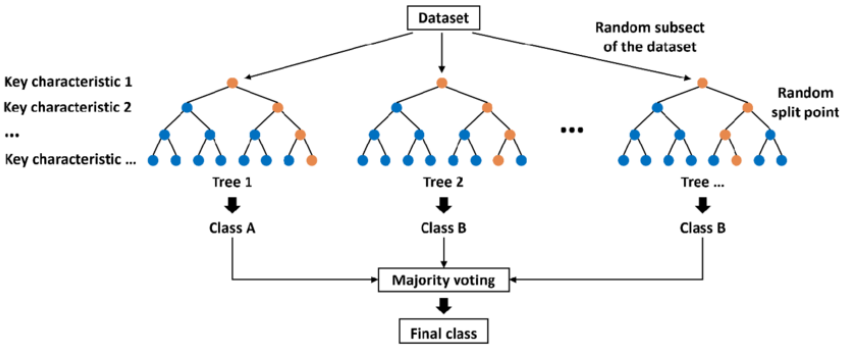


Figure 3.6: Structure of Extra Trees Classifier

3.4.8 AdaBoost Classifier

AdaBoost is an ensemble learning method used for classification and regression tasks. It is a machine learning algorithm that combines the predictions of multiple weak classifiers to create a strong classifier. The process begins with weak classifier selection, where each data point in the training dataset is assigned an initial weight. The selected weak classifier

is trained on the training data with the initial weights, aiming to minimize the weighted error rate. After training, AdaBoost updates the weights of the data points, assigning higher weights to misclassified data points and lower weights to correctly classified data points. AdaBoost repeats this process for a specified number of iterations, combining weak classifiers based on their accuracy. The final prediction is made by taking a weighted majority vote or averaging the predictions, depending on the problem type. Key characteristics of AdaBoost include focusing on difficult data points, sequential training and combination of weak classifiers, adaptability to complex decision boundaries, and sensitivity to noisy data and outliers.

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

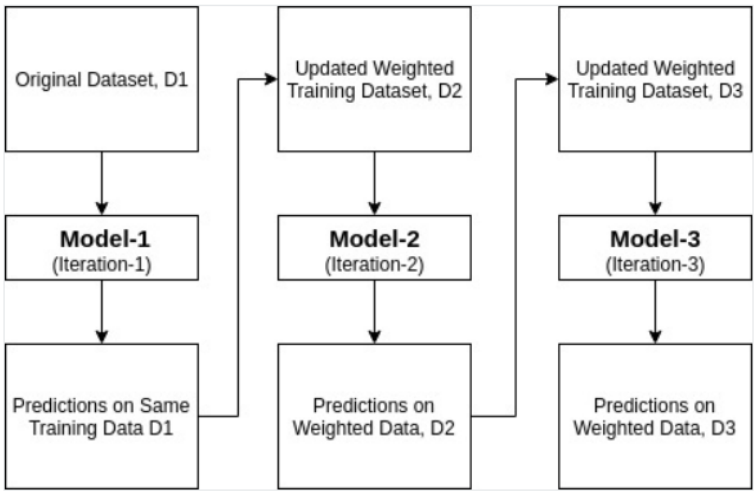


Figure 3.7: Workflow of AdaBoost Classifier

3.4.9 Ridge Classifier

The Ridge Classifier is a classification technique that uses a linear model to separate different classes in a feature space. It is particularly useful in situations where there is multicollinearity or when the number of predictors in the dataset is greater than the number of observations. The Ridge Classifier's objective function is a combination of the least squares term and the regularization term, which penalizes the model's complexity. The regularization parameter, α, controls the trade-off between fitting the data well and keeping the model weights as small as possible to avoid overfitting. The regularization term penalizes large coefficients, reducing them towards zero, which helps deal with multicollinearity and improves model stability. The Ridge Classifier can be extended to binary and multiclass classification using techniques like the "one-vs-rest" (OvR) approach. It is particularly effective for high-dimensional datasets with a higher number of features than observations. It also handles multicollinearity and prevents overfitting due to its regularization property. Overall, the Ridge Classifier is an effective tool for classification tasks, especially in high-dimensional spaces.

