**STUDENT STRESS FACTOR DETECTION THROUGH CLASSIFICATION AND FEATURE SELECTION**

### This project report is submitted to

### Silicon University, Odisha

***in partial fulfillment of the requirements for the award of the degree of***

**Bachelor of Technology**

***in***

**Computer Science and Engineering**

### Submitted by

|  |  |
| --- | --- |
| **Anubhav Mohanty** | (Regd. No. : 2001209087 ) |
| **Payal Pani** | (Regd. No. : 2001209147) |
| **Rajat Kumar** | (Regd. No. : 2001209058) |
| **Sagar Verma** | (Regd. No. : 2001209168) |

**Project Group No.: CSE 27**

### Under the Esteemed Supervision of

**Prof. (Dr) Pamela Chaudhury**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SILICON UNIVERSITY ODISHA**

**SILICON HILLS, BHUBANESWAR – 751024, ODISHA, INDIA**

**May 2024**

**CERTIFICATE**

This is to certify that the work contained in the project entitled **“Student Stress Factor Detection through Classification and Feature Selection ”**, submitted by **Anubhav Mohanty (Regd. No.: 2001209087), Payal Pani (Regd. No.: 2001209147), Rajat Kumar (Regd. No.: 2001209058)** and **Sagar Verma (Regd. No.: 2001209058)** is a record of bonafide works carried out by them under my supervisionand guidance. The contents embodied in the project is being submitted as a part of 8th semester project for the undergraduate curriculum and have not been submitted for the award of any other degree or diploma in this or any other university.

## Date:1st May 2024

## Place: Bhubaneswar

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Prof. (Dr) Pamela Chaudhury

## Senior Assistant Professor

## Department of Computer Science & Engineering

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## External Examiner



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SILICON UNIVERSITY ODISHA**

**BHUBANESWAR – 751024**

**DECLARATION**

We hereby certify that:-

1. The work contained in the project is original and has been done by ourselves under the supervision of oursupervisor.
2. The work has not been submitted to any other Institute for any degree ordiploma.
3. We have conformed to the norms and guidelines given to us by the Project Review Committee of our department.
4. Whenever we have used materials (data, theoretical analysis and text) from other sources, we have given due credit to them by citing them in the text of the project and giving their details in thereferences.

Date : 1st May 2024

Place: Bhubaneswar

|  |  |  |  |
| --- | --- | --- | --- |
| **Anubhav Mohanty** | (Regd. No. 2001209087) | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Payal Pani** | (Regd. No. 2001209147) | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Rajat Kumar** | (Regd. No.2001209058) | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
| **Sagar Verma** | (Regd. No.2001209168) | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SILICON UNIVERSITY ODISHA**

**BHUBANESWAR – 751024**

**ACKNOWLEDGEMENTS**

We would like to express our heartfelt gratitude to Prof., Dr. Pamela Chaudhury for her invaluable guidance and support throughout our research journey. Her expertise, encouragement, and feedback have been instrumental in shaping this work. We would also like to extend our appreciation to all the individuals who have provided assistance and support during this project. Their contributions have been essential in making this work possible. Lastly, we would like to thank our colleagues and fellow team members who have shared their insights and expertise, which have greatly enriched our understanding and contributed to the success of this project.

|  |  |
| --- | --- |
| **Anubhav Mohanty** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Payal Pani** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Rajat Kumar** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Sagar Verma** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SILICON UNIVERSITY ODISHA**

**BHUBANESWAR – 751024**

**ABSTRACT**

A Student stress, a prevalent concern affecting academic performance and well-being, is the focal point of this project. Leveraging advanced machine learning techniques such as classification, feature selection, and optimization, the primary aim is to develop a robust model capable of precisely identifying and quantifying stress factors across academic, personal, and external domains. Innovatively, we enhance our dataset by incorporating additional samples from Kaggle, ensuring diversity and relevance. Techniques like classification and feature optimization play a crucial role in crafting a comprehensive stress detection model. The evaluation will rely on a meticulous analysis of performance metrics, ensuring the model's accuracy and reliability. In a pioneering stride, the project introduces an interactive interface to address limitations found in existing works, empowering students with a feedback mechanism for a more personalized and effective stress detection experience. The dataset, meticulously designed to reflect the present scenario, fortifies the project's real-world applicability.

***Keywords****: Student Stress Detection, Classification Algorithm*

# LIST OF ABBREVIATIONS

## Abbreviation Description

ADAM Adaptive Moment Estimation

KNN K nearest Neighbor

ML Machine Learning

ReLU Rectified Linear Unit

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
|  |  | **Page #** |
| **Chapter4.** |  |  |
| Figure 4.1. | Stress Factor Dataset Features | 1 |
| Figure 4.2. | Correlation of features with Stress level | 3 |
| Figure 4.3.  Figure 4.4.  Figure 4.5.  Figure 4.6.  Figure 4.7.  Figure 4.8.  Figure 4.9.  Figure 4.10. | Correlation Matrix  Classification report of KNN Classification  Classification report of Logistic Regression  Classification report of Naïve Bayes Classifier  Classification report of Decision Tree Classification  Classification report of Neural Network  Accuracy scores of different algorithms  Feature Importance based on Neural Network weights | 8 |

# LIST OF TABLES

|  |  |  |
| --- | --- | --- |
|  |  | **Page #** |
| **Chapter4.** |  |  |
| Table 4.1. | Different factors contributing to student stress | 2 |
| Table 4.2. | Comparing the different models | 3 |

