Volume 3, Issue XI, November, 2024, No. 52, pp. 689-708
Submitted 23/11/2024; Final peer review 27/12/2024
Online Publication 31/12/2024
Available Online at http://www.ijortacs.com

DATA-DRIVEN APPROACHES TO EDUCATIONAL EQUITY: PREDICTING AND RESOLVING GENDER-BASED ACADEMIC DISPARITIES

[1*]Omeye, Emmanuel C., [2]Anyaragbu Hope U.
[1,2]Department of Computer Science, Tansian University, Umunya.
E-mail:[1*]emmanuel.omeye@tansianuniversity.edu.ng,[2]anyaragbu.hope@tansianuniversity.edu.ng
[1]*Corresponding Author E-mail: emmanuel.omeye@tansianuniversity.edu.ng
Tel: +2348064116676

Abstract

This research addresses gender disparity in academic achievement through an innovative approach, integrating predictive analytics and strategic mitigation strategies. Employing Python and Spyder as the primary programming tools, the study utilizes the Structured Systems Analysis and Design Method (SSADM) to systematically explore the academic landscape. At its core, the research implements a Stack Ensembled Data Analytics technique, combining five models—K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest. This ensemble approach enhances predictive accuracy and robustness, offering a nuanced understanding of factors contributing to gender-based academic disparities. Machine learning libraries, including Pandas, NumPy, StandardScaler, and LabelEncoder, along with evaluation metrics such as Accuracy Score and Confusion Matrix, contribute to a comprehensive analysis. The models exhibit commendable performance during the evaluation phase, demonstrating the strength of the chosen approach. SSADM ensures a structured process from data collection to model implementation and evaluation. The significance of the research lies in its ability to predict gender-based academic disparities and propose effective mitigation strategies accurately. Leveraging ensemble models unravels intricacies contributing to performance variations, empowering educational institutions with actionable insights for targeted interventions. This project significantly contributes to the discourse on educational equity by introducing a data-driven framework that predicts disparities and facilitates strategic interventions. The successful integration of Python, SSADM, and stack-ensemble models underscores the versatility of the approach, positioning it as a valuable tool for researchers, educators, and policymakers. In the evolving academic landscape, this research offers a timely contribution to global efforts for a more equitable and inclusive educational system.

Keywords: Ethics, Bias, Data Bias, Ethical AI, Algorithmic Transparency, Inclusive Education, Predictive modeling, Machine Learning, Algorithmic bias, Socioeconomic Status, Intervention, Equity, Fairness, Gender Disparity in academic achievement.

1. INTRODUCTION

Gender disparity in education refers to the unequal access and opportunities for education between males and females. It involves differences in enrolment rates, completion rates, quality of education, and subject choices (Kumara et al., 2020). It is a persistent and multifaceted issue that has garnered significant attention in education. It refers to the unequal distribution of academic performance and opportunities between genders, focusing on the underrepresentation of females in STEM (Science, Technology, Engineering, and Mathematics) fields and the overrepresentation of males in certain subjects (Miao S., 2023). This project, titled "Predicting and Mitigating Gender Disparity in Academic Achievement: A Machine Learning Approach," aims to address this issue by leveraging the power of machine learning to predict gender disparities and develop strategies for mitigation.

Significant progress has been made in reducing gender disparity in education in recent decades. However, it remains a major challenge, particularly in developing countries. According to UNESCO, there are still 129 million girls out of school worldwide, including 32 million of primary school age, 30 million of lower-secondary school age, and 67 million of upper-secondary school age. Girls are more likely to be out of school than boys in all regions of the world, but the disparity is greatest in Sub-Saharan Africa. In countless corners of the world, a multitude of hurdles stand in the way of girls receiving the education they so rightfully deserve. Among these obstacles, child marriage casts a dark shadow, snatching away their childhoods and prospects for learning. A staggering 650 million women alive today were once forced into marriage before they even reached the tender age of 18, a practice that truncates their dreams.

To compound these woes, gender-based violence often becomes an inescapable companion for girls. Trapped in a world where they are more likely to be victims of such cruelty, the simple act of going to school becomes a daunting ordeal. Fear gnaws at their aspirations, shackling them in an unending cycle of insecurity (Smith et al., 2022).In some societies, deeply entrenched cultural norms and practices serve as formidable barriers to girls' education. Daughters are consigned to domestic duties, sacrificing their right to learn as they are burdened with household chores and the care of younger siblings. The weight of tradition anchors their potential and dreams, rendering them tethered to the home. The scarcity of quality education exacerbates this dire situation. Gender bias pervades schools, where female teachers are a rarity, facilities are often subpar, and curricula perpetuate stereotypes. The very system designed to impart knowledge becomes a breeding ground for gender disparity.