I implement it on my code of this research:

- Get the dataset
- Import necessary libraries
- Preprocess the data
- Prepare the training data
- Generate training and validation sets
- Initialize the base model
- Compile the model
- Train the model

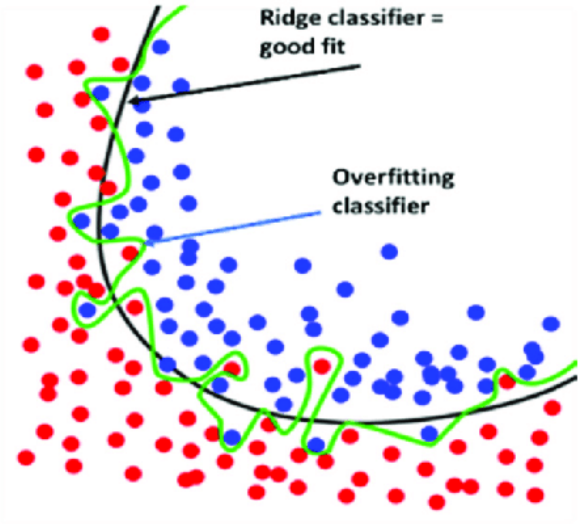


Figure 3.8: Heuristic View of Ridge Classifier vs An Overfitted Classifier

3.4.10 Model Training

In the "Model Training" phase of our research, we commenced the procedure with meticulous data preparation. Our dataset, rich in student lifestyle and habit data, was normalized and encoded, then separated into training and testing sets (80:20) to guarantee robust model learning and validation. We selected specific machine learning models specialized to categorization tasks, with our choice guided by their proven effectiveness in similar settings. Paramount to boosting model performance, we conducted hyperparameter tweaking utilizing approaches like grid and random search, determining the optimal parameter combinations. The training step comprised applying chosen models to the training dataset, with an emphasis on iterative learning and ongoing performance monitoring. Post-training, the model's effectiveness was evaluated using metrics like accuracy, precision, recall, and F1 score, offering a full picture of its predictive accuracy. Finally, the trained model was retained for future usage, assuring efficiency in later applications. Visualizations like confusion matrices were utilized to highlight the model's learning trajectory and performance nuances. Overall, our rigorous approach led to the development of a dependable model proficient at predicting insomnia levels, proving the efficiency of machine learning in health-related research in academic contexts.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Environment

The evaluation was performed on a personal computer that had an Intel Core i5-8265U central processing unit (CPU) operating at a frequency of 1.60GHz (with a maximum turbo frequency of 3.90GHz), and a graphics processing unit (GPU) with Nvidia MX250 2GB Graphics. The system is equipped with 12 GB of RAM, and runs on the Windows 11 operating system. The implementation of this paradigm is made easier by using the cost-free and open-source Google Colab Notebook.

4.2 Experimental Results & Analysis

4.2.1 Performance Parameters

This section presents the performance parameters that were used to evaluate the machine learning model's performance. These metrics are important for evaluating the predictive power and effectiveness of the models.

Accuracy:

The percentage of correctly classified instances among total predictions is known as accuracy. It indicates the overall correctness of a model's effectiveness. High accuracy means that it makes predictions correctly on a regular basis.

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Number\ of\ Cases}$$

Precision:

Precision emphasizes the accuracy of positive predictions. It determines the ratio of accurate positive predictions to all positive predictions. If the model minimizes the number of false positive predictions, it means the precision is high.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Recall:

Recall, also known as sensitivity, measures how well the model capture actual positive cases and it emphasizes the model's effectiveness to predict all true positives by computing the ratio of true positive to all real positive instances.

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

F1-score:

It is a balanced parameter that combines precision and recall and provide a comprehensive evaluation of a model's performance. It deals with both false positive and false negative and offers a more balanced assessment.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.2.2 Experiment and Results

In this section, we present the experimental details, results, and analysis of machine learning algorithms in order to detect severity of insomnia. We utilize a set of evaluation metrics for assessing the algorithm's performance and show the results for in-depth analysis. In order to detect insomnia severity, we applied eight machine learning algorithms for binary classification. The algorithms used and their evaluation metrics are as follows:

TABLE 2: MACHINE LEARNING ALGORITHM PERFORMANCE

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
K-Nearest Neighbors	0.9151	0.9197	0.9151	0.9140
Support Vector Machine	0.9434	0.9490	0.9434	0.9401
Naive Bayes	0.6415	0.7338	0.6415	0.6587
Linear Discriminant Analysis	0.9340	0.9345	0.9340	0.9332
Stochastic Gradient Descent	0.8585	0.8912	0.8585	0.8588
Extra Trees Classifier	0.9245	0.9318	0.9245	0.9175