# CONTENTS

## CONTENTDETAILS PAGE NO.

|  |  |  |
| --- | --- | --- |
| **Title Page** |  | i |
| **Certificate** | | ii |
| **Declaration** |  | iii |
| **Acknowledgements** | | iv |
| **Abstract** |  | v |
| **List of Abbreviations** | | vi |
| **List of Figures** | | vii |
| **List of Tables** |  | viii |
| **Contents** |  | ix |
|  |  |  |
| **Chapter 1.** | ***Introduction*** | **1 – 12** |
|  |  |  |
| **1.1.** | **Background** | 3 |
| **1.2.** | **Problem Statement** | 4 |
| **1.3.**  **1.4.**  **1.5.** | **Objectives of the Project**  **Proposed Model**  **Organization of the Project**  **Summary** | 6 |
|  |  |  |
| **Chapter 2.** | ***Literature Survey*** | **13 – 18** |
|  |  |  |
| **2.1.** | **Introduction** | 13 |
| **2.2.** | **Scope of the Work** | 17 |
|  | **Summary** | 18 |
|  |  |  |
| **Chapter 3.** | ***Methodology*** | **19 – 29** |
|  |  |  |
| **3.1.** | **Introduction** | 19 |
| **3.2.** | **KNN-Classification** | 20 |
| **3.3.** | **Logistic Regression** | 22 |
| **3.4.**  **3.5.**  **3.6.** | **Naïve Bayes Classifier**  **Decision Tree Classification**  **Neural Network** | 25 |
|  | **Summary** | 29 |
|  |  |  |
| **Chapter 4.** | ***Experimental Results*** | **30 – 42** |
|  |  |  |
| **4.1.** | **Introduction** | 30 |
| **4.2.** | **Dataset Exploration** | 31 |
|  | 4.2.1 Dataset Introduction | 31 |
|  | 4.2.2 Dataset Overview | 33 |
| **4.3.**  **4.4.**  **4.5.** | **Basic Data Analysis**  **Model Analysis**  4.4.1 KNN Classification  4.4.2 Logistic Regression  4.4.3 Naïve Bayes Classifier  4.4.4 Decision Tree Classification  4.4.5 Neural Network  **Performance Analysis**  **Summary** | 42 |
| **Chapter 5.** | ***Conclusion and Future Scope*** | **43 – 45** |
|  |  |  |
| **5.1.** | **Conclusion** | 44 |
| **5.2.** | **Future Scope** | 45 |
|  |  |  |
|  | **References** | **46 – 51** |
|  |  |  |
|  | **Appendix – A: List of Formulas Used** | I |
|  |  |  |

**CHAPTER 1**

**INTRODUCTION**

In recent years, there has been a growing recognition of the significant impact that stress can have on the academic performance, mental health, and overall well-being of students. The modern educational landscape is fraught with various stressors, ranging from academic pressures and social expectations to personal challenges and extracurricular commitments. As a result, there is an urgent need for effective strategies to identify, assess, and mitigate student stress factors in order to promote a supportive and conducive learning environment.

This project focuses on addressing this pressing need through the development of a sophisticated system for student stress factor detection using classification, feature selection, and optimization techniques. By harnessing the power of machine learning and data-driven methodologies, this project aims to provide educators and stakeholders with valuable insights into the underlying stressors affecting students, enabling targeted interventions and support mechanisms.

* 1. **BACKGROUND**

The prevalence of student stress in educational institutions has become a pervasive concern in recent years. Research studies have highlighted the detrimental effects of stress on student academic performance, mental health, and overall quality of life. High levels of stress can lead to increased anxiety, depression, and burnout, ultimately impairing students' ability to succeed academically and thrive in their personal lives.

Traditionally, efforts to address student stress have relied on subjective assessments, such as self-reported surveys and anecdotal observations. While these methods provide valuable insights, they often lack the scalability and precision needed to effectively identify and intervene in cases of student stress. Moreover, the complex interplay of factors contributing to student stress necessitates a more nuanced and data-driven approach to detection and intervention.

Advancements in machine learning and data analytics present a promising opportunity to overcome these challenges. By leveraging the vast amounts of data generated within educational settings, such as academic performance records, behavioral metrics, and demographic information, machine learning algorithms can uncover hidden patterns and relationships indicative of student stress factors. Furthermore, by incorporating feature selection and optimization techniques, these algorithms can enhance their predictive accuracy and efficiency, enabling more targeted and effective interventions.

* 1. **PROBLEM STATEMENT**

The prevalence of student stress is undeniable, with profound effects on academic performance and overall well-being. Despite the essential role education plays in shaping futures, student stress has become a pressing concern. Current stress detection methods fall short in terms of efficiency and personalization, presenting a pressing need for a more robust model capable of identifying stress factors across academic, personal, and external domains. This necessity is further intensified by the recent surge in student stress due to the pandemic, highlighting the urgency of incorporating new data to ensure the model's relevance and accuracy in addressing the current situation.

Moreover, existing stress detection models face limitations such as inadequate performance metrics or a lack of interactivity. Overcoming these challenges demands innovative solutions, including advanced machine learning techniques and the development of interactive interfaces. Our problem statement is to explore, analyze, and comprehend the impact of psychological, physiological, social, environmental, and academic factors on student well-being to guide wellness approaches and treatments that encourage a more supportive and healthy learning environment. By addressing these shortcomings and unlocking meaningful insights from our dataset, we can create a comprehensive stress detection model that accurately mirrors real-world scenarios and provides personalized feedback, ultimately empowering students to better manage their stress levels and enhance the overall student experience.

* 1. **OBJECTIVE AND MOTIVATION**

The primary objective of this project is to develop a comprehensive system for student stress factor detection that leverages classification, feature selection, and optimization techniques. Specifically, the project aims to:

1. Utilize machine-learning algorithms to classify and categorize different types of student stress factors based on multi-dimensional data sources.

2. Employ feature selection techniques to identify the most relevant and informative variables for stress detection, thereby improving model performance and interpretability.

3. Apply optimization techniques to enhance the efficiency and scalability of the stress detection system, enabling real-time monitoring and intervention.

By achieving these objectives, this project seeks to provide educators, administrators, and mental health professionals with a powerful tool for identifying and addressing student stress factors proactively, ultimately contributing to the overall well-being and success of students in educational settings.

## PROPOSED METHOD

* + 1. **Dataset Enhancement through Collection and Integration**

To bolster the efficacy of our stress detection model, we embark on a proactive approach of dataset enhancement. Recognizing the dynamic nature of stressors affecting students, we engage in systematic data collection from diverse sources, including academic institutions, online forums, and surveys. By harnessing the power of platforms like Kaggle, we curate a comprehensive dataset that encapsulates the multifaceted nature of student stress across academic, personal, and external domains. Through meticulous curation and integration of additional samples, we ensure the dataset's diversity and relevance, thereby enriching the model's ability to accurately identify and quantify stress factors. This strategic enhancement reinforces the project's real-world applicability, laying a robust foundation for our endeavor to develop a nuanced and effective stress detection model.

