CH1.

Introduction

Background of the Study

Technology has developed significantly in the past few years, influencing how businesses function and how they make decisions in the digital world. These advances in technology enabled the collection, storage, and processing of immense amounts of data across multiple sectors. According to the United Nations, Digital technologies have advanced more rapidly than any innovation in our history – reaching around 50 per cent of the developing world's population in only two decades and transforming societies. By enhancing connectivity, financial inclusion, access to trade and public services, technology can be a great equaliser.

In today's digital age, Data analytics has emerged as an essential decision-making tool in a variety of fields, including organization, business, government, education, and healthcare. Organizations utilize data analysis to optimize performance, establish effective policies, and offer better services. Businesses apply analytics to better understand how consumers behave and manage operations, whereas governments use data-driven insights to improve services to the public and distribute resources more effectively. In the healthcare sector, for instance, AI-enabled frontier technologies are helping to save lives, diagnose diseases and extend life expectancy. In education, virtual learning environments and distance learning have opened up programmes to students who would otherwise be excluded. Public services are also becoming more accessible and accountable through blockchain-powered systems, and less bureaucratically burdensome as a result of AI assistance. Big data can also support more responsive and accurate policies and programmes.

Statement of the Problem

Depression among students has become a major concern, yet educational institutions often lack methods that use data for identifying and understanding the factors that contribute to students' mental health issues. While various lifestyle, academic, and demographic factors are believed to contribute to depression, there is limited empirical study that utilizes all these factors to determine their overall effect on students' well-being.

Using the dataset we provided, this study approaches the analytical problem of identifying relevant variables and patterns connected with depression amongst students using exploratory data analysis and predictive modeling.

Objectives of the Study

The research being undertaken aims to address the analytical and research context challenges identified in the study. In a specific manner, this study aims to:

- This study aims to explore and analyze the Student Depression Dataset using exploratory data analysis (EDA) to unfold underlying trends, patterns, and relationships among lifestyle, academic, and demographic variables that are associated with student depression.
- The study aims to quantify depression as a categorical variable by utilizing measurable indicators derived from the dataset to successfully provide a data-driven assessment of students' mental health status.
- This study aims to clearly identify and determine the most common and influential factors contributing to depression amongst students while acknowledging the complexity arising from the interaction of multiple variables in the dataset.
- This study aims to apply the k-Nearest Neighbors (kNN) algorithm as a predictive modelling technique to easily classify students according to their depression status through similarity-based learning.
- This study aims to assess the effectiveness of kNN as a low-cost and lazy learning algorithm in supporting early recognition of depression-related patterns using appropriate classification metrics within a global and non-local student context.

In simple terms, the study mainly intends to demonstrate the applicability of data-driven analytical methods in enhancing understanding of student mental health issues in educational settings.

Scope and Delimitation

This particular study focuses on the analysis of the Student Depression Dataset derived from Kaggle and is specifically limited to the examination of demographic, academic, and lifestyle-related variables that are anchored on student depression. The study encompasses exploratory data analysis and supervised classification using the k-Nearest Neighbors (kNN) algorithm to identify patterns and classify students based on their depression status. Moreover, the research study is circumscribed to the use of a single secondary dataset and does not involve any form of primary data collection, clinical diagnosis, or real-time monitoring of student mental health.

Additionally, the study does not compare kNN with other machine learning algorithms, nor does it attempt to generalize findings beyond characteristics of the provided dataset. Ultimately, the results are intended to support analytical understanding rather than provide definitive medical or psychological conclusions.

Significance of the Study

This study focuses on the application of k-Nearest Neighbors (kNN) algorithm in identifying and analyzing factors that contribute to depression amongst students using the Student Depression Dataset from Kaggle. Moreover, this study aims to demonstrate how machine learning techniques can provide data-driven insights into mental health trends and early identification of at-risk students.

In a specific manner, this study will be beneficial to the following:

- The Information Technology (IT) Practitioners and Students

  The IT practitioners and students will gain practical knowledge regarding the implementation of machine learning algorithms, particularly k-Nearest Neighbors, in analyzing real-world datasets, like Student Depression Dataset. In addition, this study helps these individuals to better understand how to utilize data-driven analytical methods to extract meaningful insights without being limited by theoretical concepts alone.

- Business and Educational Institutions

  The business and educational institutions will benefit by comprehending how predictive analytics can help identify students at risk of depression and to encourage their development sectors to build more effective support programs. Moreover, this study helps them to make evidence-based decisions that optimize their resources while improving student well-being.

- Society

  The society will gain a vital awareness regarding the prevalence and factors that contribute to students' cause of depression. In addition, this study pushes them to better recognize how data-driven approaches can support early interventions and promote mental health awareness, ultimately contributing to societal well-being.

Review of related Literature

I. Overview of k-Nearest Neighbor in Machine Learning Literature

In the field of machine learning, the algorithm k-Nearest Neighbor (KNN) can still be considered as one of the most popular and widely used classification tools. This is because of its adaptability, flexibility, and effectiveness in various domains. This literature review looks into several academic publications released between 2021 and 2025. Focusing on the analytical principles of KNN, its theoretical strengths and limitations, key trends and enhancements, real-world application of KNN, comparative analysis with other algorithms, and challenges, as well as the research gaps within the academic papers.

In the year of 1951, KNN was actually first introduced by Fix and Hodges in the field of statistics as a non-parametric discrimination method to address the ongoing overfitting problem of the original NN (nearest neighbor) classification approach (Zhang, 2021). The proposed solution was to use several K nearest neighbors within the new/unknown data point instead of using a single data point and then gather all the classification of that neighborhood and label the new data based on the majority rule. The author also noted that when applying KNN to very mild conditions the algorithm's error rate tends to lean more on the Bayes optimality theoretical concept, which makes the algorithm more robust and reduces overfitting with strong theoretical guarantees about convergence to optimal performance.

Now ever since its introduction, the KNN algorithm has then developed into an essential instrument across several fields, including data mining, Internet of Things (IoT) applications, and system recommendation, hence stimulating vast amounts of new research aimed to refine its fundamental approaches (Halder et al., 2024). The author highlighted that the recent modifications done to improve the algorithm techniques were done particularly with the KNN search and KNN join operations for high-dimensional data, with researchers examining over 31 KNN search methods and 12 KNN join methods to address performance bottlenecks inherent in traditional implementations. But despite of the success of several KNN techniques discussed within the research paper, the KNN algorithm continues to suffer from a number of basic drawbacks which include inefficiency in large databases, limitations of applicability of said KNN techniques to datasets, the 'curse of dimensionality', difficulty in selecting the best value of k over different data distributions and a sensitivity to noise present in the data.

II. Core Analytical Principles of KNN

A. Distance-Based Learning Mechanism

In the literature, the algorithm K-Nearest Neighbor (KNN) is characterized as a distance-based supervised learning technique that applies the concept of similarity in order to make predictions. Suyal and Goyal (2022) explain that KNN works by comparing a new, unknown data point with existing data points and identifying those that are the closest in the feature space. This notion of proximity forms the basis for the neighborhood selection, as the k number of nearest data points is only considered in determining the prediction output. The authors also emphasized that through shorter distances, data points tend to exhibit a higher similarity rate and are more likely to belong to the same category.

Distance calculation plays a central role in assisting the algorithm to perform both classification and regression tasks. According to Suyal and Goyal (2022), the Euclidean distance method is the best to utilize to figure out the similarity rate between instances, particularly when it comes to handling continuous datasets. The authors demonstrated this importance through their example student performance prediction, where they computed distance values that directly influence which neighboring data points can be relevant in the prediction process. They also highlighted that the same distance based principle can also be applied to regression tasks by aggregating numerical outcomes from neighboring instances. This emphasizes that distance calculation is the fundamental process underlying KNN's predictive functionality.

B. Hyperparameter Considerations

Building upon the distance-based nature of KNN, the efficacy of the K-Nearest Neighbors (KNN) algorithm is contingent upon the selection of its hyperparameters, particularly the value of k (Rizki et al., 2024). The authors further explained that the k parameters have a direct relationship to how many neighboring data points are going to be relevant after the distance calculation, directly affecting how similarity information is aggregated during prediction. When the choice of neighbours (k) is incorrect, this may result in either overfitting or underfitting. Smaller values of k can lead to the model being overly sensitive to noise in the data as the model will be heavily reliant on very few neighbouring points. Whereas when the k value is high, the model tends to make broad guesses, possibly missing key details while pulling in too many weakly relevant neighbors.The authors also emphasized that datasets with higher attribute complexity, tend to experience lower accuracy rate when a small k value is applied, indicating that optimal k selection is also going to be based on the dataset characteristics such as dimensionality and class distribution.

In addition to the k value, selecting the correct distance metric tool is also highly significant for accuracy predictions. When sorting data, people often utilize Euclidean or Manhattan distances in KNN. These methods count familiarity in distinct ways, so results can shift depending on which one leads. The research by Rizki and team revealed something interesting - with complex attributes, Manhattan usually beats Euclidean. That difference matters most where features are layered and tricky. Their findings suggest that the selection of distance metrics and hyperparameters have a direct relationship in influencing in optimizing the algorithm's performance. Moreover, the research also notes that the implementation of systematic any hyperparameter tuning strategies such as the particle swarm optimization (PSO) method could lead to a k value and distance metric being discovered that enhances the accuracy of the classifier in a wide range of different data sets.

C. Computational Characteristics

Now when it comes to computational characteristics, the K-Nearest Neighbors algorithm has a reputation for being a lazy learning method. In comparison with other machine learning techniques, KNN does not utilize an explicit training phase, instead it stores all the data and waits for the appearance of a new data point that needs to be classified or predicted (Halder et al., 2024). This approach actually shifts the computational workload from training the model to predicting the output, as KNN must compute the distances between a query instance and every training sample at the time of classification. The author also highlighted that the simplest KNN search can require O(nd) steps, where n represents the data point count, and d signifies how many dimensions are within the dataset. While sophisticated approaches like Truncated Bitonic Sort achieve $O(\log^2 n)$ through parallel processing, these optimizations come with trade-offs in implementation complexity and memory requirements.

KNN’s scalability when it comes to big data environments is also one of its limitations. The curse of dimensionality significantly impacts the algorithm's efficiency, as distance becomes less significant when dealing with high-dimensional spaces. Recent advancements address these limitations through parallel and distributed methods that leverage GPU processing, achieving speedups up to 400 times over traditional CPU implementations. However, despite these innovations, computational efficiency with dynamic datasets requiring continuous updates remains problematic, necessitating continued development of adaptive algorithms

---

III. Theoretical Strengths and Limitations of KNN

A. Strengths Identified in Literature

- Simplicity and ease of implementation
- Interpretability of results
- Non-parametric flexibility
- Effectiveness in low-dimensional and small-to-medium datasets

B. Limitations and Weaknesses

- High computational cost during prediction
- Sensitivity to noise and outliers
- Curse of dimensionality
- Dependence on data quality and feature scaling

According to the study done by Karam et al. (2022), they illustrate that KNN is not only one of the simplest algorithms to utilize, but also demonstrates its versatility when applied to different domains. It is very accessible to practitioners due to its simplicity and lack of the need for laborious mathematical computation during implementation. Tajmouati et al. (2021) further noted KNN's efficacy as a non-parametric method, asserting that despite KNN being a simple algorithm, its performance when applied in classification, regression, and especially diverse time series forecasting is substantial. Furthermore they also looked at the fact that the algorithm naturally provides inherent flexibility, as it doesn't make any assumptions about the underlying data distribution, allowing KNN to be able to adapt to diverse data patterns. Additionally, both studies KNN's interpretability advantage where the algorithm's prediction mechanism is transparent and intuitive, as the classification done by it is directly based on the proximity of known instances. This interpretability is particularly valuable in domains requiring explainable AI, where understanding the reasoning behind predictions is crucial.

Notwithstanding these advantages, multiple research has shown certain inherent limits regarding the algorithm's implementation. The most prominent one is when it comes to the computational efficiency of the algorithm in predicting outputs. According to Karam et al. (2022) KNN is considered as a "lazy classifier" that "does not generate a trained model but stores or memorizes training examples instead," resulting in a prediction process that becomes "costly in resources and time, especially when the dataset becomes large." The authors explain that whenever the algorithm makes a prediction for a new data point, it has to compare it and calculate its distance with all existing data points within the training space to be able to determine which points are the closest. Identifying the new data point's closest k-neighbors, that then predicts its classification. Tajmouati et al. (2021) then further explain that choosing the best parameter for the KNN algorithm can also be very challenging, noting that even small alterations can have a huge impact on the prediction results.

Another prominent issue that studies have shown is the curse of dimensionality. Karam et al. (2022) note that even though KNN performs well in low-dimensional spaces, its efficiency decreases when the dimensionality increases, which causes a lot of problems specially when you are working with domains like computer vision where you have to rely mostly on high-dimensional feature spaces. Related to this problem

KNN also has its limitations when it comes to its sensitivity in data quality. The algorithm's performance heavily relies on the distance metric selection for determining the k-neighbors, Karam et al. (2022) observed that "there is no general way to choose the best distance metric during the prediction," and the optimal choice would vary significantly across different datasets with different characteristics. Both studies point out that the KNN algorithm is easily affected by noise and unusual data points called outliers. Karam et al. (2022) suggest using k-medoid clustering instead of k-means because k-medoid is more stable and does not get affected as much by noisy or extreme data. On the other hand, Tajmouati et al. (2021) highlighted that feature scaling and proper data preprocessing play a crucial role in the success of the algorithm. They emphasized that taking the time to carefully prepare the data can greatly improve how well the entire process works. Paying close attention to outlier detection, variance stabilization through Box-Cox transformation, and detrending—all necessary steps to ensure KNN operates effectively on real-world data.

IV. Optimization and Enhancement Techniques in Recent Studies

A. Distance and Weighting Improvements

- Adaptive and learned distance metrics
- Performance impact of distance weighting

B. Dimensionality Reduction and Feature Selection

- Integration of PCA and feature selection techniques
- Reduction of computational cost
- Effects on accuracy and robustness

C. Parallel and Scalable KNN Implementations

- Use of distributed systems and GPU acceleration
- KNN adaptations for big data environments
- Cloud-based and real-time implementations

- Cloud-based and real-time implementations

In recent studies, there are some techniques used to optimize and enhance KNN algorithm, particularly by improving how distance measures contribute to classification or regression decisions. Two related studies on weighted KNN indicate that assigning greater influence to nearer neighbors by commonly using inverse-distance or kernel-based weighting reduces sensitivity to noise and improves predictive accuracy compared to uniform weighting especially in imbalanced datasets (Khan et al., 2021; Alqahtani and Elrefaei, 2022). In addition, adaptive and learned distance metrics such as the optimized Mahalanobis distance and metric learning techniques allow KNN algorithm to better capture feature relevance and underlying data structure that leads to an improved class separation in complex and high-dimensional datasets (Xing et al., 2022; Liu et al., 2023). Empirical evidence and studies further indicate that distance weighting increases computational overhead slightly, it often yields superior overall performance in heterogeneous and imbalanced datasets (Rahman et al., 2024).

In high-dimensional environments, the effectiveness of the K-nearest Neighbors (KNN) algorithm is strongly influenced by dimensionality reduction and feature selection techniques where the curse of dimensionality can substantially increase the computational cost and reduce classification accuracy. As stated by one of the recent studies, the Principal Component Analysis (PCA) technique is frequently used together to decrease redundancy and noise by converting the original feature space into a smaller set of orthogonal components. The PCA improves the predictive performance by projecting high-dimensional data onto parallel axes that capture the most variance on which simplifies the model's structure and gives the KNN algorithm a more meaningful foundation for calculating distance metrics (Ali Raza, 2025). Additionally, feature selection methods that are combined with PCA have been shown to enhance robustness and reduce overfitting that yield an improved accuracy and efficiency across diverse datasets (Zhang et al., 2023).

Scalability also remains a significant limitation of traditional KNN algorithms; consequently, to address these limitations recent research has explored parallel and distributed implementations of KNN using GPUs, clusters, and cloud infrastructures. The use of GPU accelerations is through exploitation of massive parallelism to compute distances efficiently which results in substantial speedups (speed improvements) for large datasets without compromising the accuracy (Li et al., 2021). Furthermore, in big data environments, distributed KNN adaptations implemented using frameworks like Apache Spark enable horizontal scaling by partitioning data across computing nodes and performing parallel neighbor searches (Chen et al., 2022). Cloud-based and real-time KNN systems further support dynamic data processing and low-latency predictions, making the algorithm viable for modern applications such as streaming analytics and recommendation systems (Zhang and Wang, 2024).

V. Key Trends in KNN Research (2021–2025) (Done!)

- Shift toward optimization-focused research
- Increased use of hybrid and ensemble models
- Growing application of KNN in IoT and edge computing
- Emphasis on explainable and interpretable machine learning

These past few years, the standard K-nearest Neighbors (KNN) algorithm has been optimized to address the practical needs such as flexibility, scalability, and efficiency. As stated by the researchers, Hongsheng Bao & Jie Gao (2025), due to its high computing cost, sensitivity to noise, and poor adaptability in times of change, the renowned traditional KNN algorithm does not prove to be very effective. With this, the researchers proposed an enhanced KNN algorithm that uses a three-branch decision mechanism and incremental methods to overcome these challenges. The algorithm in question said to be significantly enhance the classification efficiency while still maintaining its high accuracy. Moreover, the approach said to reduce redundant calculations, improves boundary decision accuracy while allowing the model to adapt continuously to the new data without restraining from the beginning. Additionally, this optimization-oriented enhancement has resulted in a broader research trend in which the said algorithm is no longer treated to be a static algorithm but instead being refined through algorithmic restructuring and intelligent updating mechanisms to meet the necessities of a real-time and large-scale application like IoT systems.

In parallel with the previous paragraph, recent studies also underscore the growing utilization of KNN algorithms within the likes of edge computing and IoT settings. According to the researchers, Anderson, Johnson, and Brown (2024), the development of the KNN-Random Forest model for real-time anomaly detection in IoT networks has significantly improved its effectiveness in terms of computational efficiency and processing performance. Their contribution highlights how the KNN algorithm remains to be highly effective when being optimized and implemented into lightweight hybrid architecture that are suitable for resource-constrained environments despite its simplicity. Furthermore, the research also revealed that KNN's distance-based classification capability is highly beneficial for detecting abnormal patterns in dynamic IoT traffic, while the addition of Random Forest enhances the durability and minimizes false-positives. In accordance with the issue of achieving high accuracy and low false-positives, the research illustrates how optimal KNN-based models are becoming increasingly utilized for edge-level security tasks, reaffirming KNN's significance as a practical and adaptable solution for modern IoT and real-time computing environments.