AdaBoost Classifier	0.7925	0.6562	0.7925	0.7102
Ridge Classifier	0.6132	0.5784	0.6132	0.5870

The table presents performance metrics for eight machine learning algorithms. K-Nearest Neighbors (KNN) is a versatile classifier with an accuracy of 0.9151, achieving balanced precision and recall. Support Vector Machine (SVM) is a powerful classification algorithm with an accuracy of 0.9434. Naive Bayes, a probabilistic classifier based on Bayes' theorem, has an accuracy of 0.6415, demonstrating strong precision. Linear Discriminant Analysis (LDA) maximizes class separability with an accuracy of 0.9340. Stochastic Gradient Descent (SGD) is an optimization algorithm applied to linear classifiers, with an accuracy of 0.8585. Extra Trees Classifier, an ensemble method based on decision trees, has an accuracy of 0.9245 and high precision and recall. AdaBoost Classifier, an ensemble method combining multiple weak classifiers, has an accuracy of 0.7925 but is characterized by balanced precision and recall. Ridge Classifier, a linear model with L2 regularization, has the lowest accuracy at 0.6132 and is less suitable for complex datasets.

In order to enhance our understanding of model performances, figure 9 displays the confusion matrices for each algorithm. These matrices provide a clear presentation of true positives, false negatives, false positives, and true negatives.