* + 1. **Creation of Multiple Models for Comparative Analysis**

In order to develop a robust stress detection model, we propose the creation of multiple machine learning models encompassing a variety of algorithms such as K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, Neural Networks, and Decision Trees. By employing diverse modeling techniques, we aim to comprehensively explore the complex relationships within the data and capture various manifestations of stress across academic, personal, and external domains. Each model will undergo rigorous training, validation, and testing phases to assess its performance in accurately identifying and quantifying stress factors. Through comparative analysis of these models based on various evaluation metrics, including accuracy, precision, recall, and F1-score, we seek to identify the most effective approach for stress detection. This methodological approach ensures a thorough examination of different modeling strategies, ultimately facilitating the selection of the optimal model that exhibits superior performance in addressing the multifaceted nature of student stress.

* + 1. **Creation of User Interactive Interface**

The cornerstone of our project lies in the development of a user interactive interface tailored specifically to empower students in managing their stress effectively. This interface serves as a platform where students can input relevant information about their academic, personal, and external stressors through a series of user-friendly prompts and questions. Leveraging the insights gained from advanced machine learning techniques such as classification and feature selection, the interface precisely identifies and quantifies stress factors unique to each student's situation. Moreover, it goes beyond mere identification by providing personalized suggestions and measures to alleviate stress based on the individual's responses. By integrating a feedback mechanism, the interface ensures a continuous loop of improvement, adapting to the evolving needs and experiences of the students. Through this pioneering stride, we aim to empower students with the tools and resources they need to proactively manage their stress and optimize their academic and personal well-being.

* 1. **PROJECT ORGANIZATION**

**Chapter 1,** "*Introduction*", emphasizes the critical importance of addressing student stress in academia. It introduces the project's focus on developing a sophisticated system for stress detection using machine-learning techniques to enable targeted interventions and support.

**Chapter 2,** titled "*Literature Review*", examines the array of stressors affecting students in academia today, emphasizing the necessity for interventions to mitigate their impact on academic performance and mental health. It synthesizes previous studies to identify common stressors and detection methods, offering insights for practical application and future research.

**Chapter 3** named as, *“Proposed Method”*this introduces the implementation of machine learning techniques for stress detection. The methods utilized include K-Nearest Neighbors (KNN), Logistic Regression, Naïve Bayes Classifier, Decision Tree Classification, and Neural Network. Each algorithm is chosen for its ability to learn from the dataset and enhance the model's performance in identifying stress factors across academic, personal, and external domains.

**Chapter 4** titled,*“Experimental Results”,* displays the outcomes of implementing machine-learning algorithms in the stress detection model. It highlights the performance metrics such as accuracy, precision, recall, and F1-score attained by each algorithm. Through a concise analysis, the chapter evaluates the effectiveness of the methods in identifying and quantifying stress factors, offering insights into the model's performance and potential enhancements.

**Chapter 5** titled, *“Conclusion”*, signifies the culmination of our project. Through advanced machine learning techniques, we have successfully quantified stress factors across academic, personal, and external domains. Leveraging an enriched dataset from Kaggle and an interactive feedback interface, we have empowered students in stress management. This work lays a foundation for ongoing research in student well-being and highlights the potential of technology in addressing complex issues like student stress.

**SUMMARY**

In this chapter, we discussed:

* The significance of addressing student stress in academia and the urgent need for effective strategies to mitigate its impact.
* The prevalence of student stress and its detrimental effects on academic performance, mental health, and overall well-being.
* The limitations of traditional stress detection methods and the potential of machine learning and data-driven approaches to overcome these challenges.
* The project's objectives to develop a comprehensive system for student stress factor detection using classification, feature selection, and optimization techniques.
* The proposed methods including dataset enhancement through collection and integration, creation of multiple machine learning models for comparative analysis, and development of a user interactive interface.
* The organization of the project, including chapters on introduction, literature review, proposed methodology, experimental results, and conclusion.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 Introduction**

In today’s fast-paced and demanding academic environment, students often face a myriad of stressors that can significantly affect their academic performance, mental health, and overall well-being. From academic pressures to financial concerns, social relationships, and time management challenges, the sources of stress for students are diverse and multifaceted. Left unaddressed, chronic stress can lead to a range of negative outcomes, including burnout, anxiety, depression, and even academic failure. Therefore, there is a critical need to identify and understand the factors contributing to student stress in order to develop effective interventions and support strategies.

By synthesizing the findings of previous studies, this review aims to identify common stressors experienced by students, examine the various methods and technologies used to detect stress, and explore the implications of this research for practice and future research directions.

Factors contributing to student stress encompass a wide spectrum of challenges spanning academic, financial, social, and personal domains. Academic pressure, characterized by high expectations, heavy workloads, and performance anxiety, often weighs heavily on students striving for academic success. Financial concerns, including tuition fees, living expenses, and student debt, can add another layer of stress, particularly for financially independent students. Social relationships present their own set of stressors, such as peer pressure and conflicts, while time management challenges can arise from balancing various responsibilities and combating procrastination. Health and wellness issues, both physical and mental, can exacerbate stress levels, as can environmental stressors like noise, distractions, and safety concerns. Personal factors such as perfectionism and coping strategies also play a role in shaping individual responses to stress. Recognizing and addressing these multifaceted stressors is crucial for fostering a supportive learning environment that promotes student well-being and academic success.

**2.2 Scope of the Work**

The project's scope encompasses a multifaceted approach to addressing student stress, leveraging advanced machine learning techniques and innovative strategies. By focusing on the precise identification and quantification of stress factors across academic, personal, and external domains, the project aims to provide a comprehensive understanding of the complex nature of student stress. Through the incorporation of additional samples from Kaggle and meticulous dataset enhancement, the project ensures diversity and relevance, enhancing the model's accuracy and real-world applicability. Furthermore, the introduction of an interactive interface with a feedback mechanism represents a pioneering stride in personalized stress detection, empowering students to actively engage in stress management.

Moving forward, the project identifies two key areas for further progress: updating the dataset to reflect current trends and implementing a feedback mechanism alongside an interactive user interface. Continuous data collection will be vital for dataset enhancement, ensuring that the model remains relevant and effective in addressing evolving stressors. Additionally, the development of a user-friendly interface will enhance the project's practical applicability, facilitating seamless interaction and feedback from students. By prioritizing innovation and practicality, the project aims to continually advance in its mission to mitigate student stress and promote academic success and well-being.