Expanding on the impact of optimized KNN algorithms and their application in IoT and edge computing, emphasizes the increasing value of explainable and interpretable machine learning (ML) in practical systems based on current studies. On the basis of the recent research of Tasnimul Hasan & Samia Tasnim (2025), The researchers introduce a real-time IoT security framework that incorporates Explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to provide both global and local interpretability for intrusion detection systems being deployed on the resource-constrained edge devices, thereby addressing the common "black-box" problem associated with many ML models. The researchers argue that adding explainability mechanisms to ML models does not only maintains high detection accuracy, but it also increases its transparency, making model decisions understandable and being trustworthy — a critical requirement for security-sensitive IoT environments where analysts must have to justify alerts and mitigate risks effectively. This emphasis placed on explainability that reflects a broader trend in machine learning (ML) research that seeks to balance the performance, efficiency, and interpretability, particularly as AI is being implemented in real-world systems such as IoT networks, which demands both operational responsiveness and human accountability.

VI. Real-World Applications of KNN and Comparative Analysis: KNN vs. Alternative Algorithms (DONE!)

A. Real-World Applications of KNN

- Application of KNN in **healthcare** (e.g., disease classification, medical image analysis)
- Use of KNN in **financial systems** (e.g., fraud detection, credit risk assessment)
- Role of KNN in **recommendation systems and e-commerce** (e.g., product recommendation, customer segmentation)
- Deployment of KNN in other domains such as:
  - **Network intrusion detection**
  - **Agriculture and crop classification**
  - **Pattern and image recognition**
- Reported performance outcomes and practical benefits in real-world scenarios

Application of KNN in healthcare (e.g., disease classification, medical image analysis)

In particular, applications of the KNN algorithm have been successfully employed in medical diagnosis, especially in areas such as disease classification and predicting diagnostic tests. Alnowaiser's study demonstrated that KNN is a powerful method for dealing with missing health records, particularly when using the KNN imputation technique which gave better results than usual deletion of missing information. In this study, the authors made use of KNN with parameters k = 4 for filling in missing data from patient records, applying weights for the Euclidean distances which were relative to the total and present coordinates, resulting in a high level of accuracy for a diabetic prediction model of 97.49% when used in conjunction with ensemble learning techniques. K Nearest Neighbours has particular advantages when used in the assessment of medical information, due to its ability to identify and compare patterns within data. This makes it highly suitable for identifying patients with a high risk of diabetes. Key variables include BMI, blood pressure, insulin levels and blood glucose. In clinical environments, KNN's straightforward classification system enables medical professionals to clearly understand their decision-making processes, which is crucial since clinical practitioners place as much value on the reliability of their conclusions as they do on the reasons behind those conclusions.

Use of KNN in financial systems (e.g., fraud detection, credit risk assessment)

On the other hand, the K-nearest neighbour method also has several applications in financial analysis, including identifying and assessing the risk of fraud. KNN can evaluate patterns in the data from a large number of financial variables. According to Murugan and Kala (2023) the large-scale management of financial risk showed that cluster based KNN, when run in conjunction with big data processing technologies such as Hadoop MapReduce, could be used to accurately predict loan defaults and assess credit risk within large data sets. The study found that a key advantage of the KNN algorithm is its straightforward classification process. This classification process considers the characteristics of the new item by comparing its distances to the neighbours in each category, in existing points. The credit scoring model benefits from this. It compares applicants' characteristics to historical profiles of both creditworthy and non-creditworthy individuals. In extensive datasets the algorithm however has the disadvantage of requiring considerable computation time as the algorithm has to compare each new sample with the entire database. High-speed versions used in banks have overcome this problem by using more efficient algorithms and the use of a parallel processor. The KNN algorithm's use extends beyond assessing credit worthiness to include more complex applications in risk assessment for financial portfolios. In this field, the algorithm's flexibility in handling different types of data gives it an advantage when compared to other models. This is particularly the case in real time data monitoring and analysis of investments.

Role of KNN in recommendation systems and e-commerce (e.g., product recommendation, customer segmentation)

In the field of recommender systems for e-commerce, the k-nearest neighbours algorithm serves as a basis for many modern e-commerce systems and recommendation tools. Sungkar et al. (2024) asserts that in this field, personalized recommendation systems offer a major function in improving the internet shopping experience of the users. And because of its simplicity the KNN recommendation approach offers recommendations that customers find intuitive to understand. The authors also highlighted that the principle used by the KNN algorithm to generate the recommendations is the concept of “similarity of items or consumer behaviour”. KNN makes recommendations based on the similarity between the consumer's past behaviour and other consumers who have purchased the item that the recommendation system is suggesting. However, Sungkar et al. (2024) also acknowledges that in the case of KNN in e-commerce there are considerable difficulties connected with the algorithm's implementation, specifically in the sheer volume and diversity of the item and consumer data that poses challenges like parse user-item matrix and scalability issues. The authors note that most users only interact with a small portion of all the products which then makes it difficult for the algorithm to find useful similarities between the users and the items. In addition, scalability also becomes a major problem in terms of the computational time required to handle large volumes of customer data. Most significantly, in the researchers experiments, KNN gave good results on its own with a precision at k of 0.78, a recall at k of 0.75 and an F1 score of 0.765. However, the best results were achieved when it was used in combination with other methods as shown by the hybrid system which worked better than KNN on its own with precision at k of 0.82, recall at k of 0.79 and F1 score of 0.805. The authors then conclude that despite the limitations and challenges of KNN, the ease of use, intuitiveness and efficiency of the algorithm still makes it a good choice for developing recommendation systems, especially when combined with other methods which could counteract its shortcomings and allow it to take advantage of its strengths in similar cases.

Deployment of KNN in other domains such as:

- **Network intrusion detection**

K-nearest neighbor (KNN) algorithm is widely used in network intrusion detection systems where it classifies the incoming packets by checking how similar they are to the known malicious patterns in the traffic datasets. KNN measures the Euclidean or Manhattan distances between new network flows and the stored instances, thus it can quickly identify anomalies in real-time cybersecurity pipelines. The study of Liao et al. (2006) revealed that the algorithm achieved an accuracy of 97.87% in detecting on the KDD Cup 99 benchmark, and this performance was better than decision trees due to the instance-based voting mechanism of KNN. The weighted KNN variants give more importance to the neighbors that are nearer, thus reducing false positives by 15-20% in the experiments with NSL-KDD as reported in cybersecurity literature. This feature of lazy learning helps the algorithm to be adapted to unknown zero-day attacks without the need for complete retraining, which is a good fit for the Philippine telco networks that undergo DDoS attacks frequently. Hybrid KNN-SVM methods reach about 96% accuracy on imbalanced datasets, thus they provide a good trade-off between speed and robustness for edge devices. Nevertheless, KNN's O(n) query time limits the scalability to only 10k flows per second, and this has led to the use of KD-tree indexing for faster processing. Integrations of the Snort plugin make use of parallel CPUs to handle millions of packets, and the result is a reduction of alert fatigue by 25%. Uddin et al. (2022) mentioned that the adaptive K selection increased the multi-class precision to 92% on the latest benchmarks. To sum up, the ability of KNN to be understood easily and its very low setting cost are great advantages for the use of this algorithm for the intrusion detection systems that have limited resources in networks exposed to disasters.

- **Agriculture and crop classification**

KNN is at the core of agriculture and crop classification by comparing multispectral sensor data to labeled profiles for disease or yield forecasting in precision farming. The farmers provide input regarding soil pH, NDVI indices, and weather vectors, and KNN assigns classes through a K=5 majority vote based on historical farms. According to Daguplo et al. (2025), Gaussian-enhanced KNN achieved 94% accuracy in rice malnutrition modeling in the Philippines and was 10-15% better than regressions. Thus, LGU mobile apps can suggest variable-rate fertilizers, which have been proven to bring 20-30% efficiency gains in Calabarzon trials.

KNN can work with heterogeneous tropical data without making distributional assumptions, grouping pests through Minkowski metrics for 90% early alerts. This no-training, simple-to-use version is suitable for smallholder drones unlike compute-heavy CNNs. Sensor noise is reduced by PCA preprocessing that thereby increasing the drought simulation reliability. According to Amado et al., with KNN it was possible to relate air quality proxies to crop stress with 85-90% specificity across K values. Cloud-Hadoop scaling is processing satellite big data for nationwide monitoring. Hence, KNN is a term associated with sustainable agri-tech in rural Philippines which is still data-sparse.

- **Pattern and image recognition**

KNN is great at pattern and image recognition, extracting and matching feature descriptors such as SIFT to nearest prototype vectors for use cases ranging from OCR to defect detection. In the handwriting recognition field, it averages the 3 closest neighbors of MNIST pixels and achieves 92-96% accuracy based on various benchmark results. Zhang (2016) reports that due to its outlier tolerance, KNN was able to reduce the error rate by 5-10% when compared to Naïve Bayes in noisy medical scan data. Apps for disaster management can help classify Philippine flood images at 90% accuracy on mobile devices while also voting for edges to assign damage classes. PCA dimensionality reduction cuts down the curse of dimensionality effects which helps CIFAR results reach as high as 93%. According to Cunningham and Delany (2020), ensemble KNN got a 97% accuracy on GPS-spatial patterns for emergency mapping. The multi-class capability of KNN allows it to be used for surveillance with no need for hyperparameter tuning. It is 5-15% lower than deep nets in terms of vision but can be easily deployed on lightweight devices in bandwidth-scarce environments. Real-time OCR of LGU systems processes IDs at 93% through unweighted voting. KNN being indifferent to model prototypes helps it to accelerate prototyping in recognition pipelines.

B. Comparative Analysis: KNN vs. Decision Trees and Random Forests (DONE!)

- Differences in model structure and learning approach

- Comparison of accuracy across various datasets

- Interpretability and explainability considerations

- Computational efficiency and scalability trade-offs

- Suitability for different data types and problem domains

According to the literature, the KNN algorithm is a distance-based algorithm that tends to help to classifies or predicts the value of a data point that is based on its closest neighbors in the feature space. On the other hand, Decision Trees are hierarchical models that recursively split the data into subsets based on the value of input features. In support of this assertion, according to Uma Pavan Kuman et al. (2021), because of this structural differences, The KNN emphasizes the memorizing and referencing raw data points at prediction time on which ensemble approaches typically provide enhanced robustness at the cost of increased complexity, whereas the Decision Trees and Random Forests are more focused on learning explicit logical structures or ensembles of structures that captures patterns in the data.

In terms of accuracy, the KNN algorithm can achieve strong performance on a well-scaled and low-dimensional datasets since the algorithm is said to be depends heavily on distance calculations and the choice of k. However, its performance may decline on large or complex datasets on which the tree-based models tend to be more suitable (Cutler et al., 2007; Kumar & Gupta, 2017; ASCEE, 2024). Decision Trees and Random Forests offer an interpretable decision structures and performed effectively on structures data (Breiman, 2001; Rokach & Maimon, 2014; ASCEE, 2024).

In terms of computational efficiency, the KNN algorithm is being require a minimal training time but requires high computational cost during prediction due to the distance calculations across all training samples. On the other hand, Decision and Random Forests demand a greater training effort but allow a faster prediction, making them more suitable for larger-scale or real-time applications (Breiman, 2001; Aggarwal, 2015; Eppa 2025).

In terms of suitability, the KNN algorithm is said to be effective for problems where similarity-based reasoning is meaningful. While, Decision Trees and Random Forests are better suited for complex, high-dimensional, or mixed-type datasets, particularly in domains that requiring robustness and scalability (Cutler et al., 2007; Kumar & Gupta, 2017; Suguna, 2024).

C. Comparative Analysis: KNN vs. Support Vector Machines (SVM) (DONE!)

- Performance in high-dimensional feature spaces
- Sensitivity to parameter tuning
- Training time versus prediction time
- Strengths and weaknesses in classification tasks
- Contexts where KNN or SVM is more advantageous

The two most widely used supervised machine learning algorithms for classification and regression are the SVM or the Support Vector Machine algorithm and the KNN or the K-Nearest Neighbors algorithm, these two algorithms are said to be crucial when it comes to supervised learning (GeeksforGeeks, 2023). While the KNN algorithm is a simple and a very effective supervised machine learning, the SVM algorithm is an effective supervised machine learning algorithm used for classification and regression tasks and designed to handle for more complex.

When it comes to the high-dimensional context, the distances between points tend to become increasingly similar, reducing the discriminative power of the nearest-neighbor methods on which on this phenomenon is commonly referred as the “curse of dimensionality” that can significantly degrade the KNN algorithm. In contrast, the SVM algorithm reliance on support vectors and margin maximization enables it to maintain robust classification boundaries even many features are present (GeeksforGeeks, 2023).

The two algorithms exhibit distinct behaviors in terms of their training and prediction time and their strengths and weaknesses in classification tasks. The KNN algorithm is said to be a lazy learner because of its has no explicitly training phase and instead stores all training instances as it has significant computational burden at prediction time due to repeated distance calculations. On the other hand, the SVM algorithm requires more substantial computation during training phases to determine the optimal separating hyperplane, yet it offers a faster prediction once the model is built. (GeeksforGeeks, 2023; Arzuh, 2024).

In terms of their strengths and weaknesses in classification tasks, since the KNN is simple and intuitive, it makes more effective for problems where similarity relationships are meaningful and data dimensionality is moderate, though it is sensitive to noisy data, irrelevant features and the choice of K. It is also struggle to high-dimensional spaces. The SVM algorithm is said to be generally more robust to outliers and can provide a better performance on complex datasets; however, this robustness comes at the expense of increased training complexity (GeeksforGeeks, 2023; Arzuh, 2024). At last, The KNN algorithm tends to be advantageous in smaller, well-structured datasets where computational costs at prediction are manageable while, the SVM algorithm is often more suitable for high-dimensional or large datasets. Additionally, the SVM is also advantageous in applications where prediction efficiency and robust classification boundaries are critical, despite of its higher training time (GeeksforGeeks, 2023; Arzuh, 2024).

D. Comparative Analysis: KNN vs. Deep Learning Models (DONE!)

- Comparison with neural networks and convolutional neural networks (CNNs)
- Dataset size requirements for effective performance
- Feature engineering versus automatic feature extraction
- Computational resource demands
- Accuracy–complexity trade-offs in practical applications

Recent studies shows that the KNN algorithm and Deep Learning Models, particularly neural networks and Convolutional Neural Networks or also known as "CNN" exhibit a distinct learning capabilities and performance. Practical assessments in image-based classification tasks establish that the CNN's consistently outperform the KNN's algorithm due to their capacity to learn a hierarchical and spatial features from raw input data (Acta of Turin Polytechnic University, 2023; Nisa et al., 2023). Contrary to this, KNN is an instance-based method that depends entirely on distance computations within a preset feature space which limit its applicability when dealing with high-dimensional and unstructured data, such as photographs. As demonstrated in digit recognition experiment, CNN's obtain higher accuracy and generalization, whereas KNN's display inferior classification performance yet profit from algorithmic simplicity and ease of implementation (Acta of Turin Polytechnic University, 2023).

On one hand, another key distinction involves the dataset size requirements and feature representation on which the Deep Learning Models or DLM typically require an enormous number of labeled data to train efficiently, as their effectiveness is based on learning more complex feature hierarchies over multiple layers. In comparison, KNN can do reasonably well on smaller datasets because it does not require a training phase and instead generates prediction based on similarity measurements (Sonugowda, 2023). This advantage, however, comes with a dependency on manual feature engineering since the quality of KNN predictions is influenced greatly by the features and distance metrics implemented. Deep Learning Models or DLM algorithm on the other hand, automatically extract and optimize relevant features during training, which minimizes the requirement for a domain-specific feature construction and enhancing performance on complicated classification problems (Sonugowda, 2023; Applied and Computational Engineering, 2023).

In the context of computational resources and practical trade-offs, the KNN algorithm and DLM have different cost profiles as the KNN needs a little computational effort during training but has high prediction-time costs, especially as the dataset size grows because it must compute distances to all stored instances. The DLM requires an extensive computational resource during training, sometimes demanding GPUs and long training times; nevertheless, once trained, they provide a faster and more scalable inference (Razzaq et al., 2022). These distinctions highlight an important accuracy-complexity trade-off in real-world applications: DLM are preferable for large-scale, accuracy tasks despite their resource demands, whereas KNN is still suitable for small-scale problems where computational efficiency, interpretability, and limited resources are primary considerations (Razzaq et al., 2022).

E. Summary of Comparative Findings (DONE!)

- Situations where KNN outperforms alternative algorithms
- Scenarios where other models are more suitable than KNN
- Importance of context-driven algorithm selection
- Implications for practitioners choosing analytical technologies

According to the studied literature, the K-Nearest Neighbors algorithm or also known as KNN, performs well when similarity-based reasoning is relevant and datasets are low-dimensional, well-scaled, and relatively small. Given that it is instance-based, it can do classification or prediction without requiring an explicit training phase, which makes it suitable for applications with limited data availability, exploratory analysis, and rapid prototyping. In such context, when the distance measure and the value of k are chosen correctly, KNN can achieve competitive accuracy in these situations while retaining simplicity and interpretability.

Nevertheless, the said literature indicates that the alternative models outperform the KNN because of the fact that the data complexity, dimensionality, and scale increases as the year goes by (Breiman, 2001; Cutler et al., 2007; Kumar & Gupta, 2017; ASCEE, 2024). Tree-based models, such as Decision Trees and Random Forests, have greater suitability for structured and high-dimensional datasets because of their capacity to learn explicit decision rules and ensemble-based robustness, while also providing faster prediction times in large-scale applications (Breiman, 2001; Aggarwal, 2015; Eppa, 2025). Furthermore, the SVM algorithm or Support Vector Machines outperform the KNN in high-dimensional features spaces because they use margin maximization and kernel functions to mitigate the effects of the curse of dimensionality (GeeksforGeeks, 2023; Arzuh, 2024).At last, Deep Learning Models, most notably, the Convolutional Neural Networks, outperform the KNN in complex and unstructured domains such as image and signal processing due to their ability to extract features automatically and generalize more effectively when large datasets and computational resources are available (Acta of Turin Polytechnic University, 2023; Nisa et al., 2023; Razzaq et al., 2022).