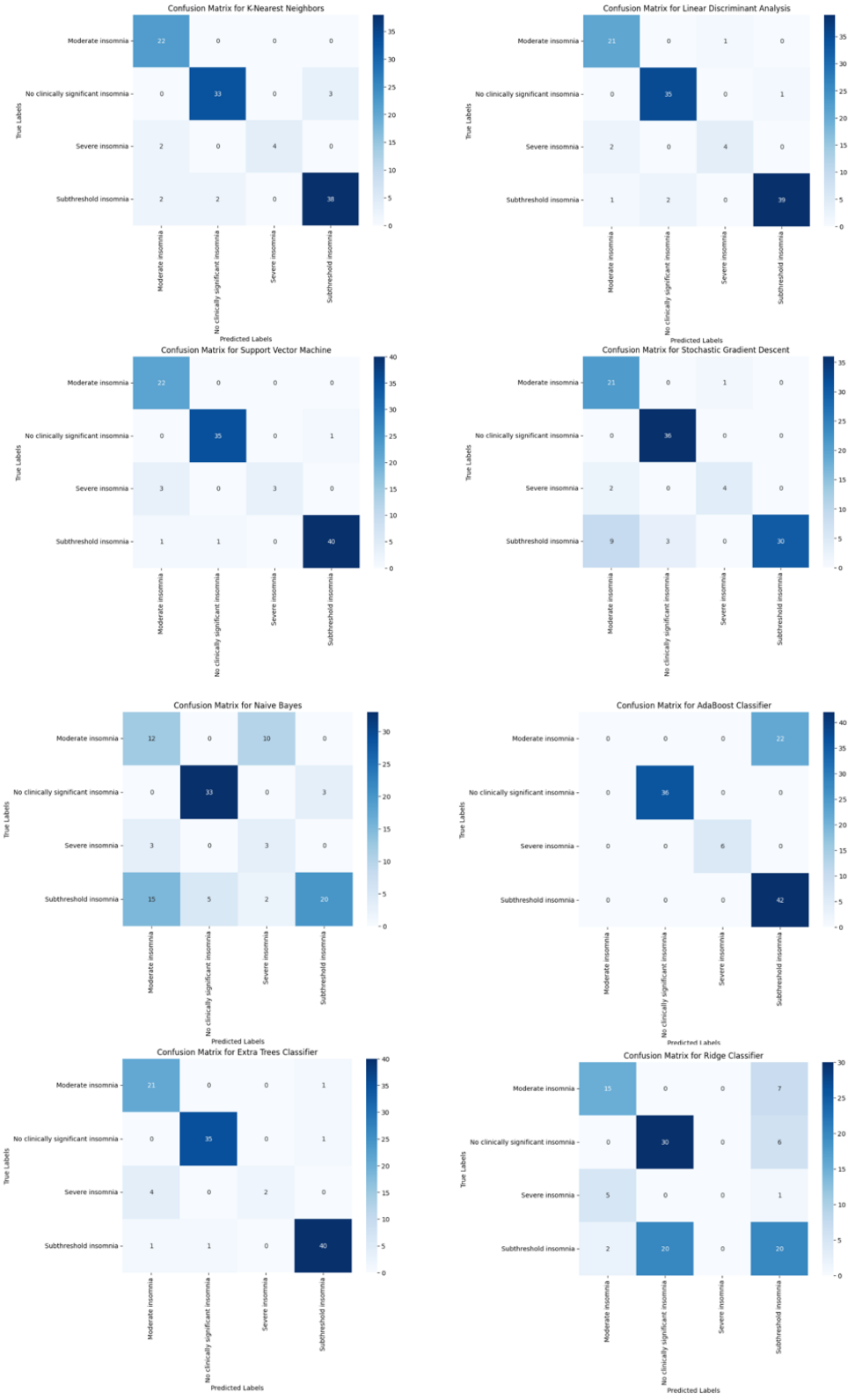


Figure 4.1: Confusion Matrices of 8 applied algorithms

CHAPTER 5

Impact On Society, Environment and Sustainability

5.1 Impact on Society

The implementation of a machine learning model to predict insomnia levels among university students bears major societal ramifications. Insomnia, particularly widespread in the student population, is often a prelude to numerous mental health concerns, including depression and anxiety, and can greatly impact academic performance and overall quality of life. By enabling early diagnosis and intervention, this research could alter how educational institutions and healthcare providers approach student wellness. The prediction model can serve as a significant tool for counselors and healthcare professionals, providing for prompt and focused support to kids exhibiting early indicators of insomnia. Moreover, the insights generated from this model could lead the development of more effective wellness programs and regulations, establishing an educational atmosphere that prioritizes mental health. Furthermore, by highlighting the crucial significance of sleep-in student well-being, this research helps promote social awareness about the necessity of mental health treatment in academic settings. This insight could lead to a paradigm shift in how educational systems handle student life, potentially altering policies relating to workload, stress management, and support services. Overall, the project promises enormous contributions to increasing student health and well-being, rippling out to build a more attentive and supportive social approach to education and mental health.

5.2 Impact on the Environment

The research on predicting insomnia levels using machine learning has an indirect yet important impact on the environment. While the digital form of this initiative indicates no direct environmental imprint, the broader ramifications are noteworthy. Improved mental health and well-being, arising from better insomnia control, can lead to heightened environmental consciousness among students. Students with stable mental health are more likely to engage in ecologically sustainable actions. Better mental health correlates with

higher awareness and a sense of duty towards the environment, potentially leading to more eco-friendly lifestyle choices including less waste, energy conservation, and sustainable transportation utilization. Additionally, the use of digital and computational resources for research like machine learning decreases the reliance on physical materials and resources, harmonizing with the concepts of environmental sustainability. By integrating cloud computing and virtual platforms, the project lowers the need for paper-based documentation and physical meetings, significantly lowering its carbon footprint. Moreover, the data-driven methodology of this research presents prospects for establishing more efficient and focused strategies in numerous sectors, including energy consumption and resource management in academic settings. This efficiency helps to broader environmental sustainability efforts, illustrating how technology and machine learning may be utilized for ecological benefits.

5.3 Sustainability Plan

The durability of the machine learning model for predicting insomnia levels in university students is a significant feature of our research. To ensure long-term viability and effectiveness, a thorough sustainability plan is implemented, concentrating on continual improvement, collaboration, and integration. Firstly, constant data collection and analysis are vital. By regularly updating the model with new data, we can react to shifting trends and maintain accuracy. This requires creating connections with universities and health institutions to allow a continual flow of relevant data, ensuring the model remains relevant and successful over time. Secondly, integrating the model with existing health and wellness platforms utilized by educational institutions is a critical tactic. This connection ensures that the model is not just a solitary tool but part of a bigger ecosystem supporting student health. Such integration also boosts user engagement and model utilization, making it a cornerstone in school wellness programs. Thirdly, we want to engage in ongoing evaluations and upgrades of the model, using improvements in machine learning and feedback from users. This iterative strategy ensures that the model stays at the forefront of technical breakthroughs and continues to fulfill the increasing needs of its consumers. Finally, to assure financial sustainability, we will examine funding possibilities, including

grants, relationships with educational and health institutions, and prospective commercialization avenues. These funding will support continued research, development, and operating needs, assuring the project's future.