**SUMMARY**

* Explored various stressors faced by students in academic environments.
* Synthesized previous studies to understand the multifaceted nature of student stress.
* Outlined the project's scope, focusing on leveraging advanced machine learning techniques for stress detection.
* Emphasized the importance of dataset diversity and the development of an interactive interface.
* Identified areas for further progress, including dataset updates and implementation of a feedback mechanism for student empowerment.

**CHAPTER 3**

**METHODOLOGY**

**3.1 Introduction**

Machine Learning is a subset of artificial intelligence that is mainly concerned with the development of algorithms that allow a computer to learn from the data and past experiences on their own. Machine Learning is the study of algorithms that can improve the performance of doing certain task.0-

**3.2 KNN – Classification**

K-Nearest Neighbors (KNN) is a popular non-parametric machine-learning algorithm used for classification tasks. It relies on proximity to labeled data points for predictions. By calculating distances between new and existing data points, KNN identifies the K closest neighbors to determine the class of the new data point. While effective for small datasets and nonlinear boundaries, KNN may be computationally expensive with larger datasets and requires careful hyper parameter tuning to prevent overfitting.

**3.3 Logistic Regression**

Logistic Regression is a type of supervised machine learning algorithm used for classification tasks. The goal of logistic regression is to model the probability of a binary outcome (e.g. 0 or 1) based on one or more independent variables. Logistic regression uses the logistic function (also known as the sigmoid function) to map the linear combination of the independent variables to a value between 0 and 1, which represents the predicted probability of the positive class. The logistic function is defined as the exponential of the linear combination divided by one plus the exponential of the linear combination. Logistic regression estimates the coefficients of the linear equation that best describes the relationship between the independent variables and the dependent variable, using maximum likelihood estimation. Once the coefficients are estimated, they can be used to predict the class of new data points by thresholding the predicted probabilities. Logistic regression is a simple yet powerful algorithm that can work well with both binary and multi-class classification problems, but it assumes that the relationship between the independent variables and the dependent variable is linear, which may not always be the case.

**3.4 Naïve Bayes Classifier**

Naive Bayes is a type of supervised machine learning algorithm used for classification tasks. It is based on Bayes’ theorem, which states that the probability of a hypothesis (or class) given the data is proportional to the product of the probability of the data given the hypothesis and the prior probability of the hypothesis. Naive Bayes assumes that the independent variables are conditionally independent given the class, which means that the presence or absence of one feature does not affect the probability of the presence or absence of another feature. This assumption simplifies the calculation of the probability of the data given the hypothesis and allows the algorithm to make predictions with high accuracy and speed. Naive Bayes can handle both binary and multi-class classification problems and can handle text classification and sentiment analysis tasks. Naive Bayes is particularly useful when the number of features is large and the interactions between the features are not important. However, Naive Bayes can suffer from the problem of zero-frequency, which occurs when a feature does not occur in the training data for a particular class, and it can be sensitive to the choice of the prior probabilities.

**3.5 DecisionTree Classification**

Decision Tree Classification is a type of supervised machine learning algorithm used for classification tasks. The goal of decision tree classification is to model the relationship between the independent variables and the dependent variable by recursively partitioning the data into smaller subsets based on the values of the independent variables. The partitioning is done in a way that maximizes the reduction of the impurity (or entropy) of the dependent variable within each subset. A node in the decision tree represents each partition, and the leaf nodes represent the final predictions. Decision tree classification can handle both binary and multi-class classification problems, and it is particularly useful when there are non-linear relationships between the independent and dependent variables or when the interactions between the variables are important. Decision tree classification can also handle missing data and outliers in the dataset.

**3.6 Neural Network**

Neural network classification is a potent tool in the realm of machine learning, adept at handling intricate relationships between input features and target variables. Unlike traditional algorithms, neural networks employ interconnected layers of neurons, each layer refining the data representation as it passes through. Beginning with an input layer that receives the dataset's features, the network then progresses through hidden layers where complex transformations occur, guided by activation functions like ReLU or sigmoid. These transformations allow the network to capture nuanced patterns in the data, crucial for tasks with non-linear relationships. Finally, the output layer generates predictions, often utilizing activation functions like softmax for multi-class classification. Throughout training, the network fine-tunes its parameters to minimize a predefined loss function, optimizing its ability to make accurate predictions. Neural network classification excels in scenarios with complex datasets, offering robust performance and adaptability to various classification tasks.

**SUMMARY**

* Introduction to Machine Learning: ML develops algorithms for computers to learn independently from data and past experiences, enhancing performance.
* KNN – Classification: KNN is a non-parametric algorithm for classification, relying on proximity to labeled data points. Effective for small datasets and nonlinear boundaries but computationally expensive and requires tuning.
* Logistic Regression: Logistic Regression models binary outcomes based on independent variables. Assumes a linear relationship, simple yet powerful.
* Naïve Bayes Classifier: Naive Bayes is based on Bayes' theorem, assuming feature independence given the class. Efficient for large datasets and text classification but may face zero-frequency problems.
* Decision Tree Classification: Decision Tree partitions data into subsets to maximize impurity reduction. Handles non-linear relationships, missing data, and outliers.
* Neural Network: Neural networks use interconnected layers to capture complex data relationships. Excelling in capturing patterns, they offer robust performance across various tasks.

**CHAPTER 4**

**EXPERIMENT RESULTS**

**4.1 Introduction**

In this chapter, we focus on experimental analysis to delve deeper into the issue of student stress in educational contexts. Recognizing the significant impact of stress on academic performance and well-being, our aim is to dissect the various factors contributing to student stress through rigorous experimentation.

We seek to uncover the intricate relationships between psychological, physiological, social, environmental, and academic factors, aiming to derive actionable insights to inform interventions and support systems. Despite the inherent challenges, our commitment to advancing understanding and alleviating student stress remains unwavering.

Through collaborative efforts and interdisciplinary engagement, we endeavor to contribute to the ongoing discourse on student well-being, ultimately striving to foster a more nurturing and conducive learning environment. Join us as we embark on this empirical journey towards a resilient and flourishing student community.