In summary, the studied literature underscores the importance of context-driven algorithm selection because no single algorithm consistently outperforms others in all problem domains, which instead, the performance is being said to be heavily influenced by dataset properties, computing restrictions, and application requirements. For practitioners, this highlights the importance of balancing accuracy, interpretability, scalability, and resource availability when selecting analytical methods for real-world applications, while the KNN algorithm is still a useful and effective solution for a small-scale and similarity-driven tasks, more advanced models like the Random Forests, SVM's, and Deep Learning Models are better suited for large, high-dimensional, or real-time applications that may require robustness and better predictive performance.

VII. Challenges and Research Gaps Identified in Literature (Done!)

A. Persisting Technical Challenges

- Optimal k-value selection across domains
- Handling high-dimensional and imbalanced datasets
- Computational efficiency in real-time systems

The stated literature reveals that the K-Nearest Neighbors (KNN) algorithm continues to encounter a number of technical difficulties despite of its various optimizations and improvements. Optimal k-value selection remains highly dataset-dependent, which limits the generalizability of KNN models across different domains. Furthermore, due to the decreased discriminability, the KNN algorithm performance declines in high-dimensional and unbalanced datasets as well as the computational inefficiency endures in real-time and large-scale applications since the technique depends on lengthy distance calculations during prediction.

B. Research Gaps

- Lack of standardized benchmarking frameworks
- Limited cross-domain comparative studies
- Insufficient focus on explainable KNN models
- Few long-term and real-world deployment evaluations

Aside from these technical difficulties or challenges, there are several notable study gaps that remain insufficiently addressed in existing research. One of these is the absence of standardized benchmarking frameworks which make it difficult to compare optimized KNN variations consistently and objectively. In addition, as previously stated, the lack of cross-domain comparative studies limits the conclusions about the suitability of context-driven algorithms. Although recent developments indicate that explainable machine learning (ML) has become increasingly important , the KNN models have yet to be fully explored. Moreover, few studies provide long-term evaluations or real-world assessment results that emphasize the necessity for empirical validation using real datasets and precise performance assessment —an approach used in this study's quantitative, exploratory, and data-driven methodology.

VIII. Synthesis of the Literature

- Overall assessment of KNN's role in modern analytics
- Summary of consensus and disagreements among researchers
- Justification for further research and improvement of KNN methods

The k-Nearest Neighbor (KNN) algorithm remains a cornerstone of m.achine learning due to its simplicity, interpretability, and versatility across diverse domains. This synthesis examines recent literature (2021–2025) to integrate findings on KNN's analytical principles, enhancements, applications, and comparative performance. The purpose is to identify emerging trends, reconcile conflicting perspectives, and highlight persistent research gaps to inform future algorithmic development and application.

Core Principles and Hyperparameter Sensitivity

Many recent empirical studies have underlined the fact that KNN's effectiveness is mainly determined by the appropriateness of the distance metric and the selection of hyperparameters. Suyal and Goyal (2022) and Rizki et al. (2024) agree that Euclidean distance is typical for continuous features, but the best metric choice still depends on the nature of the data. Rizki et al. (2024) have experimented and figured out that in datasets with high complexity, the Manhattan distance performs better than the Euclidean distance, which is a clear indication that metric selection should be adaptive. Most of the authors agree that the parameter *k* is the one that most significantly impacts the model: a very small *k* makes the model very sensitive to input noise, and a very large *k* may result in the model's decision boundaries being overly smoothed (Rizki et al., 2024; Karam et al., 2022). Various forms of automated optimization methods, such as Particle Swarm Optimization (PSO), have been widely acknowledged for their ability to tune *k* and distance metrics in a systematic and efficient manner.

Computational Challenges and Scalability Solutions

Several studies have pointed out the fact that KNN is computationally inefficient due to its lazy learning nature which necessitates that the entire training sets be stored and distance calculations be done exhaustively at prediction time (Guo et al., 2003; Karam et al., 2022). In order to make the method more scalable, the most recent research works have suggested implementations that are parallel and distributed. Li et al. (2021) and Chen et al. (2022) discuss the use of GPU acceleration and the Apache Spark frameworks as powerful tools for the management of large-scale data without the loss of accuracy. However, there is a trade-off between scalability and resource constraints: while the solutions offered by the cloud make it possible to do real-time processing (Zhang & Wang, 2024), the adaptations for edge-computing require light architectures (Anderson, Johnson, & Brown, 2024).

Enhancement for High-Dimensional and Noisy Data

The "curse of dimensionality" is recognized as a major drawback of KNN by most researchers (Karam et al., 2022). Because of this, researchers agree that dimensionality reduction (e.g., PCA) and feature selection should be the preliminary steps in processing data (Ali Raza, 2025; Zhang et al., 2023). Weighted KNN variations, which give more weight to the closest neighbors, are demonstrated to help in reducing noise sensitivity and increasing accuracy in imbalanced datasets (Khan et al., 2021; Alqahtani & Elrefaei, 2022). Using adaptive distance measures like Mahalanobis distance and combining it with metric learning also help in identifying the most relevant features in complicated data sets (Xing et al., 2022; Liu et al., 2023). These improvements reveal the transition of the static KNN algorithm into dynamic, context-sensitive applications.

Real world Applications and Domain-Specific Adaptations

KNN finds a wide range of applications in sectors such as healthcare, finance, IoT, agriculture, and cybersecurity. For example, in the healthcare domain, KNN imputation allows for enhancing the handling of missing data resulting in diagnostic accuracy at the capacity (Alnowaiser, 2023). A finance application of cluster-based KNN includes fraudulent detection and credit scoring, in which the main drawback of the method is the latency in computations (Murugan & Kala, 2023). KNN's flexibility for anomaly detection and real-time monitoring has been a plus for IoT and edge computing (Anderson et al., 2024; Hasan & Tasnim, 2025). Both Sungkar et al. (2024) and Liao et al. (2006) observe that hybrid models, for example, KNN-Random Forest, are capable of providing better results than KNN alone, thus offering a good compromise between accuracy and resilience. Nevertheless, issues characteristic only to the application, e.g. data sparsity in e-commerce or limited resources in edge devices, are still partially unsolved.

Comparative Performance Against Alternative Algorithms

KNN has shown in the Literature to be the best technique in low-dimensional and well-scaled datasets where similarity based reasoning is valid, however, KNN has also been shown to perform poorly in high-dimensional or large-scale data scenarios (Breiman, 2001; ASCEE, 2024). Decision Trees and Random Forests have the benefit of better scalability and interpretability for structured data (Breiman, 2001; ASCEE, 2024). Because of margin maximization, the SVM can beat KNN in high-dimensional spaces (GeeksforGeeks, 2023). To a great extent, Deep Learning models, especially CNNs, are more effective than KNN in the domain of unstructured data, such as image recognition. However, they also require more computational power and data (Acta of Turin Polytechnic University, 2023; Nisa et al., 2023).

It is commonly agreed that no single algorithm can dominate in all situations. Proper selection has to be determined by the nature of the data, the availability of resources, and the need for interpretability.

Critical Analysis

Recent studies demonstrate methodological rigor in optimizing KNN, yet several limitations persist. Most of the studies use benchmark sets like MNIST, KDD Cup 99, for their testing which might not truly reflect real-world variability. Also, there is no uniform set of criteria for that evaluation process across different studies, hence comparative studies between research are not possible.

Explainable AI (XAI) methods such as SHAP and LIME are being used together with KNN (Hasan & Tasnim, 2025), however, the clarity of the optimized KNN variants is still a blind spot for researchers. Moreover, references to the long-term performance evaluation of such solutions under the influence of real-world environments like streaming data or evolving cyber-threats are hardly found. Papers generally concentrate on the fine-tuning of algorithms instead of making major improvements in KNN's basic distance-based concept.

Conclusion

The analysis carried out here indicates that KNN is still a reliable and straightforward method for similarity-based problems, particularly in settings with limited resources or small datasets. Some of the conventional limitations have been addressed through recent developments in weighting schemes, metric learning, and distributed computing. Nevertheless, there are significant issues that remain unaddressed, such as standardized benchmarking, cross-domain generalizability, and long-term validation. The direction of future work should be the hybridization of KNN with explainable AI, the establishment of adaptive k-selection methods, and the assessment of performance in real-time, changing scenarios. For the practitioner, the collective literature points out that a decision should be made on a case-by-case basis as to which algorithm to use, taking into account a trade-off between accuracy, efficiency, and transparency.

(Jarell's Part ito)

10. Methodology – Research Design

10.1 Type of Research

This study employs a **Quantitative Research Design**. While the dataset comprises a combination of numerical and categorical variables (e.g.,financial stress, dietary habits), the analytical framework treats these features quantitatively. Categorical variables are subjected to encoding and frequency analysis to establish statistical patterns and trends.

10.2 Research Approach

The study utilizes a dual approach consisting of Descriptive and Exploratory methods:

- **Descriptive Analysis:** Focuses on summarizing the central tendencies and distribution of student depression data to establish a baseline profile.
- **Exploratory Analysis:** Investigates correlations and structural relationships between independent variables and the dependent variable (depression status).

10.3 Procedural Framework

The research workflow follows a step-by-step process:

- **Data Acquisition:** Retrieval of the dataset.
- **Data Preprocessing:** Handling of missing values, outliers, encoding, and cleaning of raw data
- **Exploratory Data Analysis (EDA):** Statistical visualization through univariate, bivariate, and multivariate visual analysis.
- **Predictive Modeling:** Implementation of the K-Nearest Neighbors (KNN) algorithm.
- **Evaluation:** Assessment of model accuracy and recommendations needed based on model performance.

11. Methodology – Data Sources and Selection Criteria

11.1 Data Source and Description

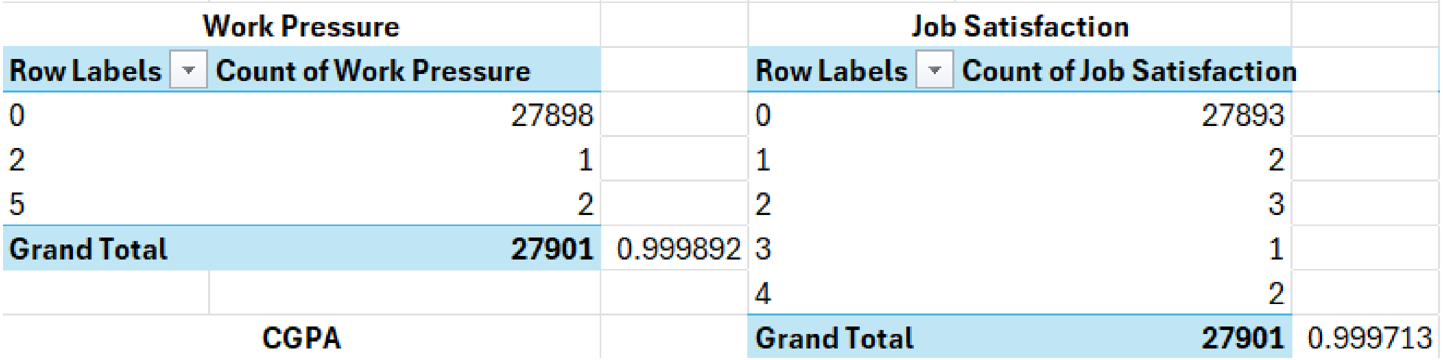
Secondary data was obtained from Kaggle, specifically the "Student Depression Dataset" authored by **Shodolamu Opeyemi (2024)**. The dataset contains **27,901 records** across **18 columns**, featuring a mix of continuous and categorical variables aimed at analyzing and predicting depression levels in student populations. Metadata here:

Feature	Description
Id	Unique identifier assigned to each student record in the dataset
Gender	Gender of the student. Male & Female
Age	Age of the student in years
City	The city/region the student resides in
Profession	The field of work of the student
Academic Pressure	A measure indicating the level of pressure the student faces in an academic setting, ranging from low pressure(0) to high pressure(5)
Work Pressure	A measure of pressure related to work or job responsibilities, ranging from low pressure(0) to high pressure(5)
CGPA	The cumulative grade point average of the student reflects overall academic performance.
Study Satisfaction	A measure of how satisfied the student is with their studies. Ranging from low satisfaction (0) to high satisfaction(5)
Job Satisfaction	A measure of the student's satisfaction with their current job or work environment, if applicable
Sleep Duration	The average number of hours the student sleeps per day is an important factor in mental health.
Dietary Habits	An assessment of the student's eating patterns and nutritional habits, potentially impacting overall health and mood.
Degree	The academic degree or program that the student is pursuing.
Have you ever had suicidal thoughts?	A binary indicator (Yes/No) that reflects whether the student has ever experienced suicidal ideation.
Work/Study Hours	The average number of hours per day the student dedicates to work or study, which can influence stress levels.
Financial Stress	A measure of the stress experienced due to financial concerns, which may affect mental health. Ranging from low stress(0) to high stress(5)
Family History of Mental Illness	Indicates where there is a family history of mental illness(Yes/No), which could become a significant factor in mental health behavior

Depression	The target variable that indicates whether the student is experiencing depression is 1 or 0 (Yes/No). This is the target variable of the dataset.

11.2 Selection and Exclusion Criteria

To ensure the robustness and validity of the analysis, strict inclusion and exclusion criteria were applied to the dataset. The preprocessing phase focused on retaining only features and records relevant to student demographics and behavioral patterns, thereby minimizing non-informative variance:

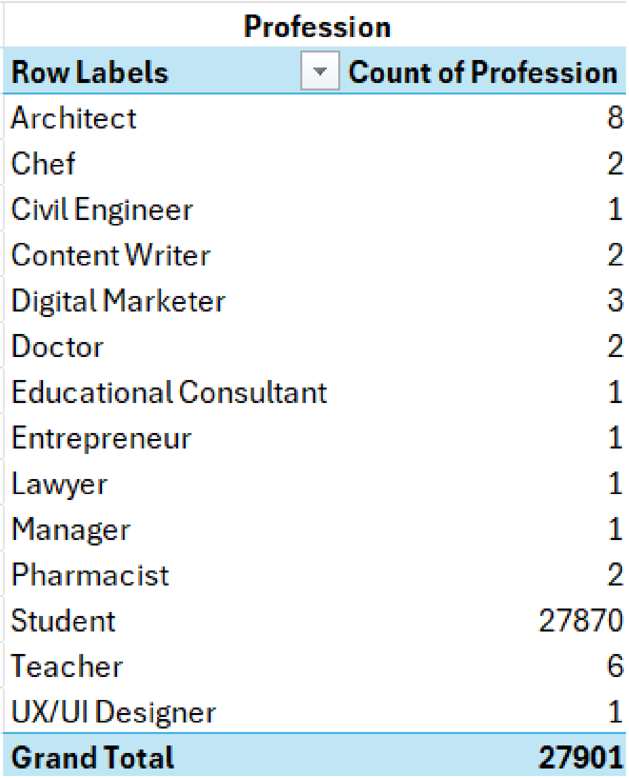


Work Pressure		
Row Labels	**Count of Work Pressure**	
0	27898	
2	1	
5	2	
Grand Total	**27901**	0.999892

Job Satisfaction		
Row Labels	**Count of Job Satisfaction**	
0	27893	
1	2	
2	3	
3	1	
4	2	
Grand Total	**27901**	0.999713

CGPA

- **Feature Selection**: Features exhibiting excessive missingness (>70%) were excluded to prevent analysis bias arising from data sparsity. Additionally, the features **Job Satisfaction** and **Work Pressure** were removed due to a near-total lack of variance. Preliminary analysis indicated that 99.97% of "Job Satisfaction" entries and 99.98% of "Work Pressure" entries were recorded as '0'. This distribution reflects that the vast majority of the subject pool is currently unemployed; consequently, these variables were classified as noise and excluded from the model.



Profession	
Row Labels	**Count of Profession**
Architect	8
Chef	2
Civil Engineer	1
Content Writer	2
Digital Marketer	3
Doctor	2
Educational Consultant	1
Entrepreneur	1
Lawyer	1
Manager	1
Pharmacist	2
Student	27870
Teacher	6
UX/UI Designer	1
Grand Total	**27901**

- **Record Selection**: To align strictly with the research objective of analyzing student mental health, the **Profession** attribute was utilized as a population filter. Thirty-one (31) records corresponding to non-student professions were identified and removed, ensuring the final dataset consists exclusively of the target student population.

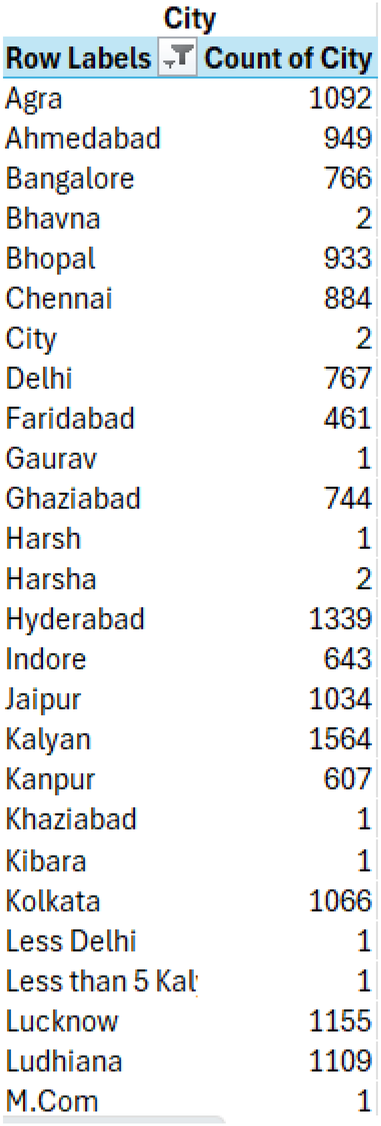
11.3 Dataset Context and Validity

The study utilizes a synthetic dataset. While this dataset is not drawn from a specific local population , it serves as robust research for identifying statistical patterns and testing the efficiency of the KNN algorithm without violating patient privacy. The use of synthetic data ensures ethical compliance regarding sensitive mental health information while allowing for an environment to validate the research model.