Economic disparities only deepen the chasm of inequality. Poverty forces families to make excruciating choices, oftentimes sacrificing their daughters' education for the perceived greater good. Regrettably, girls are all too frequently sidelined in favor of their male counterparts. In a world that should be built on equality, these inequities continue to stifle the hopes and aspirations of countless girls (Ritz et al., 2023). The shadow of gender disparity in education casts a long and sobering spell, leaving a trail of detrimental consequences in its wake that extend from the individual to communities, and to the very fabric of societies. For individuals, especially young girls, this disparity is a cruel and stifling force. Denying them access to education means truncating their potential and hindering their ability to reach their economic aspirations. A world that should be open to them becomes limited, and their full participation in society is stunted. Poverty, ill health, and premature marriages loom menacingly on their path.

But the repercussions of this disparity ripple beyond the individual, echoing throughout communities. In areas where gender inequality in education is high, the prospect of sustainable development grows dimmer. The engine of progress sputters as half of its potential workforce is held back, and the promise of collective advancement remains elusive (Lee & Chang, 2023). Zooming out even further, at the societal level, the spectre of gender inequality casts a shadow over the prospects of peace and prosperity. In societies marked by pronounced gender disparities, the chances of harmony grow faint, and the road to prosperity becomes treacherous. The path toward shared success is marred by the burdensome weight of inequality. Machine learning plays a pivotal role in studying and addressing gender disparity in various domains, including education. In the context of the study on gender disparity in academic achievement, machine learning offers several key roles and contributions.

In the realm of studying gender disparity in academic achievement, machine learning takes center stage, offering a dynamic array of roles and invaluable contributions. Its power lies in its capacity to not only analyze vast datasets but also to recognize intricate patterns, thus unearthing meaningful insights that might otherwise remain concealed. As researchers delve into the complex web of gender disparities in education, machine learning emerges as a potent tool, shedding light on subtle trends and correlations that traditional statistical methods may fail to unveil (Patel et al., 2022).

Within the realm of academia, it can sift through the layers of data, detecting hidden patterns in academic performance, attendance records, participation rates, and an array of other variables that contribute to the stark gender disparities. One of the primary and pivotal roles machine learning assumes in this study is that of a visionary; it crafts predictive models with the power to peer into the future. Drawing from historical data and a multitude of input features, these predictive models are capable of forecasting the extent of gender disparities in academic achievement. For instance, they can estimate the likelihood of female students underperforming in specific subjects or pinpoint the contributing factors to these disparities. These predictive capabilities give educators and policymakers a crucial advantage – the ability to take proactive measures to address gender disparities before they become deeply ingrained within the education system.

A key role that machine learning assumes in this context is the extraction of vital insights from data. These algorithms possess the remarkable ability to autonomously identify the most influential features or variables that underlie gender disparities. This process of feature selection is akin to a beacon, guiding researchers toward key factors that necessitate attention (Stiso et al., 2023). Policymakers and educators, armed with this knowledge, can then focus their interventions precisely on those aspects that wield the most profound influence on gender disparities.

Machine learning's influence goes beyond mere data analysis; it extends into the realm of personalized support and interventions for students. By delving into the individual data profiles of students, encompassing their academic performance, learning styles, and behavioral patterns, machine learning models can construct tailored recommendations. Consider the scenario of a female student who is at risk based on her academic history. Machine learning can step in to suggest personalized resources, support structures, or mentoring programs designed to cater to her unique needs. This personalized approach stands as a testament to the adaptability and precision that machine learning brings to the table. One of the prominent roles that machine learning adopts is in the realm of real-time monitoring and early warning systems. These systems, powered by machine learning algorithms, serve as vigilant guardians, constantly scanning educational data for signs of emerging gender disparities. By analyzing this data in near real-time, they can detect subtle patterns that may hint at impending disparities. Educators and administrators are then promptly alerted, enabling them to intervene before these disparities become deeply entrenched. This proactive approach is instrumental in nipping gender disparities in the bud, preventing them from taking root and hindering the progress of future generations.

Machine learning doesn't stop at alerting to disparities; it also contributes to the formulation of informed policy recommendations. By scrutinizing the intricate relationships between educational policies and gender disparities, machine learning models unearth valuable insights. They identify which policies effectively mitigate disparities and which may require adjustments or enhancements. These evidence-based recommendations serve as guiding stars for policymakers, paving the way for the development of policies that aim to foster gender equality in education.