**4.2 Dataset Exploration**

**4.2.1 Dataset Introduction**

The dataset titled "Student Stress Factors: A Comprehensive Analysis" is now available on Kaggle. This dataset offers valuable insights into the underlying causes and consequences of student stress, providing researchers and educators with a wealth of information. By collecting data from students, we have significantly expanded the dataset, adding approximately 308 new data points. This nearly 28% increase in dataset size enhances its relevance to the present scenario, allowing for a more comprehensive understanding of the variables contributing to student stress and their interactions.

**4.2.2 Dataset Overview**

The dataset comprises a comprehensive collection of data points representing diverse factors influencing student stress. With approximately 20 carefully selected features, the dataset highlights the most significant contributors to a student's stress levels. These features are chosen methodically, considering five primary factors: psychological, physiological, social, environmental, and academic. Each factor encompasses a range of variables, offering a holistic understanding of students' experiences and the stressors they encounter.

The dataset includes variables representing different factors contributing to student stress:

|  |  |
| --- | --- |
| Psychological Factors | anxiety\_level, self\_esteem, mental\_health\_history, depression |
| Physiological Factors | headache, blood\_pressure, sleep\_quality, breathing\_problem |
| Environmental Factors | noise\_level, living\_conditions, safety, basic\_needs |
| Academic Factors | academic performance, study load, teacher\_student\_relationship, future\_career\_concerns |
| Social Factors | social\_support, peer\_pressure, extracurricular\_activities, bullying |

Table 4.1: Different factors contributing to student stress

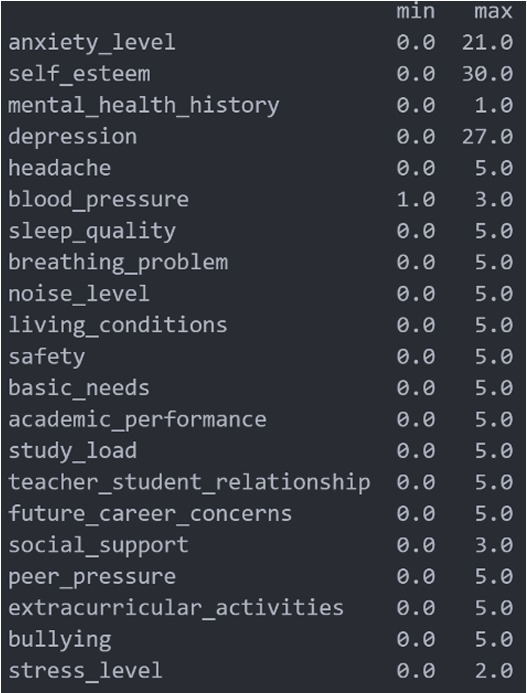


Figure 4.1: Stress Factor Dataset Features

These variables offer insights into students' surroundings, physical and mental health, academic circumstances, and social interactions, collectively painting a comprehensive picture of potential stressors. Analyzing these factors can enhance our understanding of student stress and its complex interactions.

**4.3 Basic Data Analysis**

Basic data analysis entails examining data to reveal insights and patterns. This process encompasses tasks such as cleansing data for accuracy, summarizing key features with descriptive statistics, and presenting data visually using charts and graphs. Exploratory data analysis aids in understanding relationships among variables, while inferential statistics enable making inferences about a population from sample data. Additionally, utilizing Pearson correlation allows for understanding the relationship between features and stress levels. In essence, basic data analysis establishes a groundwork for deeper investigation and informed decision-making.