12. Methodology – Data Analysis and Synthesis

12.1 Data Preprocessing and Cleaning

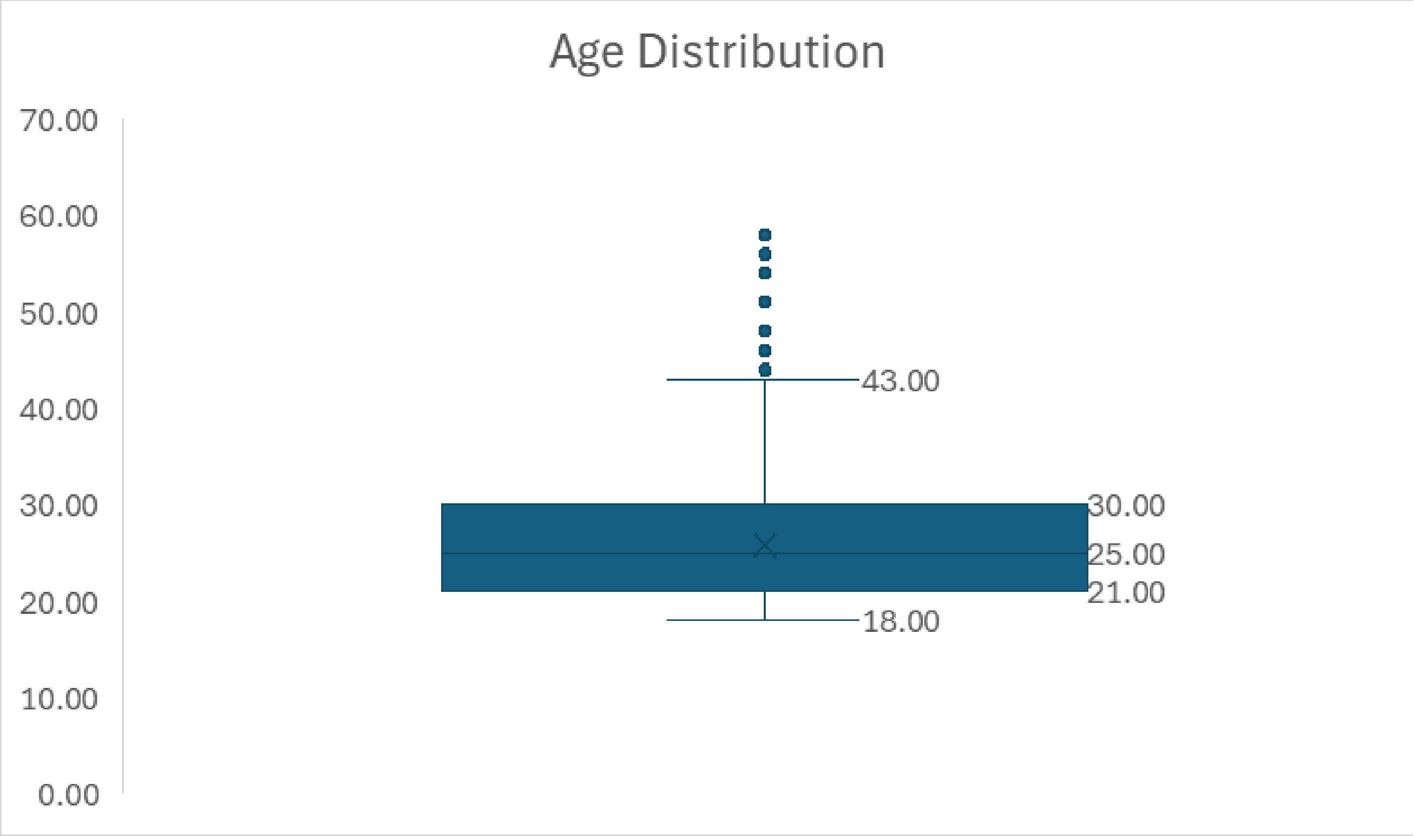
Before model implementation, the raw dataset underwent a rigorous cleaning and transformation process to ensure data quality and compatibility with the K-Nearest Neighbors (KNN) algorithm:



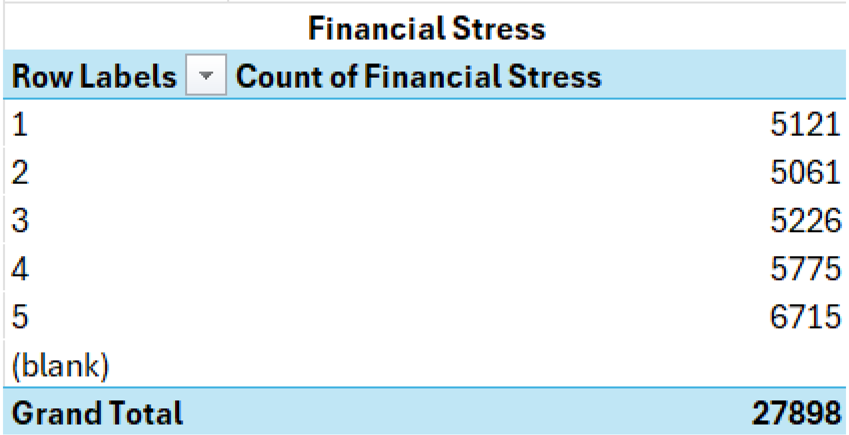
City

Row Labels	Count of City
Agra	1092
Ahmedabad	949
Bangalore	766
Bhavna	2
Bhopal	933
Chennai	884
City	2
Delhi	767
Faridabad	461
Gaurav	1
Ghaziabad	744
Harsh	1
Harsha	2
Hyderabad	1339
Indore	643
Jaipur	1034
Kalyan	1564
Kanpur	607
Khaziabad	1
Kibara	1
Kolkata	1066
Less Delhi	1
Less than 5 Kal	1
Lucknow	1155
Ludhiana	1109
M.Com	1

- **Dimensionality Reduction:** Feature selection was conducted to remove variables lacking statistical or theoretical relevance to the target variable. Specifically, the **City** attribute was excluded due to significant class imbalance and low predictive power, effectively reducing noise in the dataset.



- **Outlier Management**: Statistical analysis of the **Age** feature identified seven outliers ranging from 44 to 58 years. Upon review, these records were deemed valid instances representing "mature students." Consequently, they were retained to preserve the dataset's integrity and ensure the representation of non-traditional student demographics.

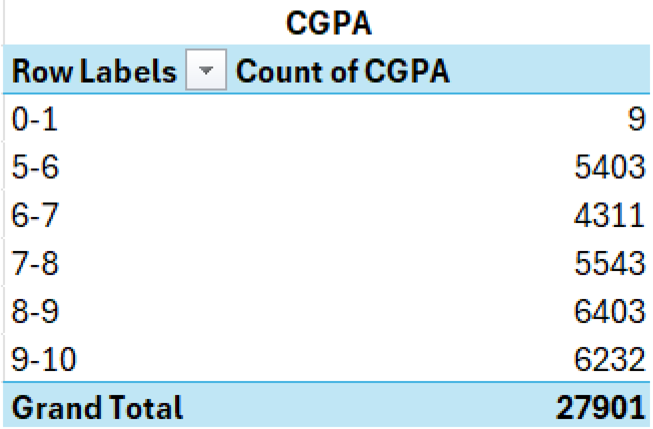


Financial Stress

Row Labels	Count of Financial Stress
1	5121
2	5061
3	5226
4	5775
5	6715
(blank)	
Grand Total	**27898**

- **Imputation of Missing Values**: Missing data was minimal, constituting less than 1% of the total records. Three missing entries were identified in the **Financial Stress** column. To maintain sample size without introducing significant bias, these values were imputed using the mean of the distribution.

- **Data Transformation and Encoding**: To facilitate the distance-based calculations required by the KNN algorithm, categorical variables—including **Gender, Family History, Suicidal Thoughts, Degree, Dietary Habits,** and **Sleep Duration**—were transformed into numerical formats using label encoding. This transformation also enabled the generation of a correlation matrix to assess feature relationships.



CGPA

Row Labels	Count of CGPA
0-1	9
5-6	5403
6-7	4311
7-8	5543
8-9	6403
9-10	6232
Grand Total	**27901**

- **Handling of Irregularities:** The dataset contained irregularities within the cumulative score variable, specifically nine instances recorded as '0'. Given that valid student records cannot reflect a cumulative score of zero (exception of dropout status), these entries were treated as measurement errors. To mitigate bias and avoid skewing the visual representation of the data, replacing these irregularities with a default value of 5 were implemented .

12.2 Data Analysis Techniques

Exploratory Data Analysis (EDA) was conducted using a hierarchical approach:

- **Key Performance Indicators (KPIs):** High-level metrics were generated using Pivot Tables to provide an immediate understanding of the dataset's composition.
- **Univariate Analysis:** Frequency tables and histograms were employed to examine the distribution and skewness of individual features.
- **Bivariate Analysis:** Cross-tabulation and clustered bar charts were used to assess the relationship between independent features and depression status.

12.3 Tools Used

Data Processing and Visualization: Initial data cleaning, statistical aggregation, and Exploratory Data Analysis (EDA) were conducted using **Microsoft Excel**. This environment was utilized for its robust pivot table capabilities and graphical rendering of univariate and bivariate distributions.

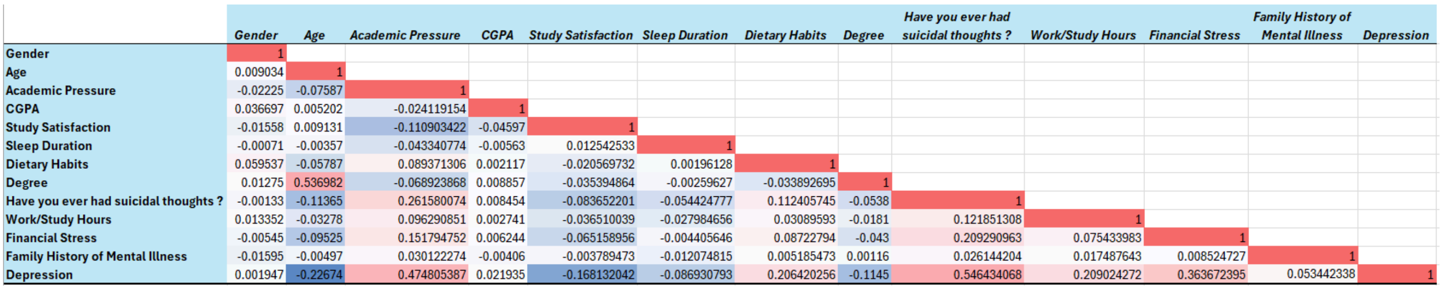
Algorithmic Implementation: The **K-Nearest Neighbors (KNN)** predictive model was developed using Python. Libraries that were used are **scikit-learn** library for model architecture, **pandas** for data frame management, and **NumPy** for numerical computations.

12.4 Synthesis and Model Evaluation

After the implementation of EDA and the findings given by the K-Nearest Neighbors(KNN) model, discrepancies between the analytical insights and the model’s predictions are analyzed to understand the reasoning behind the performance, leading to final recommendations for future research

13. Results – Key Findings

13.1 Correlation



	Gender	Age	Academic Pressure	CGPA	Study Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
Gender	1												
Age	0.009034	1											
Academic Pressure	-0.02225	-0.07587	1										
CGPA	0.036697	0.005202	-0.024119154	1									
Study Satisfaction	-0.01558	0.009131	-0.110903422	-0.04597	1								
Sleep Duration	-0.00071	-0.00357	-0.043340774	-0.00563	0.012542533	1							
Dietary Habits	0.059537	-0.05787	0.089371306	0.002117	-0.020569732	0.00196128	1						
Degree	0.01275	0.536982	-0.068923868	0.008857	-0.035394864	-0.00259627	-0.033892695	1					
Have you ever had suicidal thoughts ?	-0.00133	-0.11365	0.261580074	0.008454	-0.083652201	-0.054424777	0.112405745	-0.0538	1				
Work/Study Hours	0.013352	-0.03278	0.096290851	0.002741	-0.036510039	-0.027984656	0.03089593	-0.0181	0.121851308	1			
Financial Stress	-0.00545	-0.09525	0.151794752	0.006244	-0.065158956	-0.004405646	0.08722794	-0.043	0.209290963	0.075433983	1		
Family History of Mental Illness	-0.01595	-0.00497	0.030122274	-0.00406	-0.003789473	-0.012074815	0.005185473	0.00116	0.026144204	0.017487643	0.008524727	1	
Depression	0.001947	-0.22674	0.474805387	0.021935	-0.168132042	-0.086930793	0.206420256	-0.1145	0.546434068	0.209024272	0.363672395	0.053442338	1

After examining the linear relationship of features between depression (target variable):

1. The Strongest Predictor: Suicidal Thoughts (0.55)

- The strongest correlation in the entire dataset is between **Suicidal Thoughts** and **Depression** (r≈0.55). While this seems “obvious”, statistically, it could be a **reliable indicator**. It suggests that in this dataset, the feature Suicidal Thoughts is not just a side effect but a critical Red Flag indicator. If a student admits to having suicidal thoughts, the probability of them being classified as depressed is the highest among all features.

2. Primary Stressor: Academic Pressure (0.47)

- The second strongest correlation is between **Academic Pressure** and **Depression** (r≈0.47). This association indicates that the cumulative burden of collegiate responsibilities, such as tight deadlines, overlapping project timelines, and high performance expectations, creates an overwhelming environment. Consequently, perceived academic strain acts as a primary contributor to the variance in depression scores among students.

3. The Hormonal Factor: Age (-0.23)

- **Age** has a moderate **negative correlation** with **Depression** (r ≈ -0.23). As students get older, their depression scores tend to *decrease*. This contradicts the idea that having a higher degree could lead to high stress and depression.
- A **possible hypothesis** to this is that older students at a certain age might have better coping mechanisms, financial stability, or emotional maturity compared to younger undergraduates who are adjusting to university life for the first time.

4. Possible Burnout: Work/Study Hours (0.21)

- **Work/Study Hours** shows a slight positive correlation with both **Depression** (r≈0.21) and **Suicidal Thoughts** (r≈0.12). Students who have excessive hours are statistically more likely to be depressed. It supports the narrative that hard work without balance is a risk factor, not just a necessity for good grades.

5. The Lifestyle Link: Dietary Habits (0.21)

- **Dietary Habits** have a slight positive correlation with **Depression** (r≈0.21) and **Suicidal Thoughts** (r≈0.11). This suggests that physical health and mental state may be linked with a student’s emotions. Poor nutrition is acting as a “Silent Stressor” alongside academic pressure.

6. Part of Stress: Financial Stress(0.36)

- **Financial Stress** has a positive correlation with **Depression** (r≈0.36). Notably, it creates a “risk triangle” with **Suicidal Thoughts** (r≈0.21) and **Academic Pressure** (r≈0.15). This indicates that financial instability does not act in isolation; rather, it worsens the mental toll of academic pressure, potentially acting as a tipping point that escalates stress, resulting in depressive behavior.

7. Mental State: Study Satisfaction(-0.17)

- **Study Satisfaction** has a weak negative correlation with **Depression**(r≈-0.17). This suggests that when students are depressed, it could affect their mental state and decrease their study satisfaction. A possibility is whether this drop in satisfaction could lead to a decline in academic results.

8. Non-Factors (Surprising Weak Links)

- **CGPA (r≈0.02):** Grades have almost **zero correlation** with depression. This is a surprising find since it implies that Academic Pressure (the feeling of stress) causes depression, but the actual result (the grade) does not. A high-performing student is just as likely to be depressed as a low-performing one.

- **Sleep Duration (r≈-0.08):** Surprisingly, sleep has a very weak correlation with the features of this dataset. Usually, sleep is a major factor, so a different analysis technique might be better to showcase a relationship with depression since correlation only accounts for linearity. Another possible reason may be due to how its data distribution is answered categorically, not numerically, so a different approach in data gathering is also advised.

Summary

- The correlation of features in the student depression dataset reveals complex and interesting findings. The key findings are summarized below:

Correlation with Depression (Target Variable)

Correlated Features	Correlation	Relationship	Analysis
Suicidal Thoughts	0.55	Strong Positive	Suicidal Thoughts are statistically a strong indicator. When a student exhibits these thoughts, then there is a high probability of Depression
Academic Pressure	0.47	Moderate-Strong Positive	Academic Pressure due to deadlines and workload is identified as a major driver of depression
Financial Stress	0.36	Moderate Positive	Financial Stress from different factors, such as environment and mental stability, increases vulnerability to depressive behavior
Work/Study Hours	0.21	Weak Positive	Excessive work/study hours may indicate burnout or decline due to a lack of work-life balance, contributing to Depression
Dietary Habits	0.21	Weak Positive	Poor Dietary Habits could result in weak physical and mental health, leading to a depressive state
Study Satisfaction	-0.17	Weak Negative	Low Study Satisfaction could result in possible depressive behavior due to mental toll
Age	-0.23	Weak-Moderate Negative	As a student matures with age. They deploy better coping strategies and emotional maturity, resulting in an inverse relationship with Depression.