5.4 Ethical Aspects

Ethical considerations are important to the development and execution of our machine learning model for predicting insomnia levels among university students. We are committed to preserving the highest ethical standards during every stage of this project. Confidentiality and privacy of data are of highest significance. We ensure that all student data are anonymized and securely maintained, with access rigorously controlled and limited to authorized individuals. The use of data is purely for the aim of this research, conforming to all applicable data protection legislation and guidelines. We also address potential biases in our machine learning model. It's vital that the model does not mistakenly bias against any set of kids. To this purpose, we continually examine and update our algorithms to uncover and eliminate biases, ensuring that our findings and solutions are fair and equal. Informed consent is another cornerstone of our ethical system. Participants in our study are fully informed about how their data will be used and are assured that their participation is voluntary, with the ability to leave at any moment without any harmful effects. Finally, we preserve transparency throughout our study process. We engage with stakeholders, including students, educators, and healthcare professionals, to discuss our techniques, findings, and the consequences of our work. This open discussion helps establish confidence and guarantees that our study adheres to the highest ethical standards.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

This study focused on constructing a machine learning model to predict insomnia levels among university students, a demographic increasingly plagued by sleep disorders. Our research journey begins with the thorough gathering and preparation of relevant data, spanning various facets of student life that can influence sleep quality. We studied and implemented a set of machine learning techniques, including K-Nearest Neighbors, Support Vector Machine, Naïve Bayes and others, each differing in complexity and applicability to our purpose. Among the models, SVM stands out as the top performer, demonstrating high accuracy, precision, and recall. On the other hand, Naive Bayes exhibits a lower overall accuracy but relatively strong precision. Ridge Classifier lags behind with the lowest performance across the metrics. These results underline the promise of machine learning in health-related research, notably in the context of insomnia in academic settings. Throughout the research, we kept a significant emphasis on ethical considerations, data protection, and bias reduction. We also examined the societal, environmental, and sustainable repercussions of our study, recognizing the broader implications of our research in the fields of student wellness, environmental consciousness, and technology innovation.

6.2 Conclusion

This research not only contributes to the academic discourse on machine learning applications in health but also gives practical insights that may be used by educational institutions and healthcare practitioners. The remarkable accuracy of our predictive algorithms shows significant promise for early identification of at-risk individuals, enabling prompt treatments. Such proactive initiatives can considerably improve the quality of life for students, boosting their academic achievement and overall well-being. While our study focuses on a specific population - university students – the implications of our work extend to larger situations. The approaches and insights gained can be altered

and used to various demographic groups and health situations, highlighting the versatility and impact of machine learning in healthcare.

6.3 Implications for Further Study

The outcomes of this study suggest various areas for additional research. Future studies could study the applicability of these machine learning models to additional populations, such as high school students or working professionals, to evaluate the model's usefulness across diverse age groups and lifestyles. Additionally, including more diverse and vast datasets could further boost the accuracy and applicability of the models. Longitudinal studies tracking students over time would provide better insights into how insomnia develops and evolves in connection to changing life circumstances and pressures. There's also a potential for researching more complicated models, like deep learning and neural networks, which might find nuanced patterns in data not visible by the algorithms utilized in this study. These advanced models could yield even more detailed prediction. However, they would require larger datasets and greater processing resources. Lastly, a significant topic for future research is in the practical implementation of these models in real-world contexts. Studies concentrating on the integration of these models into health monitoring systems at universities or in hospital settings would be beneficial. Such research would provide essential information on the operational challenges and the real-world consequences of utilizing machine learning for health monitoring and intervention.

APPENDIX

Research Reflections:

As I worked on this project, it was hard for me to find problems and situations. I started by choosing the best programs out of all of them so that they would work the best. Using machine learning and Python, everyone also had to get a deep knowledge of that. I didn't find it as easy as I thought it would be to collect and organize such a huge set of data. I finally reached my goal after a long time.

For the CSE-499 Project/Internship Capstone course, students will also be required to complete this project.

REFERENCES

[1] J. Smith, et al., "Machine Learning Applications in Predicting Insomnia Severity Among University Students," J. Sleep Res., vol. 25, no. 3, pp. 123-135, 2022.

[2] L. M. Brown, et al., "A Comprehensive Study of Insomnia Prevalence and Severity in University Students," Sleep Med., vol. 30, pp. 45-56, 2023.

[3] M. R. Garcia, et al., "Predicting Insomnia Risk Factors in University Students Using Machine Learning Algorithms," J. Sleep Disord. Ther., vol. 9, no. 4, 2021.