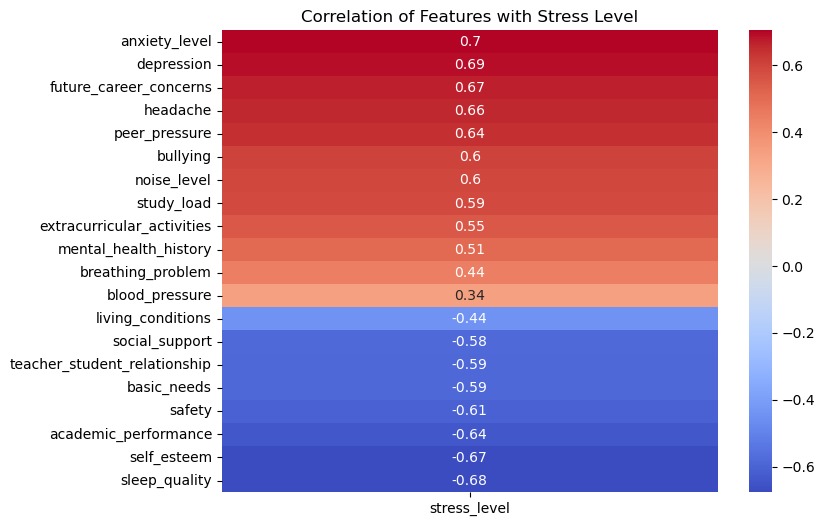
****

Figure 4.2:Correlation of features with stress level

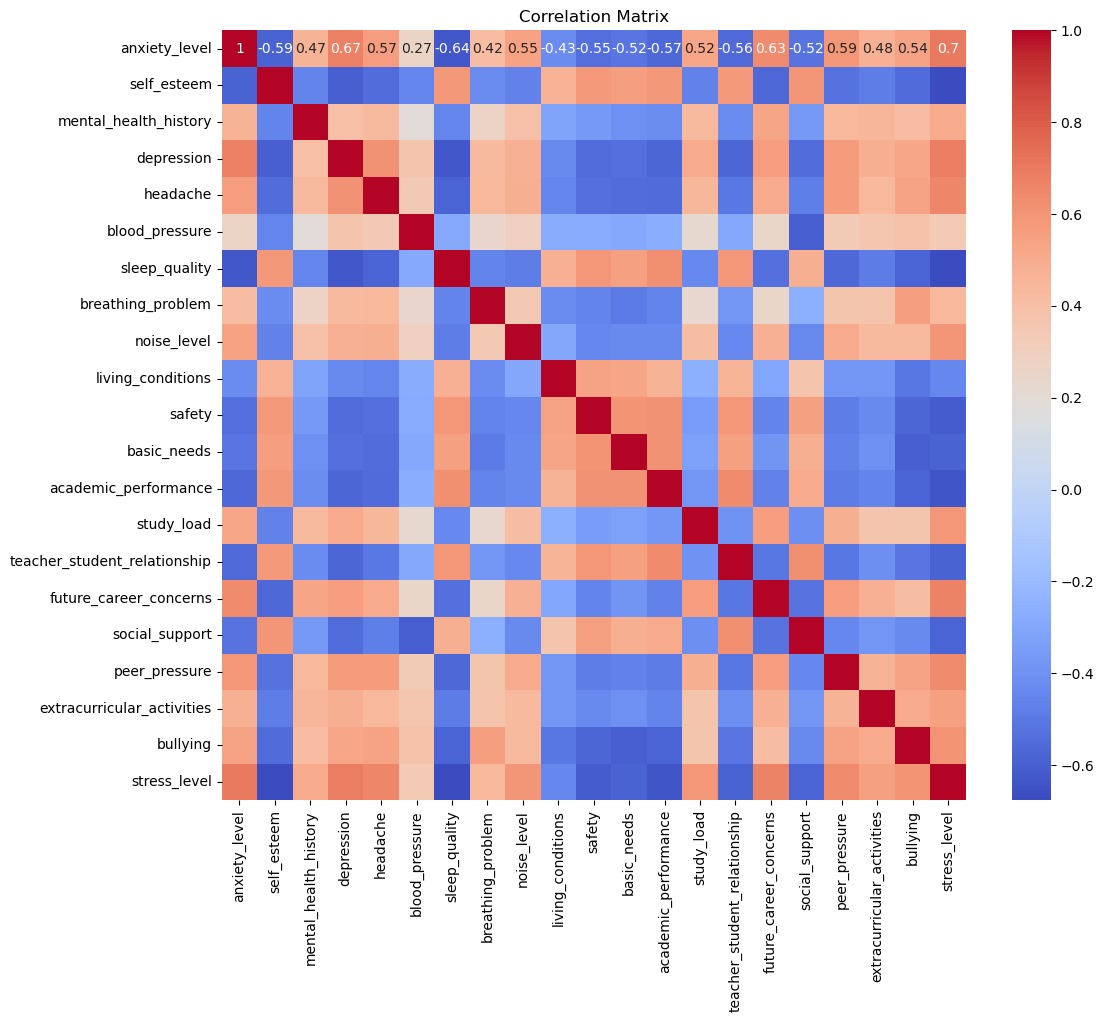


Figure 4.3: Correlation Matrix

**4.4 Model Analysis**

**4.4.1 KNN – Classification**

KNN (K-Nearest Neighbors) classification was utilized in this research paper because of its simplicity and flexibility. KNN is a non-parametric algorithm that does not make any assumptions about the underlying data distribution and can handle both continuous and categorical data.

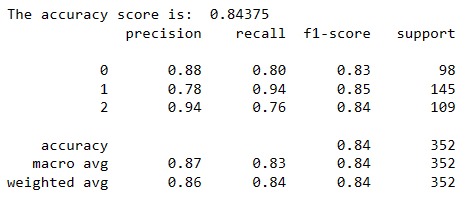


Figure 4.4: Classification report of KNN Classification

**4.4.2 Logistic Regression**

Logistic regression is a popular and widely used algorithm for binary classification problems, where the goal is to predict one of two possible outcomes. In this research paper, logistic regression was chosen for the binary classification task due to its simplicity, interpretability, and effectiveness in handling linearly separable datasets. Logistic regression estimates the probability of the positive class as a function of the input variables, using a sigmoid function that outputs values between 0 and 1 predicting whether or not the person has Parkinson.

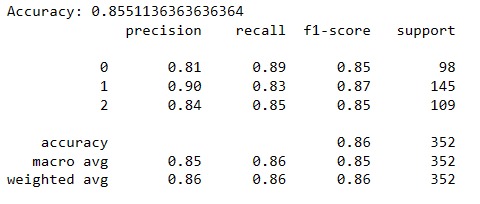


Figure 4.5: Classification report of Logistic Regression

**4.4.3 Naïve Bayes Classifier**

Naive Bayes classification was used due to its simplicity and effectiveness in handling high-dimensional datasets. Naive Bayes is a probabilistic algorithm that is based on the Bayes theorem and the assumption of conditional independence between features. This assumption simplifies the probability calculation by allowing the algorithm to estimate the probability of each feature independently, given the class variable.

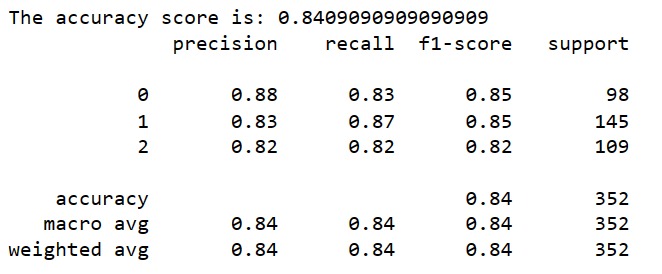
****

Figure 4.6: Classification report of Naïve Bayes Classifier

**4.4.4 Decision Tree Classification**

Decision Tree Classification models the relationship between independent and dependent variables by recursively partitioning data into subsets, aiming to maximize impurity reduction. It is versatile, handling both binary and multi-class classification and suitable for non-linear relationships and interactions between variables. Additionally, it can manage missing data and outliers effectively.

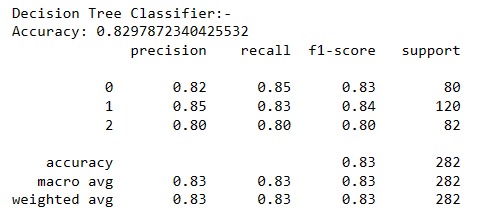


Figure 4.7: Classification report of Decision Tree Classification

**4.4.5 Neural Network**

The neural network architecture is defined with an input layer, two hidden layers employing ReLU activation functions, and an output layer utilizing softmax activation for multiclass classification. Following model compilation with the Adam optimizer and categorical cross-entropy loss function, training commences on the training data. Once trained, the model predicts probabilities for the test data. True labels are binarized for evaluation, and accuracy is computed by comparing predicted labels to true labels. Ultimately, the code prints the accuracy score, providing a measure of the model's performance in classifying unseen data.