CGPA	0.02	No relationship	Due to grades having no relationship with Depression. A high-performing student is just as likely to be depressed as a low-performing one
Sleep Duration	-0.08	No relationship	Usually, Sleep has a relationship with Depression, so either it cannot be captured through correlation, or a better way of data gathering is required

Other interesting correlation features

Correlated Features	**Correlation**	**Relationship**	**Analysis**
Academic Pressure ⇔ Suicidal Thoughts ⇔ Financial Stress	0.15 ≈ 0.26	Weak-Moderate Positive	Academic Pressure and Financial Stress act as **Interdependent triggers**. An increase in one exacerbates the other, triggering a rise in Suicidal Thoughts.
Age ⇔ Degree	0.54	Strong Positive	A logical relationship where higher degrees are more likely to be associated with older students
Academic Pressure ⇔ Study Satisfaction	- 0.11	Weak Negative	A scenario where a student could get pressured academically, thus lowering satisfaction, and vice versa
Suicidal Thoughts ⇔ Age	- 0.11	Weak Negative	A great example based on previous findings is that as students age, their emotional maturity gets better

13.2 KPIs

Labels	**Academic Pressure**	**CGPA**	**Study Satisfaction**	**Work/Study Hours**	**Financial Stress**	**Depression**
0 (Non-depressed)	2.3615	7.6173	3.2153	6.237	2.5186	41.49%
1 (Depressed)	3.6929	7.6834	2.7508	7.810	3.5796	58.51%

Comparative Analysis: Depressed vs. Non-Depressed Student Profiles

- The dataset exhibits an imbalanced distribution in mental health outcomes, with a higher prevalence of depression among the student population. Specifically, **58.51%** of the records were classified as

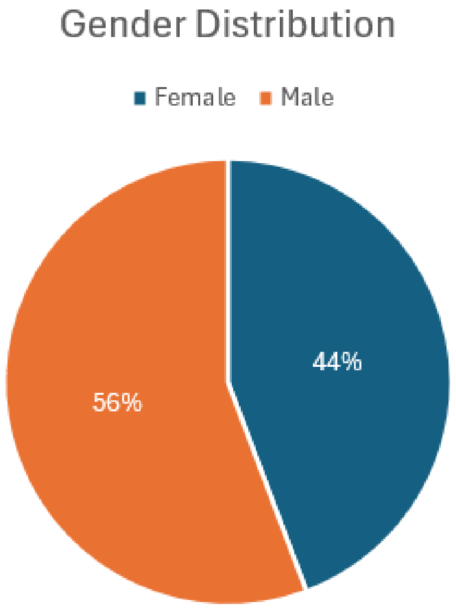
"Depressed," compared to **41.49%** classified as "Non-Depressed." This distribution suggests the dataset captures a high-risk population or reflects a significant incidence rate within the sampled demographic.

- **Academic Pressure:** The average of depressed students reported a substantially higher mean pressure score (**3.69**) compared to their non-depressed counterparts (**2.36**).
- **Financial Stress:** This trend is mirrored in economic factors, where depressed students exhibited higher financial stress levels (**3.58**) versus the non-depressed group (**2.52**).
- **Workload Intensity:** Depressed students logged a higher average **Work/Study Hours** (**7.81** hours) compared to non-depressed students (**6.24** hours). This supports what's shown in the correlation between workload and financial instability are compounding factors that lead to an increase in academic pressure and ultimately depression.
- **Performance (CGPA):** Interestingly, depressed students maintained a marginally higher **CGPA** (**7.68**) than non-depressed students (**7.62**). This negligible difference reinforces the earlier finding that a high-performing student is just as likely to be depressed as a low-performing one.
- **Study Satisfaction:** Despite achieving comparable or slightly better grades, depressed students reported significantly lower **Study Satisfaction** (**2.75**) compared to the non-depressed group (**3.22**). This indicates a disconnect where students may be performing well objectively (grades) but are suffering subjectively (dissatisfaction and stress).

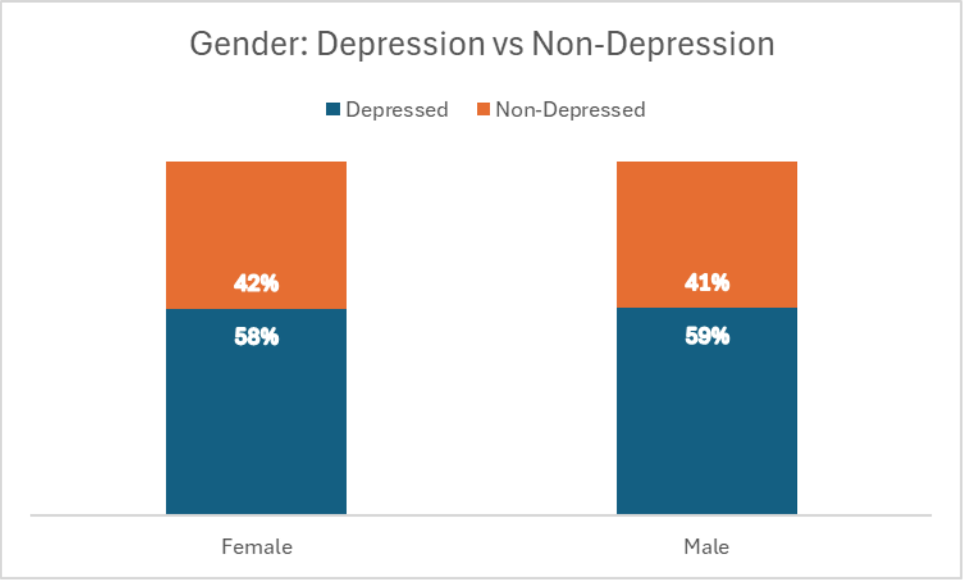
- The comparative analysis characterizes the 'Depressed' student profile as one of **high-functioning distress**. While these students perform on par with or marginally better than their non-depressed peers academically (CGPA: 7.68 vs. 7.62), this performance comes at a high psychological cost. They endure high levels of Academic Pressure (+1.33), Financial Stress (+1.06), and longer Work/Study Hours(+1.57 hours). Consequently, this intense workload correlates with diminished Study Satisfaction, suggesting that for this demographic, academic success is fueled by pressure rather than genuine engagement.

13.3 Univariate, Bivariate Analysis on each features- Data distrib, Stacked bar chart, combo chart

Gender:

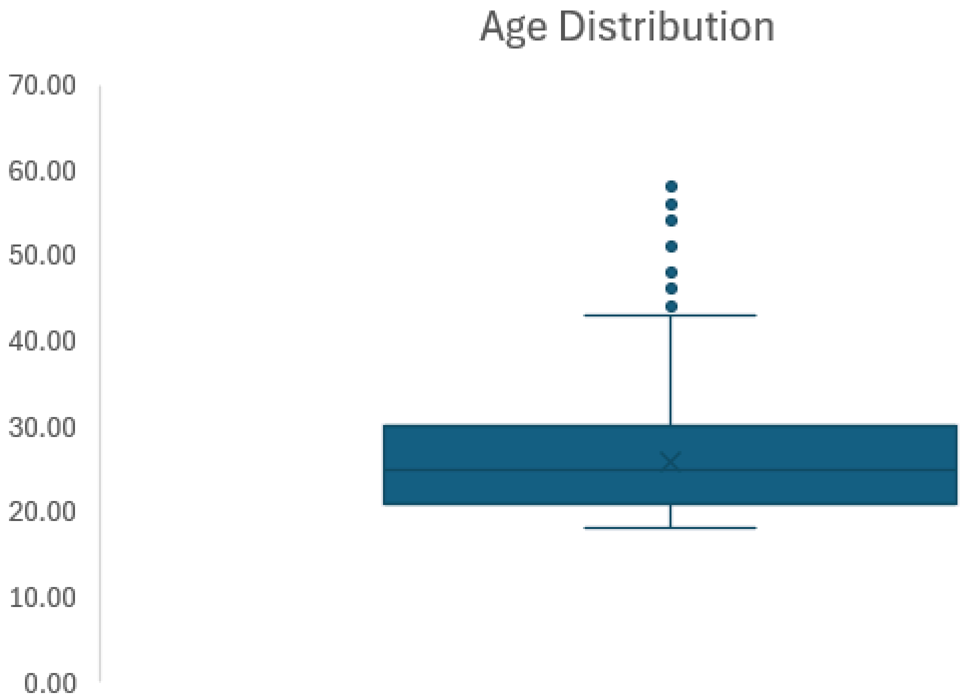


- Based on the analysis, there are **15529 (56%) male respondents** and **12341(44%) female respondents**. The dataset is skewed to be male respondents.



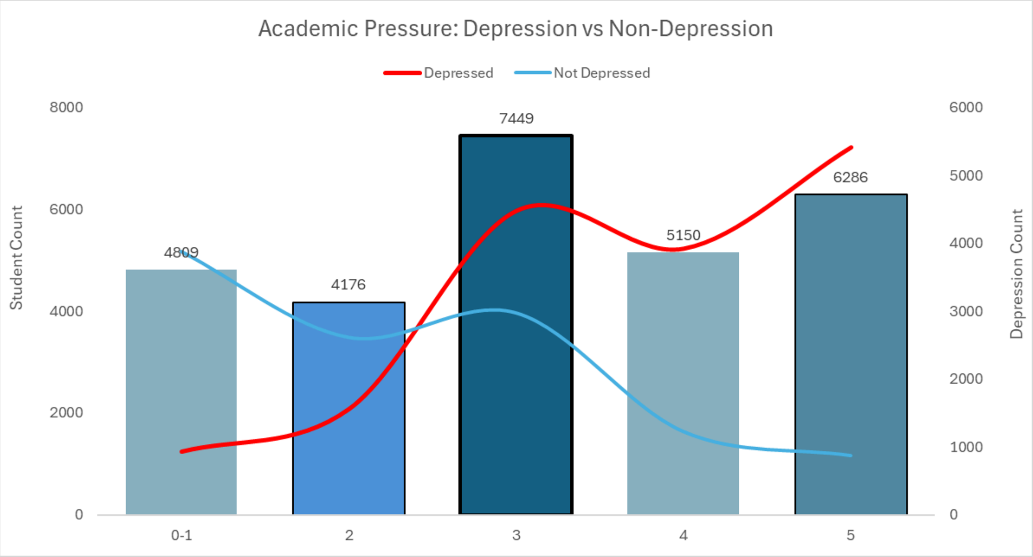
- Despite the skew towards male respondents, the rate of depression within each gender is consistent, with males having **59%** classified as depressed and females having **58%** classified as depressed, having only a 1% difference.

Age:



- The age distribution is concentrated between ages **21(Q1)** and **30(Q3)**. This means that the dataset represents traditional undergraduate and postgraduate students, the **median age at 25** reflects the standard age population distribution. There are also **7 outliers ranging from (44-58)** years old, above the **upper bound of 43**. Instead of treating them as errors/outliers, they are classified as **“Mature Students”** instead. Since age has a **negative correlation** with depression (r≈-0.23) and correlates with higher emotional resistance, removing the outliers may artificially skew the dataset towards a younger demographic that is more depression prone, resulting in a possible biased model.

Academic Pressure:



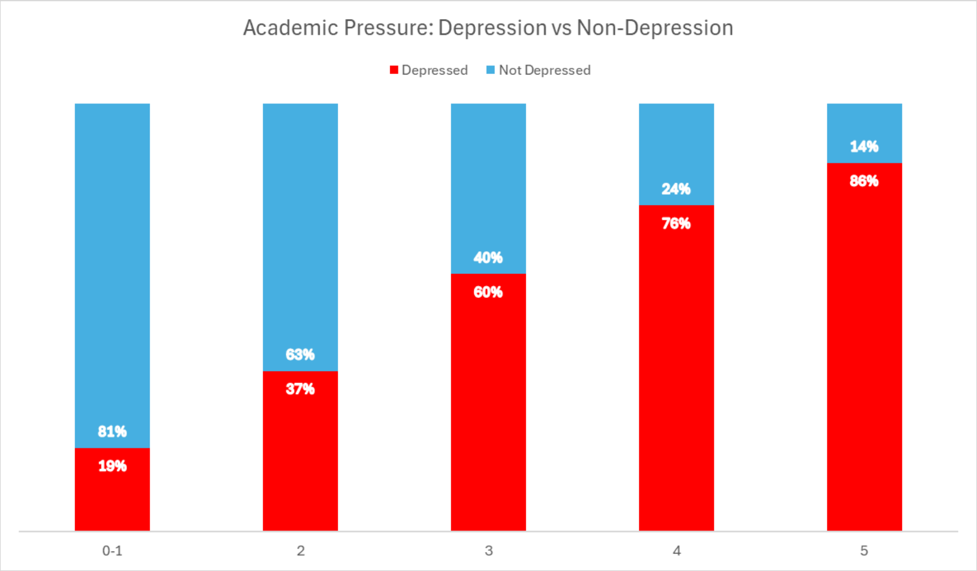
- Due to how small the total number of students that answered “0” was, for visualization purposes, 0 and 1 responses were aggregated.

Univariate Analysis:

- The highest frequency of students is from response “3” (Moderate Pressure), with 7,449 students.
- The second-highest frequency is from response “5” (High Pressure) with 6,286 students. This indicates that there is a slightly higher left skew of students under high academic pressure, based on the dataset.

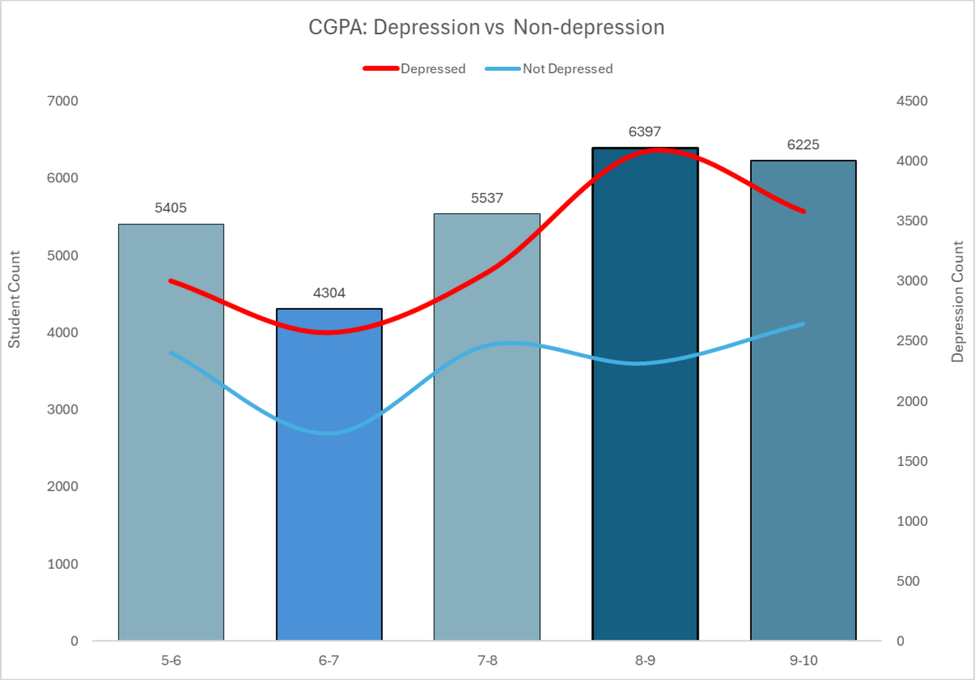
Bivariate Analysis:

- As Academic Pressure increases, the count of Non-Depressed students rapidly decreases, while the count of “Depressed” students rapidly decreases.
- The threshold seen at response “2”, where the probability of being non-depressed statistically surpasses the probability of being depressed, was suddenly overturned at response “3”, jumping up to 60%.



The analysis of Academic Pressure identifies response “3” (Moderate Pressure) as the ‘Tipping Point’, transitioning to a majority of depressed students present than not. While responses “0-1” have a high protection rate(81% non-depressed students), the sudden shift happens where depression suddenly becomes the majority outcome (60%). This trend escalates to response “5”, creating a high-risk environment where nearly 9 out of 10 (86%) are classified as depressed, confirming that Academic Pressure is a strong indicator for mental health decline.

Grade Point Average (CGPA):

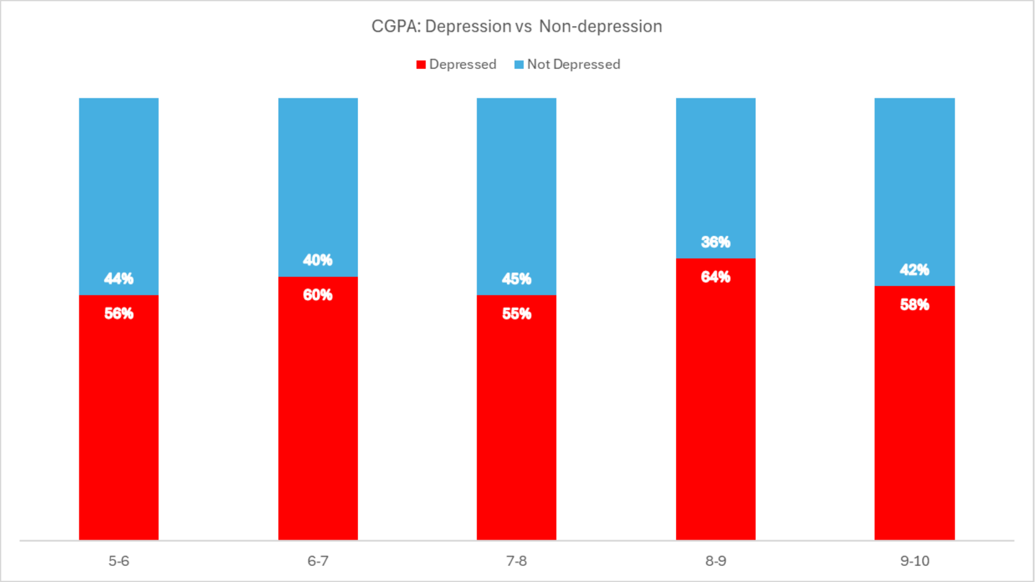


Univariate Analysis:

- The highest frequency is the 8.0 - 9.0 CGPA bracket at 6,397 students, followed by the 9.0 - 10.0 bracket at 6,225 students
- The data is slightly left skewed, meaning the lower performance tiers of 5.0 - 7.0 represent a minority of the dataset, indicating that a lot of the students have higher grades than not.