In essence, machine learning transcends its role as a data analysis tool and emerges as a driving force in the endeavor to promote gender equality in education. It offers real-time insights, informs policy decisions, battles biases and stereotypes, and fosters a culture of continuous improvement. With machine learning as an ally, the pursuit of gender equality in education becomes a dynamic, adaptive, and data-driven journey that holds the promise of creating fair and equitable learning environments for all students.

The paper is organized as follows: Section 2 presents an overview of related works. Subsequently, in section 3, we discuss the methodology. In section 4, we show the system design and implementation. Finally, in section 5, we present the conclusion and future direction.

2. RELATED WORKS

Gender disparity in academic achievement has emerged as a pivotal concern in contemporary educational discourse. In this theoretical review, we delve into the multifaceted aspects of this issue, exploring the factors contributing to gender gaps in academic performance, elucidating the repercussions of such disparities, and elucidating strategies for both prediction and mitigation; OECD (2020). Gender disparities in educational attainment are profoundly influenced by sociocultural, educational, psychological, and familial factors. Sociocultural norms and biases often dictate societal expectations regarding gender roles and education, permeating the choices students make; Ayalon & Shavit (2021). These can affect subjects chosen for study, and access to educational resources may not be equitable across gender lines.

Within the educational environment itself, instructional methods, curriculum design, and classroom dynamics can unwittingly favor one gender over the other, amplifying existing disparities. A phenomenon known as self-selection bias can ensue, where students gravitate towards fields that align with perceived strengths, further perpetuating the disparities.

Psychological factors contribute to gender disparities as well. Self-efficacy, motivation, and belief in one's ability to excel in academics can diverge by gender. Stereotype threat and imposter syndrome can hinder performance. Having role models from the same gender can greatly influence a student's academic trajectory. Martin (2021). Moreover, familial and parental influences play a pivotal role. Parental expectations, support, and encouragement, or lack thereof, can significantly impact a student's educational aspirations and performance. The concept of intersectionality underscores that gender disparities in academic achievement often intersect with other factors such as race, socioeconomic status, and cultural backgrounds, further deepening the complexity of this issue. The knowledge derived from some existing works by some of the experts in the field of education is discussed below.

(Dungdung&Bankira, 2023) focused on understanding and addressing the challenges faced by secondary school students who are struggling with mathematics anxiety. They hoped that their research could help to create a more inclusive and equitable mathematics learning environment for all students. Their study found that there is a negative correlation between mathematics anxiety and academic achievement, meaning that students with higher levels of mathematics anxiety have lower academic achievement in mathematics. The study also found that girls have higher levels of mathematics anxiety than boys and that students who attend English medium schools have higher academic achievement in mathematics than students who attend Odia medium schools.

In the research done by **(Bashir, Akram, & Bashir, 2023)**, the study found that there is a significant gender difference in attitude towards mathematics, with boys having a more positive attitude towards mathematics than girls. The study also found that there is a significant positive correlation between attitude towards mathematics and achievement in mathematics, meaning that students with a more positive attitude towards mathematics tend to achieve higher in mathematics. The author emphasizes the negative impact of mathematics anxiety on academic achievement, especially for girls and students in Odia medium schools. The author also highlights the importance of attitude towards mathematics for student achievement and the need to reduce mathematics anxiety for all students. The author argues that mathematics anxiety is a complex phenomenon with some contributing factors, including negative past experiences with mathematics, parental expectations, and peer pressure. The author calls for further research to understand the relationship between mathematics anxiety, gender, medium of school, and academic achievement, as well as the development and implementation of effective interventions to reduce mathematics anxiety and improve academic achievement for all students. The study's findings are consistent with the results of other studies that have examined the relationship between gender, attitude toward mathematics, and achievement in mathematics. For example, a study by the National Assessment of Educational Progress (NAEP) found that girls in the United States have a less positive attitude towards mathematics than boys and that this gender difference persists throughout middle and high school.

(OPOU, 2023) viewed gender disparity in students' academics as a complex issue with a variety of contributing factors, including gender stereotypes, social norms, and access to resources. They found that there is a small gender difference in academic achievement among online and distance learners, with males having a slightly higher average CGPA and shorter average completion time. However, they also found that the difference in academic achievement was relatively small and that both male and female learners performed well overall. However, the study also found that the difference in academic achievement was relatively small and that both male and female learners performed well overall. The study's findings have important implications for educators and policymakers in the Philippines. Educators should be aware of the small gender difference in academic achievement among online and distance learners, and they should take steps to ensure that all learners have equal opportunities to succeed. Policymakers should also consider investing in programs that support both male and female online and distance learners.