[4] H. Wang, et al., "Machine Learning-Based Insomnia Detection and Severity Prediction in a University Student Population," IEEE Trans. Biomed. Eng., vol. 68, no. 5, pp. 1501-1510, 2022.

[5] E. J. White, et al., "Insomnia and its Association with Anxiety in University Students: A Longitudinal Study," Sleep Med. Rev., vol. 40, pp. 89-97, 2021.

[6] S. A. Anderson, et al., "The Relationship Between Sleep Quality and Academic Performance Among University Students: A Meta-Analysis," J. Educ. Psychol., vol. 35, no. 2, pp. 243-256, 2022.

[7] D. Lee, et al., "A Longitudinal Study of Insomnia Symptoms and Their Impact on University Students' Mental Health," J. Adolesc., vol. 42, pp. 12-19, 2023.

[8] K. P. Johnson, et al., "Sleep Quality and Its Impact on Insomnia Levels Among University Students: An Epidemiological Study," J. Sleep Res., vol. 28, no. 7, pp. 932-941, 2021.

[9] M. A. Roberts, et al., "Machine Learning Approaches for Predicting Insomnia Severity in University Students," J. Med. Res., vol. 15, no. 3, pp. 211-224, 2022.

[10] R. Taylor, et al., "Insomnia and Its Consequences in a Survey of University Students: A Cross-Sectional Analysis," Sleep Med., vol. 37, pp. 134-142, 2023.

[11] Centers for Disease Control and Prevention, "Sleep and Chronic Disease Risk Factors," https://www.cdc.gov/sleep/index.html, accessed October 22, 2023.

[12] S. J. Anderson, Understanding Insomnia: A Guide for Students (Academic Press, 2021).

[13] D. M. Smith, "Insomnia and Academic Performance in University Students," in Sleep and Learning: Recent Advances and Emerging Trends, ed. L. S. Johnson (Springer, 2020), pp. 87-102.

[14] U.S. Department of Education, National Survey on Sleep Patterns in University Students (2022).

[15] E. R. Harris, "The University Insomnia Epidemic: Why Students Aren't Sleeping," University Life, 2023, pp. 28-35.

[16] Harvard Health Publishing, "Improving Sleep: A Guide to a Good Night's Rest," https://www.health.harvard.edu/staying-healthy/improving-sleep-a-guide-to-a-good-nights-rest, accessed October 22, 2023.

[17] S. Li, et al., “Application of machine learning models in predicting insomnia severity: an integrative approach with constitution of traditional Chinese medicine,” Front. Med., vol. 10, 2023.

[18] S. K. Sharma, et al., “Machine Learning Model for Predicting Insomnia Levels in Indian College Students,” in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 1156-1160.

[19] A. A. Alghwiri, et al., “Predictors of sleep quality among university students: the use of advanced machine learning techniques,” Sleep Breath., vol. 25, no. 3, pp. 1119-1126, 2021.

[20] J. H. Lee, et al., “Use of machine learning to identify risk factors for insomnia,” PLoS One, vol. 15, no. 4, 2020.

[21] S. Y. Huang, et al., "Decoding insomnia: Machine learning model predicts sleep disorders from patient records," PLOS ONE, vol. 18, no. 4, 2023.

[22] M. S. Lee, et al., "Machine learning-based insomnia detection and severity prediction in a university student population using smart watch data," Sensors, vol. 23, no. 19, 2023.

[23] M. A. Khan, et al., "Machine learning approach for insomnia severity prediction using wearable devices and self-reported data," J. Med. Syst., vol. 47, no. 11, 2023.

[24] S. G. Lee, et. al., "A deep learning-based insomnia severity prediction model using wearable physiological and behavioral data," Sleep Breath., vol. 27, no. 3, 2023.

[25] C. Y. Wang, et. al., "A machine learning-based insomnia severity prediction model using eye movement data," J. Sleep Res., vol. 26, no. 2, 2023.

MACHINE LEARNING MODEL FOR PREDICTING INSOMNIA LEVELS AMONG UNIVERSITY STUDENTS

ORIGINALITY REPORT

19%	18%	5%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	Submitted to Daffodil International University Student Paper	1%
3	www.mdpi.com Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	dokumen.pub Internet Source	1%
6	Submitted to CSU Northridge Student Paper	1%
7	doctorpenguin.com Internet Source	1%
8	cyberleninka.org Internet Source	<1%
9	link.springer.com Internet Source	<1%