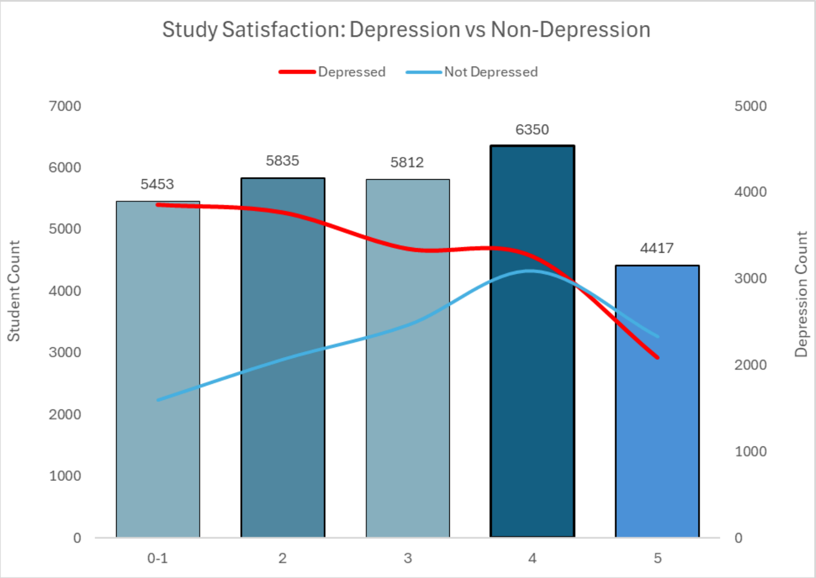
Bivariate Analysis:

- Based on the trendlines, regardless of grade bracket, the volume of depressed students consistently exceeds the volume of non-depressed students
- The fluctuation of the trend lines mirrors the number of students per grade bracket, indicating that depression is present across all grade bracket groups rather than a specific one.



Based on the analysis of CGPA, the thinking that academic competence means a student is not likely to have depression is debunked. Despite the group bracket of 8.0 - 9.0 having the highest frequency and being regarded as a high-achieving grade, it also has the highest percentage of depressed students. High grades do not separate students with depression; rather, the pressure to maintain such a high standing throughout all grade-giving activities. This aligns with previous findings that objective academic success is closely linked to mental health.

Study Satisfaction



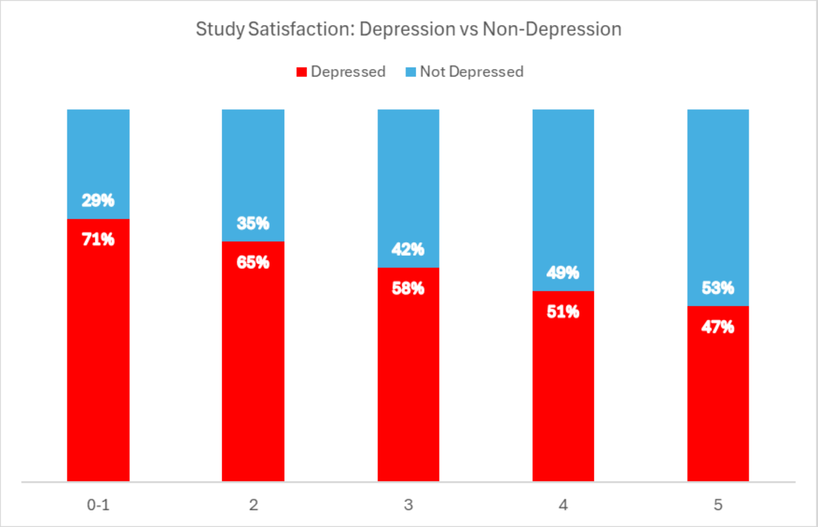
- Due to how small the total number of students that answered "0" was, for visualization purposes, 0 and 1 responses were aggregated.

Univariate Analysis:

- The highest frequency is response "4" at 6,350 students meaning a lot of students are quite satisfied in terms of studying
- The lowest frequency is response "5" at 4,417 students, suggesting that while students do not actively hate their current coursework, few are fully satisfied.

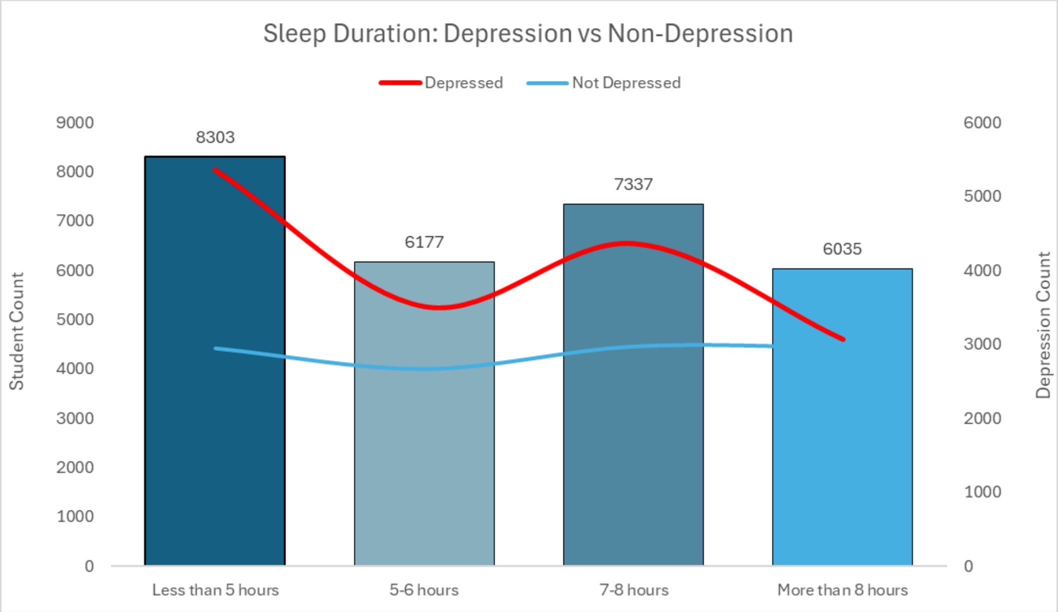
Bivariate Analysis:

- The trend line illustrates an inverse relationship towards depressed students, where as study satisfaction increases, the count of depressed students declines, while the count of non-depressed students increases.
- The intersection point shown between response "4" and "5" marks the transition where non-depressed students statistically outnumber depressed students, indicating that as students are satisfied with their coursework, the more likely a student is to be non-depressed.



Based on the analysis of Study Satisfaction, it can be statistically regarded as a good protective factor against depression due to the linear reduction in depression as satisfaction increases by 24% (From 71% to 47%). Critically, response "5" is the only group where non-depressed students are the majority (53%). We can say that focusing on increasing academic engagement and study satisfaction could just be as effective as focusing on reducing academic pressure and stress (through course relevance and reducing workload).

Sleep Duration



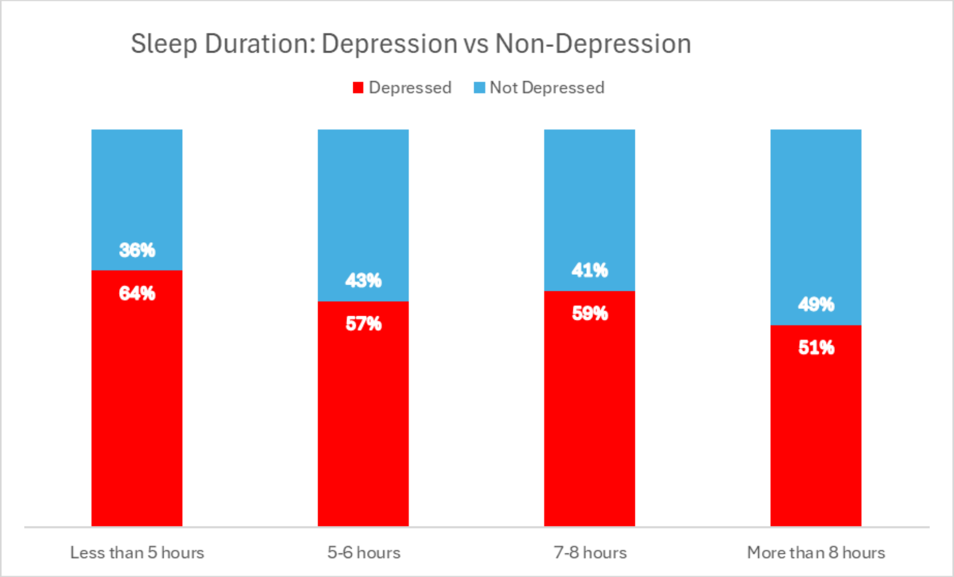
- The ‘others’ column has been omitted due to only having 10 instances and a lack of information on what it means.

Univariate Analysis:

- The highest frequency is “Less than 5 hours” at 8,303 students, while the lowest frequency is “More than 8 hours” at 6035 students.
- There seems to be a right skew in students’ sleeping patterns, meaning that the dataset consists of a significant number of students under sleep deprivation.

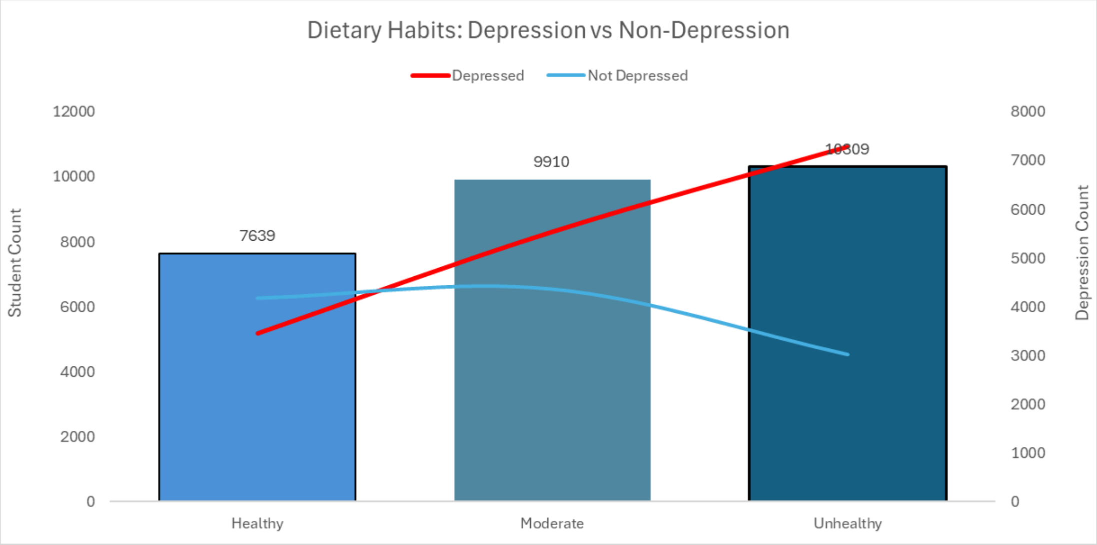
Bivariate Analysis:

- The line trend indicates that the volume of depressed students is higher than non-depressed students across all sleep schedules



Based on the analysis of Sleep Duration, we can see why the correlation is lower than expected (r≈-0.09). While sleep deprivation of <5 hours has the largest amount of depressed students at 64%, up to the optimal sleeping range for teenagers >8 hours reduced to 51%. The 13% reduction indicates that while having the appropriate amount of sleep needed to counteract mental health decline, it is insufficient to counteract the negative impact of other features, such as academic pressure or financial stress. Therefore, we were able to identify why sleep duration is not a strong indicator for depression in this dataset. A well-rested student is statistically less likely to be depressed, but remains a high risk if other critical indicators are present.

Dietary Habits

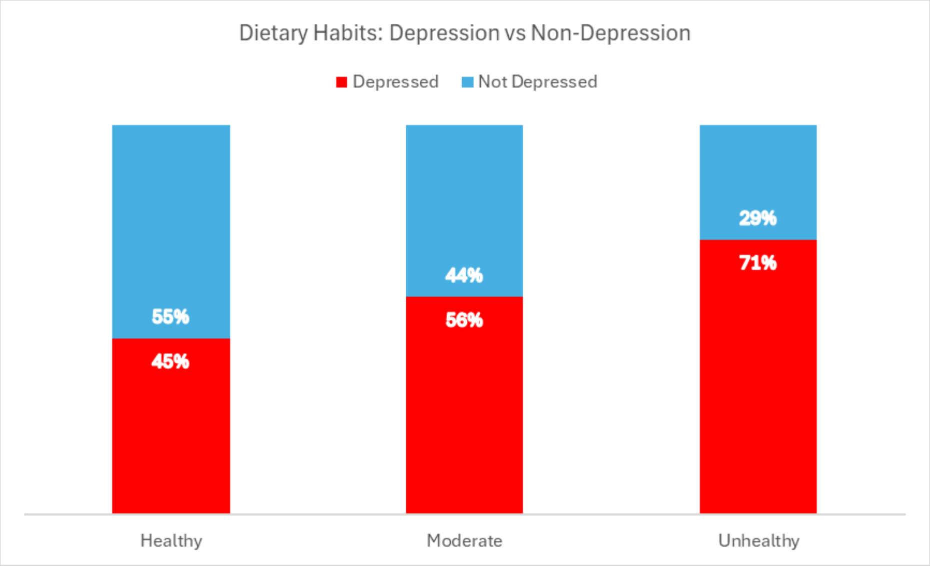


Univariate Analysis:

- The highest frequency of students' dietary habits is "Unhealthy" at 10,309, which is a concerning amount of students having poor physical health maintenance
- The lowest frequency of students' dietary habits is "Healthy" at 7,639, indicating that having a nutritious diet is a challenge for the majority of students in this dataset

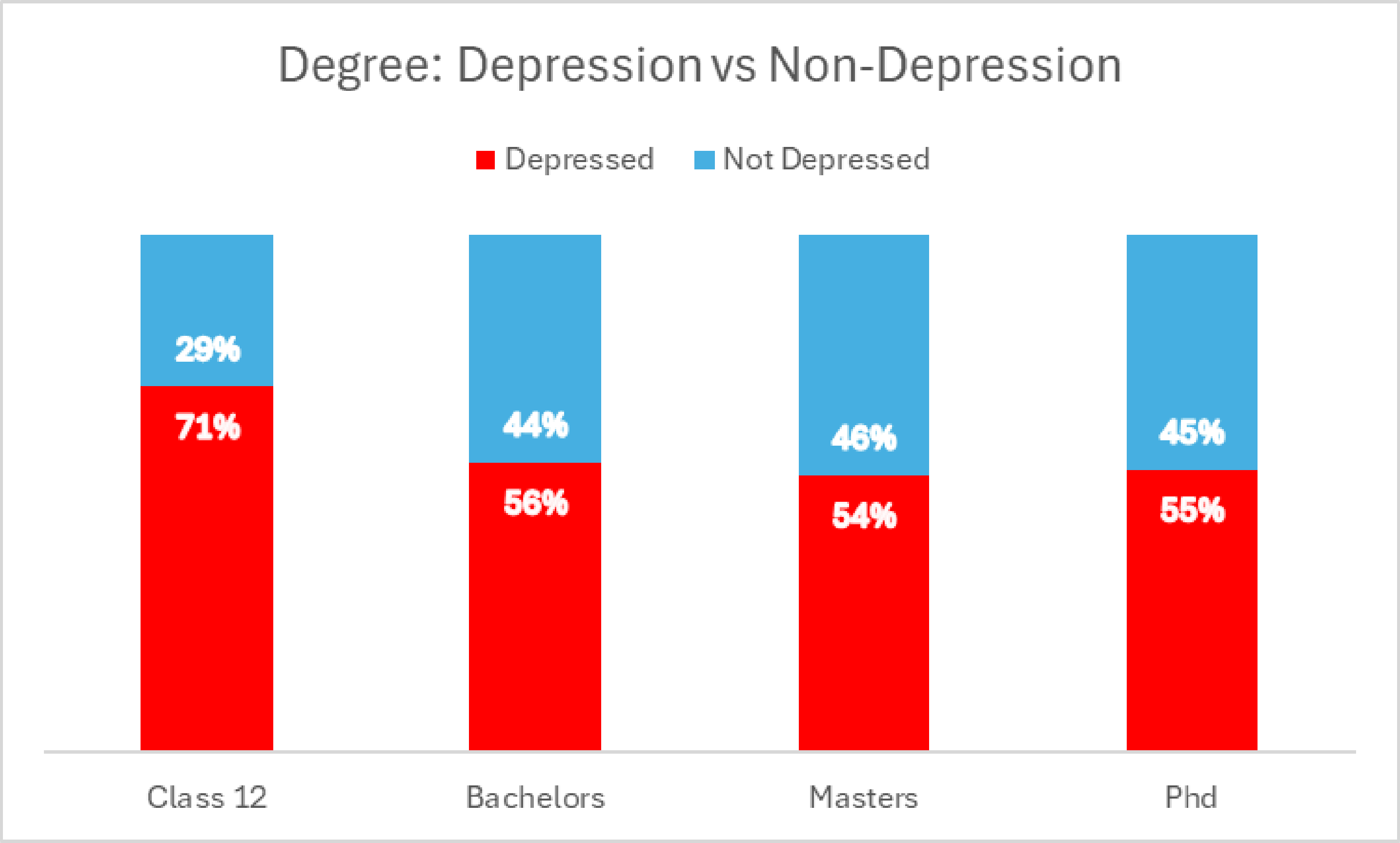
Bivariate Analysis:

- In the "Healthy" group, the number of non-depressed students is higher than depressed students. This shows that good nutrition can be a useful indicator of where students could be regarded as non-depressed.
- In the "Unhealthy" group, the gap between depressed and non-depressed students widens significantly, suggesting that dietary habits have a linear relationship to depression and an inverse relationship to non-depression



Based on the analysis of Dietary Habits, it could be regarded as a fundamental factor in terms of a student's mental health. The 26% increase in depression from Healthy to Unhealthy(45% to 71%) indicates that Dietary Habits correlate well with students' mental well-being. Additionally, the "Healthy" category is also one of the few groups where non-depress is the majority (55%). This shows that having a nutritional and well-balanced diet could be a key factor to bolster student resilience against depression