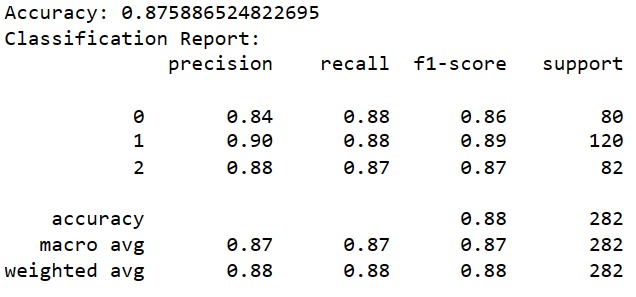


Figure 4.8: Classification report of Neural Network

**4.5 Performance Analysis**

Based on the results obtained from applying various algorithms - Naive Bayes, Decision Tree, KNN, Logistic Regression, and Neural Network - to the dataset, it is evident that each algorithm yields differing levels of accuracy. Naive Bayes and Decision Tree classifiers achieved accuracy scores of 84.09% and 82.97%, respectively, while KNN and Logistic Regression classifiers performed slightly better with accuracy scores of 84.37% and 85.51%, respectively. Notably, the Neural Network model achieved the highest accuracy score of 87.58% among all the algorithms tested.

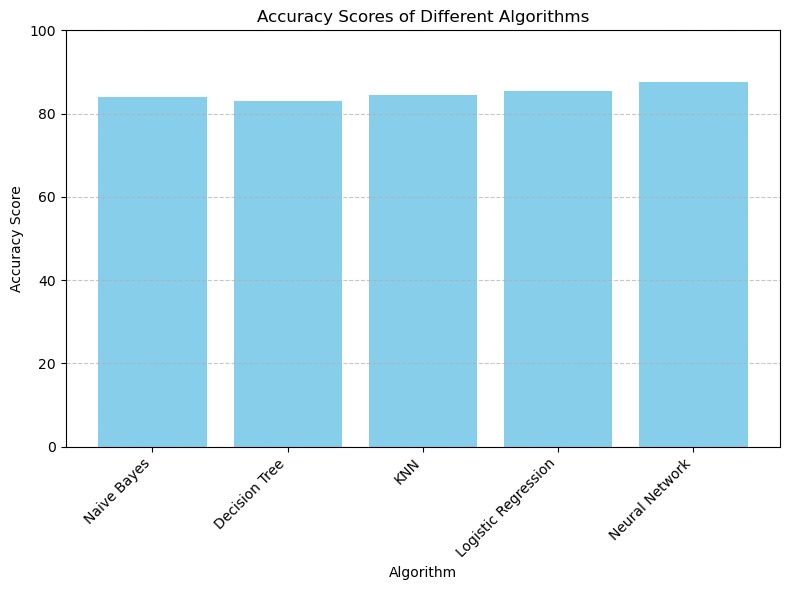


Figure 4.9: Accuracy Scores of Different Algorithm

This comparison highlights the varying effectiveness of different algorithms in accurately predicting the target variable in the dataset. While Naive Bayes and Decision Tree classifiers exhibit decent performance, KNN and Logistic Regression classifiers demonstrate slightly higher accuracy. However, the Neural Network model stands out as the most accurate among all the algorithms tested.

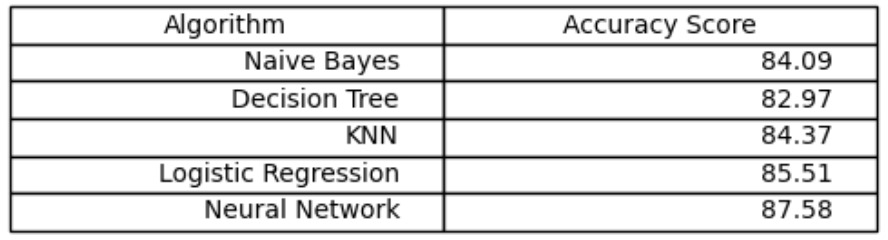


Table 4.2: Comparing the different models

As the Neural Network emerges as the top-performing model among the algorithms tested, it underscores its efficacy in accurately predicting student stress levels. In terms of feature importance, insights from Figure 10 illustrate that mental\_health\_history, blood\_pressure, and anxiety\_level are identified as the most influential features, to the predictive accuracy of the model. This underscores the critical role of mental health indicators and physiological factors in understanding and predicting student stress. Additionally, social support emerges as another key determinant, in the model. These findings underscore the holistic approachrequired in addressing student stress, highlighting the importance of considering both psychological well-being and social dynamics in educational settings.

We developed a user interface utilizing HTML and CSS for the front-end design, coupled with Flask for the back-end functionality. Flask provides a web framework that allows us to integrate Python code seamlessly into the application. We trained and saved a neural network model using Python's scikit-learn library, and then serialized it into a pickle file. This pickle file serves as a compact representation of the trained model. In our Flask application, we load this pickle file and use it to make predictions based on user input. The input from the user interface is passed to the Flask server, where it undergoes any necessary preprocessing before being fed into the neural network model for prediction. The predicted results are then sent back to the user interface for display. This setup enables us to create an interactive web application where users can input their data and receive predictions from the neural network model in real-time

**SUMMARY**

* The dataset "Student Stress Factors: A Comprehensive Analysis" on Kaggle offers insights into causes and consequences of student stress.
* Approximately 308 new data points were added, enhancing dataset relevance by nearly 28%.
* Dataset features are categorized into psychological, physiological, social, environmental, and academic factors.
* Basic data analysis involves cleaning data, descriptive statistics, and visualizations.
* Model analysis includes KNN, Logistic Regression, Naïve Bayes Classifier, Decision Tree Classification, and Neural Network.
* Neural Network emerges as the top-performing model with an accuracy score of 87.58%.
* Feature importance highlights mental\_health\_history, blood\_pressure, anxiety\_level, and social support as influential factors in predicting student stress levels.
* Overall, the findings emphasize the importance of addressing student stress holistically, considering psychological well-being and social dynamics in educational settings.