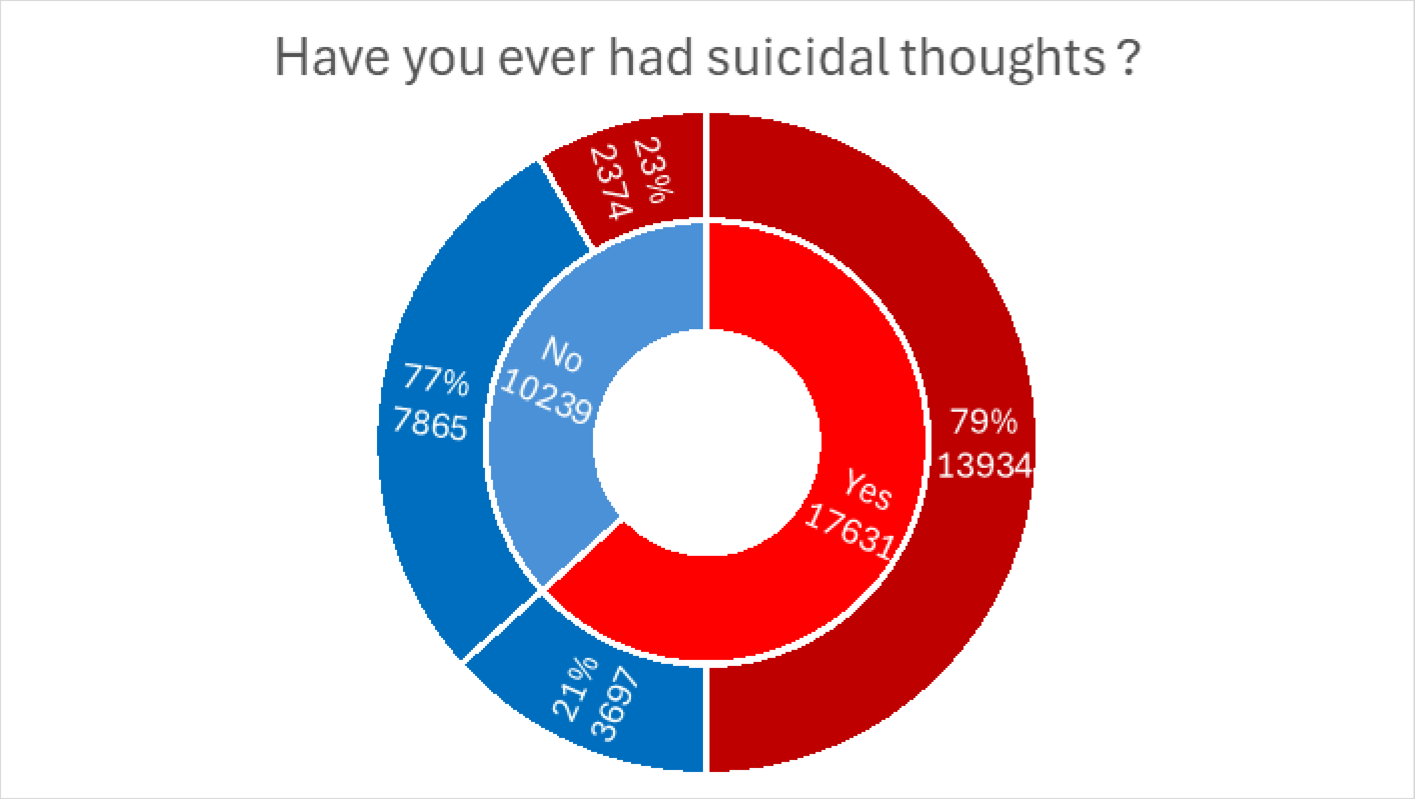
Degree



Due to a severe imbalance in the data distribution of the degree feature, a stacked bar chart is used instead. Based on the analysis, the group "Class 12" has the highest percentage of depressed students at 71%. A possibility due to the huge comparison from other groups is that these students may face unique types of stressors, such as life-altering career decisions, university entrance pressures, and changing to a college environment, which makes them vulnerable to stress and negative thoughts.

Once students have matured and entered university (Bachelors, Masters, PhD), the depression risk appears to stabilize, consistently at 54% - 56%. The difference between students in Class 12 and those who entered university suggests that the anticipation of higher education is more mentally taxing than participating in it. Based on previous findings, the negative correlation between Age and Depression could mean that students of older age have better coping mechanisms and emotional maturity, which explains the fact that depression levels stabilize during a student's university. It confirms that the most dangerous timeframe for a student's mental health to be at risk is not the highest level of education, but rather the anticipation of doing so.

Have you ever had suicidal thoughts ?

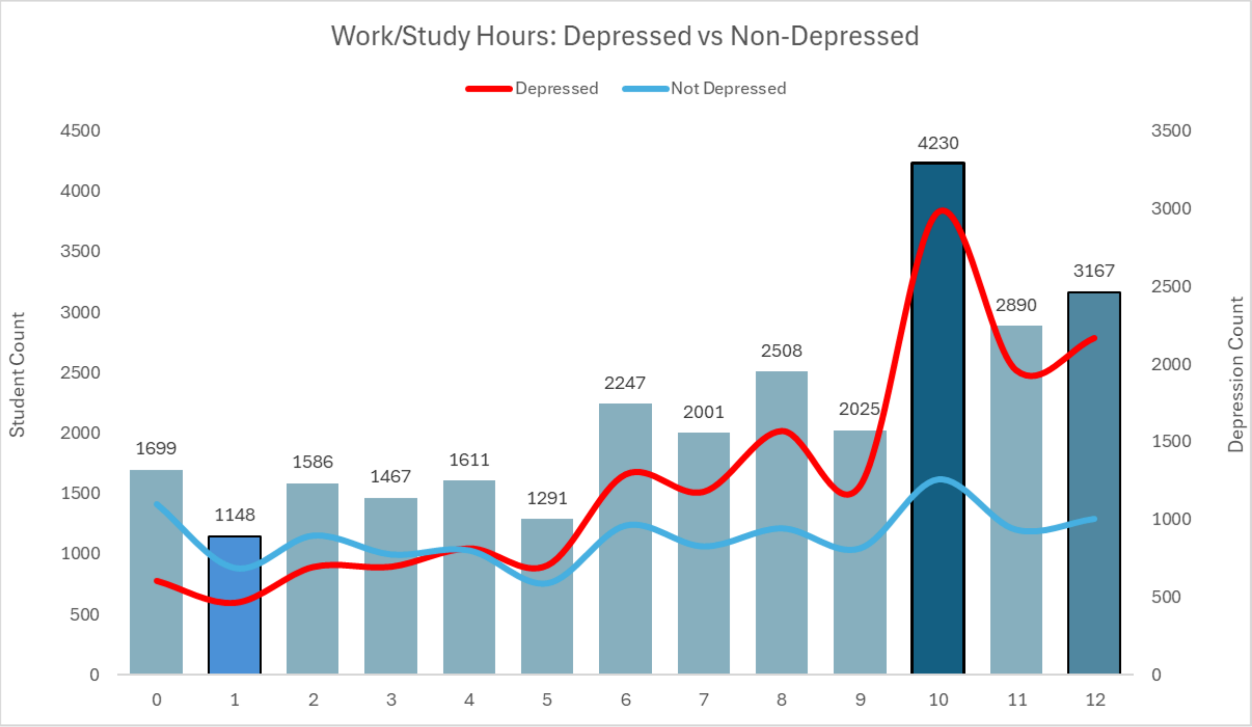


Based on the analysis of Suicidal Thoughts, 63% of the students experience suicidal thoughts, while 37% do not. Among the students who admitted to having suicidal thoughts, 79%(13,934) were classified as Depressed, which confirms that a possible link is created when a student experiences suicidal tendencies; they are more likely to exhibit depressive behavior and vice versa.

A surprising find is that among those who admitted to having suicidal thoughts, 21%(3,697) were not classified as depressed. This suggests that students like these may have suicidal thoughts due to other features like academic pressure or financial stress, rather than chronic clinical depression. Despite these students not having depressive behavior, they are still just as important since there is a possibility that they are actively suicidal but do not meet the criteria of being depressed.

Another surprising find is that those who admitted not having suicidal thoughts, 23% (2,374), were classified as depressed. This means that a minority of the students in the dataset likely manifest depression, not because of suicidal thoughts, but other factors like sleep deprivation or dietary neglect.

Work/Study Hours

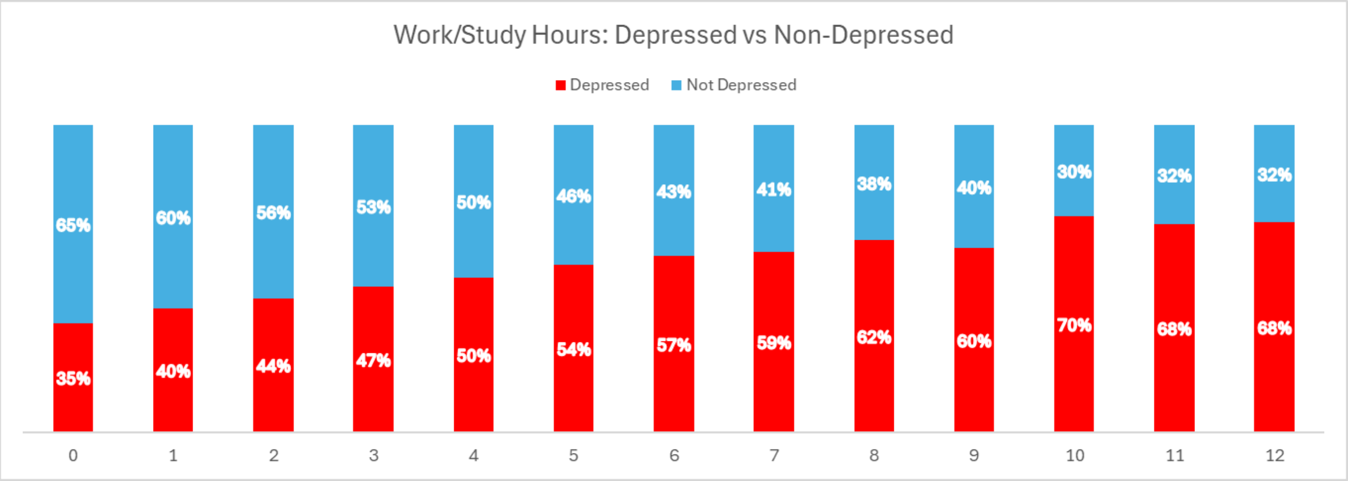


Univariate Analysis:

- The highest frequency is response “10” at 4,230 students, while the second-highest is response “12” at 3,167 students.
- The lowest frequency is response “1” at 1,148 students. The graph shows a left skew, indicating that a large portion of students maintain a demanding daily schedule of schoolwork, likely leaving minimal time for rest or recreation.

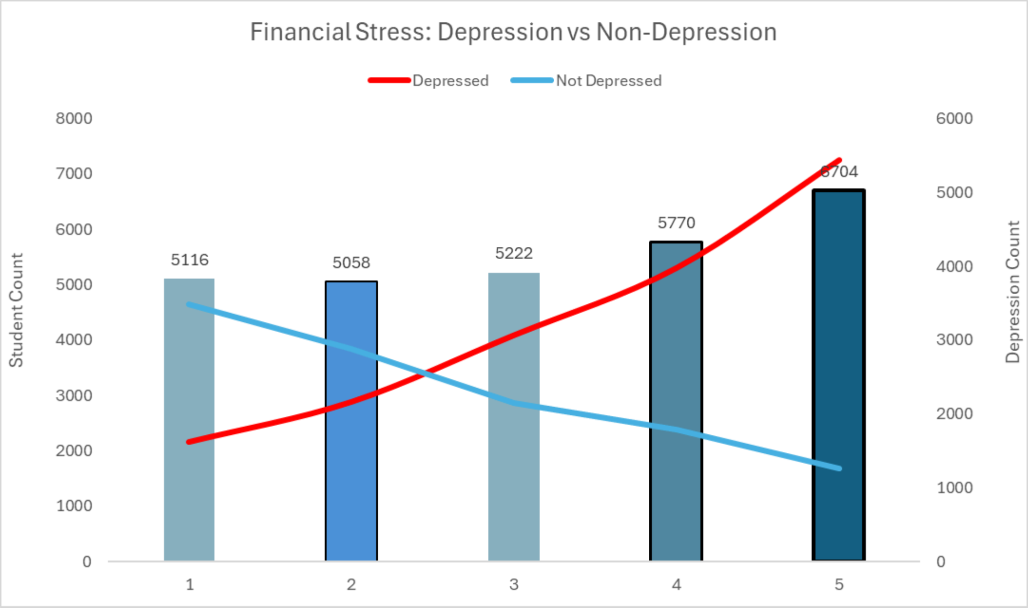
Bivariate Analysis:

- At around groups 0 - 3 hours, the amount of non-depressed count exceeds the depressed counts, showing that students with minimal workload are safer from depressive behavior
- The intersection at group 4 and beyond shows that the increasing trend of depressed students increases while the non-depressed students stagnate.



Based on the analysis of Work/Study Hours, this confirms that there is a linear positive relationship between workload duration and mental health decline. The data identifies 4 hours as the threshold where depression probability shifts from minority to majority (50%). The increase of depression from 35% to a peak of 70% at 10 hours provides strong support that overloading on academic workload increases depressive behavior.

Financial Stress

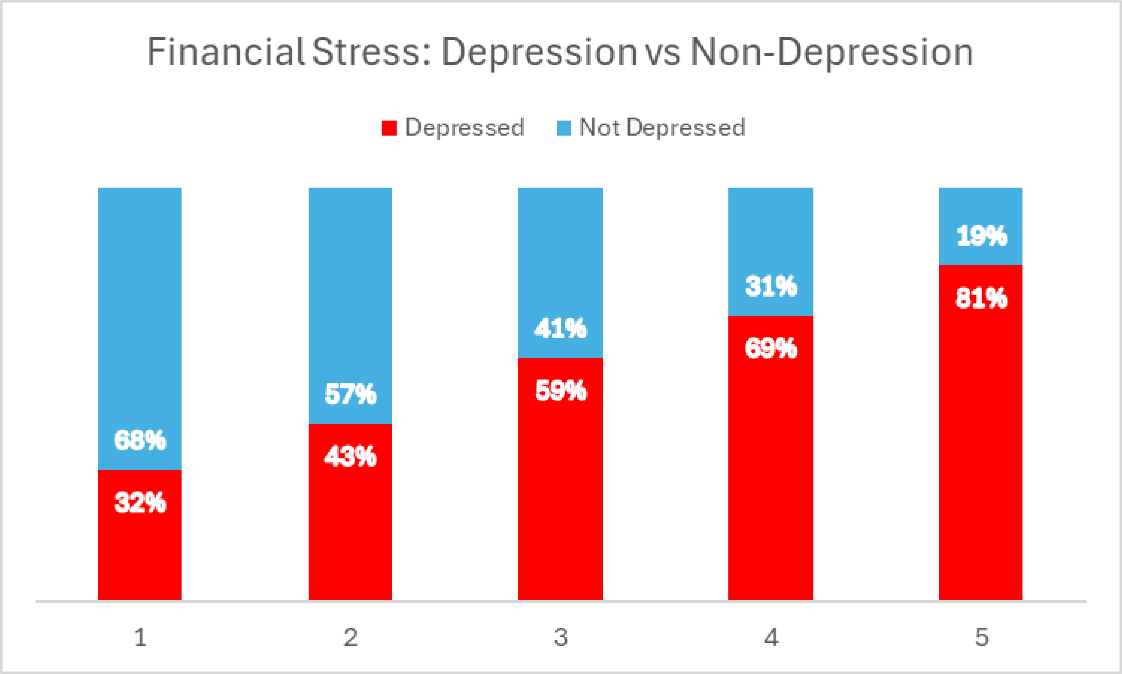


Univariate Analysis:

- The highest frequency is response “5” at 6,704 students and the second-highest is response “4” at 5,770 students. This shows that the distribution is slightly left skewed with more of the students experiencing higher levels of financial stress

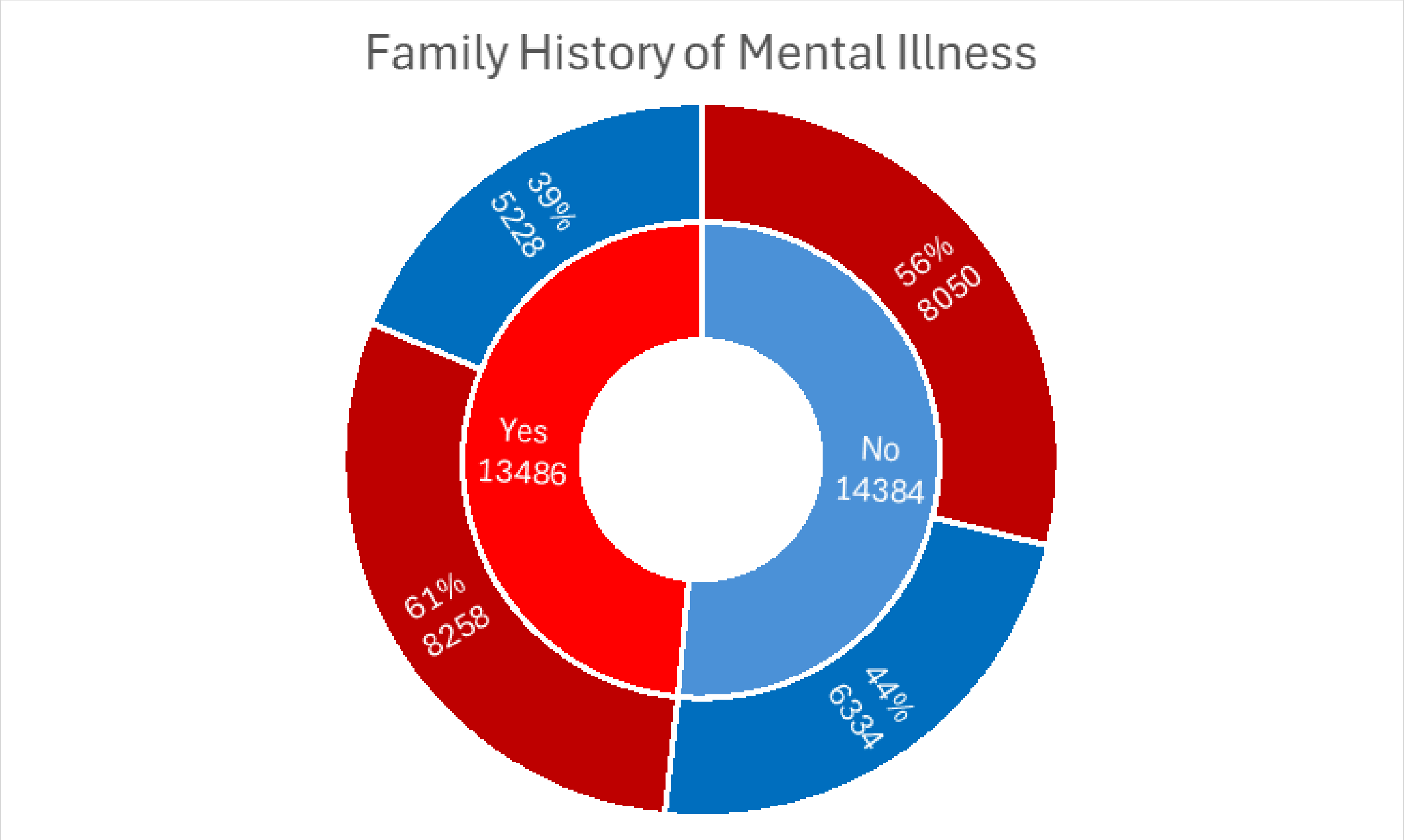
Bivariate Analysis:

- At responses “1” and “2”, the amount of non-depressed students exceeds the number of depressed students. However, an intersection between lines happens between responses “2” and “3”. Beyond this point, the majority will be depressed students
- The relationship between Financial Stress and Depression is linear, while Non-depression is inverse. This indicates that as students' financial stress increases, so does the possibility of depressive behaviors and vice versa.



Based on the analysis of Financial Stress, it has a strong indicator for students if they possess depressive behavior, and has nearly the same severity as Academic Pressure. Financial Stress has a strong positive relationship with Depression, and every incremental increase is more than 10% (11% - 17%). The massive increase of 49% (32% to 81%) means that a student experiencing financial burden could be experiencing a decline in their mental health.

Family History of Mental Illness



Based on the analysis of Family History of Mental Illness, 52% (14,384) of students said no and 48%(13,486) of students said yes. While the number of students who admitted yes is 61%, its counterpart is only 56% which is only a small difference. This indicates that the students in this dataset can still be depressed whether or not the person's family has a history of mental illnesses. This suggests that environmental factors are potent enough to act as primary causes of mental health decline rather than genetics from family, specifically in terms of depression.

13.4 Findings Summary

The analysis reveals that the strongest predictors that correlate with depression are Suicidal Thoughts, Academic Pressure, and Financial Stress (r≈0.36 to 0.55). Based on the correlation between these 3 features(r≈0.15 to 0.26), it seems that there is a possibility of multicollinearity due to how the severity of the features looks nearly identical when compared.

We have learned that although depressed students have slightly higher grades than non-depressed students (7.68 > 7.62), they are statistically more pressured(3.69 > 2.36), stressed (3.58 > 2.52), overworked (7.81 > 6.24 hours), and have lower study satisfaction(2.75 < 3.22) suggesting that this demographic academic success is fueled by pressure and stress rather than genuine engagement.

Non-factors, which are CGPA(r≈0.02) and Sleep Duration(r≈-0.08). CGPA depression levels revolve around 55%-64% showing us that high grades do not separate students with depression because it creates the pressure to keep a high standing across the entirety of their school life. Therefore, a high-performing student is just as likely to be depressed as a low-performing one. Sleep Duration depression levels revolve around 51% - 64% showing us that having poor sleep does, in fact, have an impact on mental decline. Both were regarded as Non-factors due to not having enough impact.

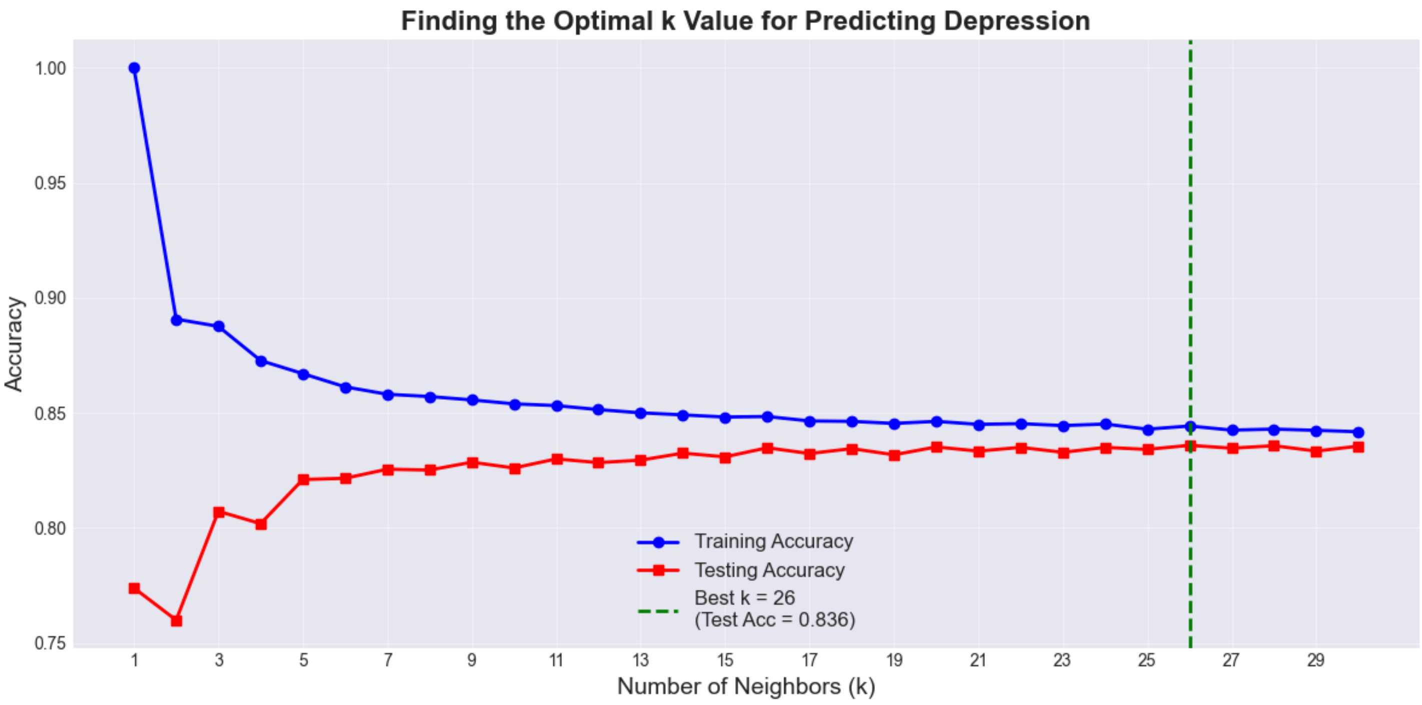
Age has a negative relationship to Depression(r≈0.23), whereas students' age, their tendency to have depressive behaviors decrease. A logical explanation can be that as they grow older, their emotional maturity and coping mechanisms have improved, which makes them more tolerant compared to younger students.

Students who have a higher percentage of depression came from "Class 12", around 71% due to the possibility that the anticipation of a higher education could be more mentally taxing and stress-inducing compared to participating in it.

From CGPA and Work/Study Hours, the highest frequency came from those who have higher grades (8.0 - 9.0 at 6,397) and study hours (10 at 4,230). However, both also have the highest percentage of depression rate of 64%(CGPA) and 70%(Work/Study Hours), indicating that having academic success by overworking and getting high grades imposes severe mental decline, resulting in possible burnout and depression. It is important to learn to be efficient while also taking care of yourself in the process.

13.5 Model Performance

13.5.1 Comparative Performance Evaluation



k-value	Training Accuracy	Testing Accuracy
1	1.000	0.7739
2	0.8907	0.8072
3	0.8876	0.8072
4	0.8727	0.8018
5	0.867	0.821
6	0.8613	0.8215
7	0.8581	0.8255
8	0.8571	0.8251
9	0.8556	0.8285
10	0.8539	0.826
11	0.8531	0.83
12	0.8514	0.8283
13	0.85	0.8294
14	0.8491	0.8325
15	0.8482	0.8309
16	0.8484	0.8348
17	0.8465	0.8323
18	0.8463	0.8344
19	0.8454	0.8318
20	0.8463	0.8352
21	0.845	0.8334
22	0.8453	0.835
23	0.8444	0.8328
24	0.8451	0.8350
25	0.8429	0.8341
26	0.8443	0.8359
27	0.8425	0.8346
28	0.8429	0.8357
29	0.8424	0.8334
30	0.8418	0.8355

Key Findings of Systematic Evaluation

The evaluation of k-values from 1 to 30 produced the following results:

- The highest testing accuracy observed was 0.8359 at k=26, achieving peak testing performance.
- Testing accuracy remained consistently high (between 0.83 and 0.836) across k-values **11 through 30.**
- Performance stabilized after k=11, with minimal fluctuation in testing accuracy.
- Testing up to k=30 confirmed that performance had plateaued, ensuring no further improvement would occur with larger k.

13.5.2 Classification Report

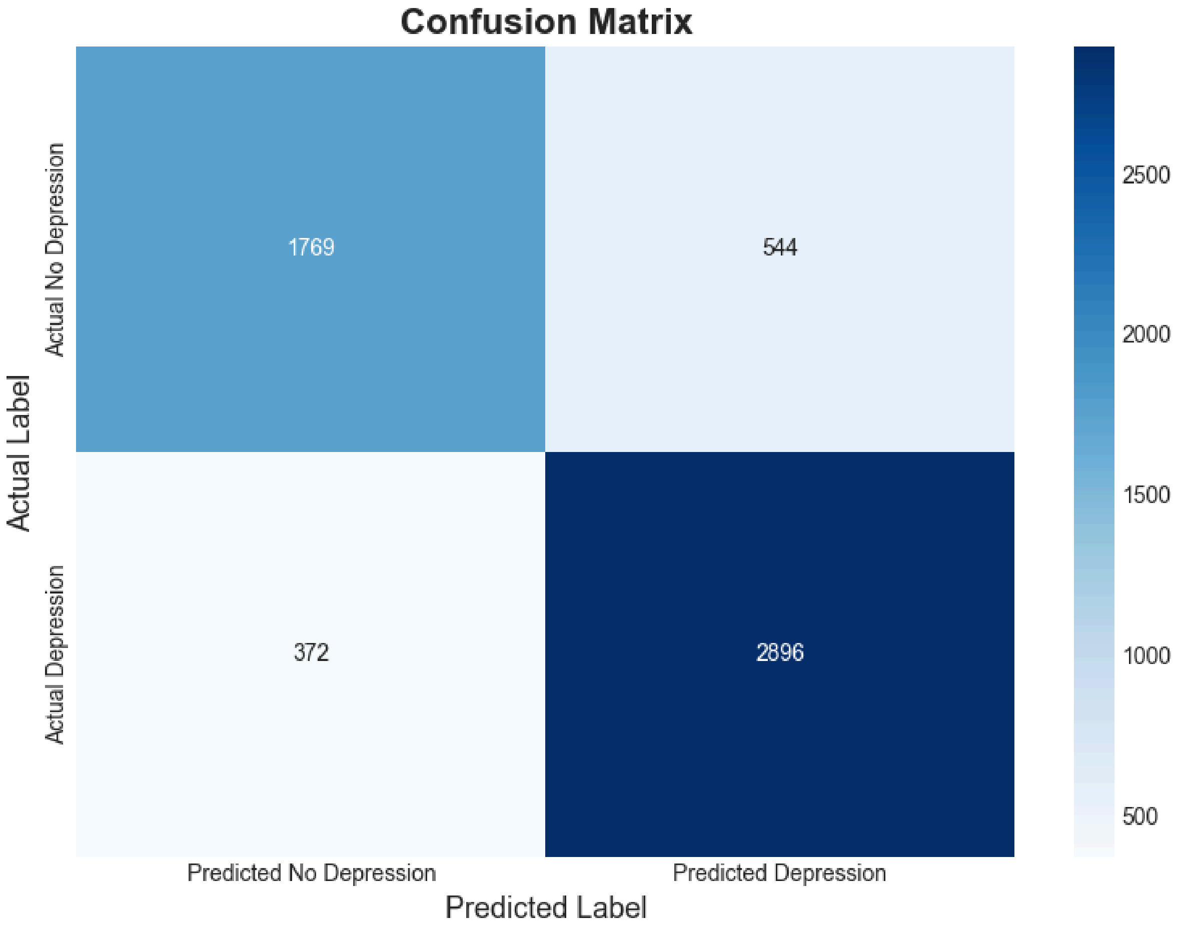
	precision	**recall**	**f1-score**	**Support**
No depression	0.83	0.76	0.79	2313
Depression	0.84	0.89	0.86	3268
accuracy			0.84	5581
Macro avg	0.83	0.83	0.83	5581
Weighted avg	0.84	0.84	0.83	5581

The model achieved an overall accuracy of 0.84, correctly classifying 84% of the 5,581 total instances, creating a strong general classification performance.

Performance by Class

- “No depression” class: Shows balanced precision (0.83) and recall (0.76), resulting in an F1-score of 0.79.
- “Depression” class: Demonstrates high recall (0.89) and strong precision (0.84), yielding an F1-score of 0.86.
- Depression has a higher recall(0.89), precision(0.84), and f1-score(0.86) compared to No Depression due to the majorty of the dataset having Depressed students. Making it a good indicator to see if a student is having depressive behavior but the low recall(0.76) of No Depression may be prone to misclassify students as Depressed.

13.5.3 Confusion Matrix



Metrics:

True Positives (TP)	2896	Correctly predicted depression
True Negatives (TN)	1769	Correctly predicted no depression
False Positives (FP)	544	Predicted depression but actually no depression
False Negatives (FN)	372	Predicted no depression but actually depression

The model demonstrates a stronger ability to correctly identify depression cases than non-depression cases, consistent with the earlier classification report showing higher recall for the depression class. The number of false positives (544) indicates some over-prediction of depression among actual non-depression cases, which may reflect the model's tendency toward higher sensitivity for detected depression.

10. Methodology – Research Design

10.1 Type of Research

This study employs a **Quantitative Research Design**. While the dataset comprises a combination of numerical and categorical variables (e.g., financial stress, dietary habits), the analytical framework treats these features quantitatively. Categorical variables are subjected to encoding and frequency analysis to establish statistical patterns and trends.

10.2 Research Approach

The study utilizes a dual approach consisting of Descriptive and Exploratory methods:

- **Descriptive Analysis:** Focuses on summarizing the central tendencies and distribution of student depression data to establish a baseline profile.
- **Exploratory Analysis:** Investigates correlations and structural relationships between independent variables and the dependent variable (depression status).

10.3 Procedural Framework

The research workflow follows a step-by-step process:

- **Data Acquisition:** Retrieval of the dataset.
- **Data Preprocessing:** Handling of missing values, outliers, encoding, and cleaning of raw data
- **Exploratory Data Analysis (EDA):** Statistical visualization through univariate, bivariate, and multivariate visual analysis.
- **Predictive Modeling:** Implementation of the K-Nearest Neighbors (KNN) algorithm.
- **Evaluation:** Assessment of model accuracy and recommendations needed based on model performance.

11. Methodology – Data Sources and Selection Criteria

11.1 Data Source and Description

Secondary data was obtained from Kaggle, specifically the "Student Depression Dataset" authored by **Shodolamu Opeyemi (2024)**. The dataset contains **27,901 records** across **18 columns**, featuring a mix of continuous and categorical variables aimed at analyzing and predicting depression levels in student populations. Key features include:

- **Demographics:** Age, Gender.
- **Academic Metrics:** Academic Pressure, CGPA, Study Satisfaction, Work/Study hours, Degree
- **Lifestyle Factors:** Sleep duration, Dietary habits, Financial Stress
- **Clinical History:** Suicide Thoughts, Family History of Mental Illness
- **Target Variable:** Depression

11.2 Selection and Exclusion Criteria

To ensure the integrity of the analysis, strict inclusion and exclusion criteria were applied:

- **Relevance:** Only features and records directly about student demographics and behaviors will be retained to minimize noise and irrelevant variance.
- **Data Completeness:** Features exhibiting **>70% missing values** were excluded from the study to prevent analysis bias due to sparsity.
- **Target Population:** Records are filtered to ensure the exclusive analysis of student subjects, aligning with the research objectives.

11.3 Dataset Context and Validity

The study utilizes a synthetic dataset. While this dataset is not drawn from a specific local population (Low Fidelity), it serves as a robust research tool for identifying statistical patterns and testing the efficiency of the KNN algorithm without compromising patient privacy. The use of synthetic data ensures ethical compliance regarding sensitive mental health information while allowing for a "Clean Lab" environment to validate the research model.

12. Methodology – Data Analysis and Synthesis

12.1 Data Preprocessing and Cleaning

Before analysis, the data underwent rigorous cleaning procedures:

- **Feature Reduction:** Variables with no theoretical or statistical correlation to the target variable were dropped.
- **Handling of Irregularities:** Outliers and missing values constituting <1% of total records were removed to maintain data quality without significantly reducing sample size.
- **Transformation:** Categorical data types were converted (encoded) into numerical formats to facilitate correlation analysis and distance-based calculations required by the KNN model.

12.2 Data Analysis Techniques

Exploratory Data Analysis (EDA) was conducted using a hierarchical approach:

- **Key Performance Indicators (KPIs):** High-level metrics were generated using Pivot Tables to provide an immediate understanding of the dataset's composition.
- **Univariate Analysis:** Frequency tables and histograms were employed to examine the distribution and skewness of individual features.
- **Bivariate Analysis:** Cross-tabulation and clustered bar charts were used to assess the relationship between independent features and depression status.
- **Multivariate Analysis:** Combo charts were utilized to visualize complex interactions between three or more variables simultaneously.

12.3 Tools Used

Data Processing and Visualization: Initial data cleaning, statistical aggregation, and Exploratory Data Analysis (EDA) were conducted using **Microsoft Excel**. This environment was utilized for its robust pivot table capabilities and graphical rendering of univariate and bivariate distributions.

Algorithmic Implementation: The **K-Nearest Neighbors (KNN)** predictive model was developed using Python. The libraries used are **scikit-learn** for model architecture, **pandas** for data frame management, and **NumPy** for numerical computations.

12.4 Synthesis and Model Evaluation

After the implementation of EDA and the findings given by the K-Nearest Neighbors(KNN) model, discrepancies between the analytical insights and the model's predictions are analyzed to understand the reasoning behind the performance, leading to final recommendations for future research

13.Results – Key Findings
13.1 KPIs
13.2 Univariate Analysis
13.3 Bivariate Analysis
13.4 Multivariate Analysis
13.5 Findings Summary
13.6 Model
13.7 Accuracy Evaluation