**CHAPTER 5**

**CONCLUSION AND FUTURE SCOPE**

**5.1 Conclusion**

The conclusion of this project highlights the comprehensive evaluation of key factors contributing to student stress and the application of different machine learning techniques to analyze and predict stress levels. By employing methods such as KNN, Logistic Regression, Naive Bayes, and Neural Network, the project demonstrated the ability to classify students into stress level categories based on their features. Additionally, Decision Trees and Random Forest approaches were effective in identifying influential factors on students' stress levels. Overall, the project provided valuable insights into student stress, its prediction, and the factors influencing it.

The outcomes of the study underscore the importance of understanding these factors for developing targeted stress management interventions. Future research could delve deeper into the causal mechanisms of these relationships and explore interventions aimed at increasing social support and academic performance to boost students' self-esteem and reduce stress levels.

Furthermore, the project emphasizes the complex interplay of various factors contributing to student stress levels, highlighting the need for an integrated approach to address student stress. This approach should encompass not only academic factors but also social support and self-esteem, reflecting the multifaceted nature of student well-being.

**5.2 Future Scope**

In future iterations of the project, enhancing effectiveness and improving capabilities can be achieved through various avenues. Firstly, integrating real-time data analytics can enable prompt responses to student stress by incorporating data streams from mood tracking apps or wearable devices, providing instant feedback and personalized coping strategies. This proactive approach ensures timely intervention and support, fostering a responsive and supportive environment. Additionally, personalized recommendations can tailor stress management techniques based on individual profiles, including academic performance and personal interests, acknowledging diverse experiences and coping mechanisms. Exploring sentiment analysis techniques, particularly in analyzing text data from student essays or social media, offers insights into emotional well-being, facilitating targeted interventions through advanced natural language processing algorithms. These enhancements promise to create a more adaptive and effective system for supporting student well-being.

With additional time, a more extensive exploration of hyper parameter tuning for all models could provide valuable insights into their performance. Experimenting with a broader range of parameters, particularly considering the success of Decision Trees and Random Forests opens avenues for exploring additional ensemble methods like AdaBoost or Gradient Boosting to enhance predictive accuracy. Furthermore, while the current research questions primarily focus on predicting stress levels and identifying significant contributing factors, further investigation could delve into identifying combinations of features from different domains that collectively have a greater influence on predicting stress levels. Additionally, exploring variations or ensemble approaches for the models, such as bagging or boosting techniques for Decision Trees and Random Forests, different distance metrics for KNN, and varied architectures for Multi-Layer Perceptron, presents promising directions for future research to deepen our understanding and improve the effectiveness of stress prediction models.

**REFERENCES**

[1] P. Manjunath, Twinkle S, P. Shreya, V. Ashok and Dr. S. Sultana. “Predictive Analysis of Student Stress Level using Machine Learning”. **2021**.

[2] Qicheng Chen and Boon Giin Lee. “Deep Learning Models for Stress Analysis in University Students”. **2023**.

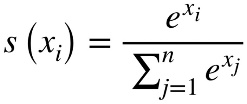
[3] Wu J, Kuan G, Lou H, Hu X, Masri MN, Sabo A, Kueh YC. “The impact of COVID-19 on students' anxiety and its clarification: a systematic review.” Front Psychol. **2023**.

[4] Fred Johansson, Ann Rudman, and Margreth Grotle. “Depression, anxiety and stress among Swedish university students before and during six months of the COVID-19 pandemic”. **2021.**

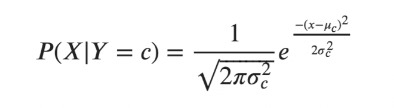
[5] https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis

**APPENDIX – A**

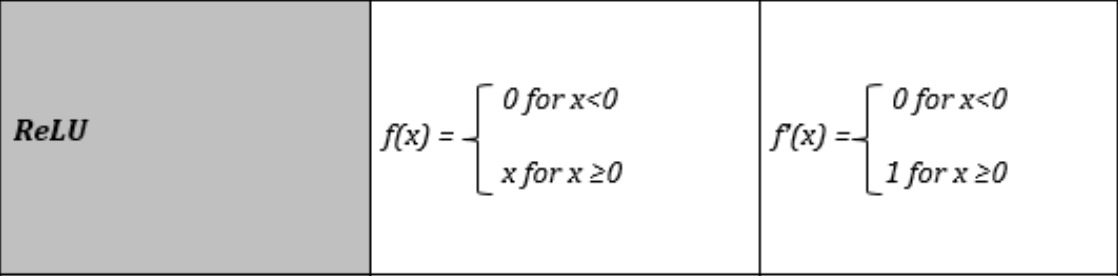
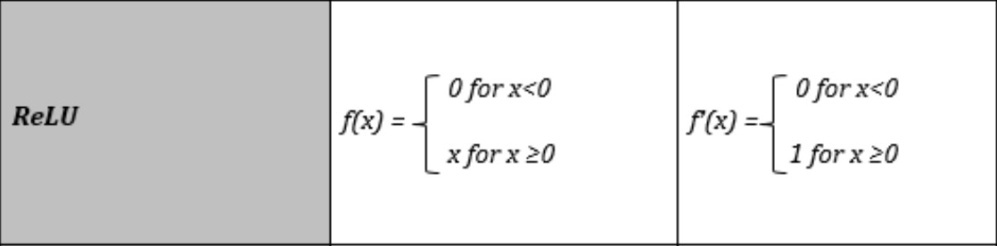
**LIST OF FORMULAS USED**



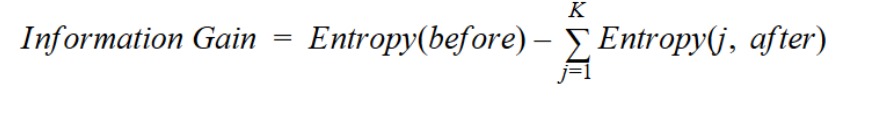
1. Softmax function:



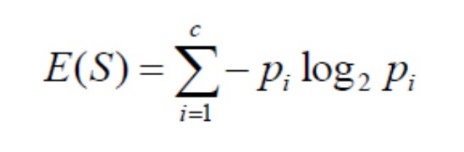
2. Naïve Bayes :



1. ReLU function :



1. Decision Tree :





Department of Computer Science and Engineering

Silicon University, Odisha

Silicon Hills

Bhubaneswar **–**751 024

Odisha, India