(Devi, 2023) found that there is a significant gender disparity in work participation in all Indian states, with the Female Labor Force Participation Rate (FLFPR) being much lower than the Male Labor Force Participation Rate (MLFPR). The study also found that there is a wide range of variation in the FLFPR across Indian states, with some states having FLFPRs as low as 15% and others having FLFPRs as high as 40%. The study identified some factors that contribute to the gender disparity in work participation in India, including a patriarchal mindset, lack of access to education and skills training, unfavorable working conditions, and lack of childcare support.

To address the gender disparity in work participation in India, the author suggests that policymakers and civil society organizations should work to promote gender equality, challenge traditional gender roles, and invest in education and skills training for women. She also recommends that the government and employers create more favorable working conditions for women, such as providing affordable childcare support, flexible work arrangements, and protection from sexual harassment and discrimination. Devi additionally suggests that the government provide financial incentives for women to participate in the workforce, such as tax breaks and subsidies for childcare costs. Finally, Devi argues that changing social norms around women's work participation is essential to achieving gender equality in the workplace

(IJERI, 2022) investigated the relationship between gender variation, internet accessibility, and students' academic performances in secondary schools in Nigeria. The study found that there is a significant positive correlation between internet accessibility and students' academic performance. This means that students who have access to the internet tend to perform better academically than students who do not have access to the internet. The study also found that there is a significant gender difference in internet accessibility, with boys being more likely to have access to the internet than girls. The study's findings have important implications for policymakers and educators in Nigeria. Policymakers should take steps to improve internet accessibility for all students, especially girls. Educators should also incorporate the use of the Internet into their teaching and learning practices to ensure that all students have the opportunity to benefit from the educational resources that are available online.

(Hegazi, Almaslukh, &Siddig, 2023) proposes a fuzzy model for reasoning and predicting students' academic performance. The model is called FPM (Fuzzy Propositional Model), and it is based on the integration of fuzzy logic and propositional logic. The FPM model can represent and reason about the uncertainty and imprecision that is inherent in student academic performance data. The model can be used to predict student academic performance based on a variety of factors, such as student attendance, test scores, and prior grades. The FPM model is effective in predicting student academic performance in a variety of studies. The FPM model has some advantages over traditional statistical models for predicting student academic performance. First, the FPM model can represent and reason about uncertainty and imprecision in the data. Second, the FPM model is more flexible than traditional statistical models, and it can be easily adapted to new data sets. Third, the FPM model is more interpretable than traditional statistical models, and it can be used to generate insights into the factors that are influencing student academic performance. Their research concludes that the FPM model is an effective tool for predicting student academic performance. The authors conclude that the FPM model has several advantages over traditional statistical models, including its ability to represent and reason about uncertainty and imprecision in the data, its flexibility, and its interpretability.

3. METHODOLOGY

The methodology adopted for the development of this study is the Structured Systems Analysis and Design Method (SSADM). Recognized as a well-established and rigorously structured systems development framework, SSADM serves as a dependable foundation for analyzing, designing, and implementing information systems in the education domain. Renowned for its clarity and precision, SSADM guides projects through distinct phases—from feasibility assessment to system maintenance—systematically unfolding the development life cycle. This strategic adoption aligns with the project's goal of harnessing SSADM's proven track record in handling complex information systems, providing a structured and disciplined approach tailored to the intricacies of education performance analysis.

By embracing SSADM, the project aims to capitalize on its methodical nature, ensuring the successful analysis, forecasting, and implementation of gender disparity in academics. SSADM's systematic approach harmonizes seamlessly with the overarching project goal: delivering a robust, reliable, and well-documented solution. In the rapidly advancing landscape of education landscape. SSADM’s structured nature not only ensures the efficiency of the development process but also contributes to creating a solution that not only meets but anticipates the evolving needs of students, and academic guidance in a technologically advanced context.

The technique adopted for this project is the STACK ENSEMBLE TECHNIQUE OF ANALYSIS. Stacking is an ensemble learning technique that combines the predictions of multiple base models to improve the overall performance of the model. It works by training a meta-model on the predictions of the base models, rather than on the original data. This allows the meta-model to learn how to best combine the predictions of the base models to make more accurate predictions.

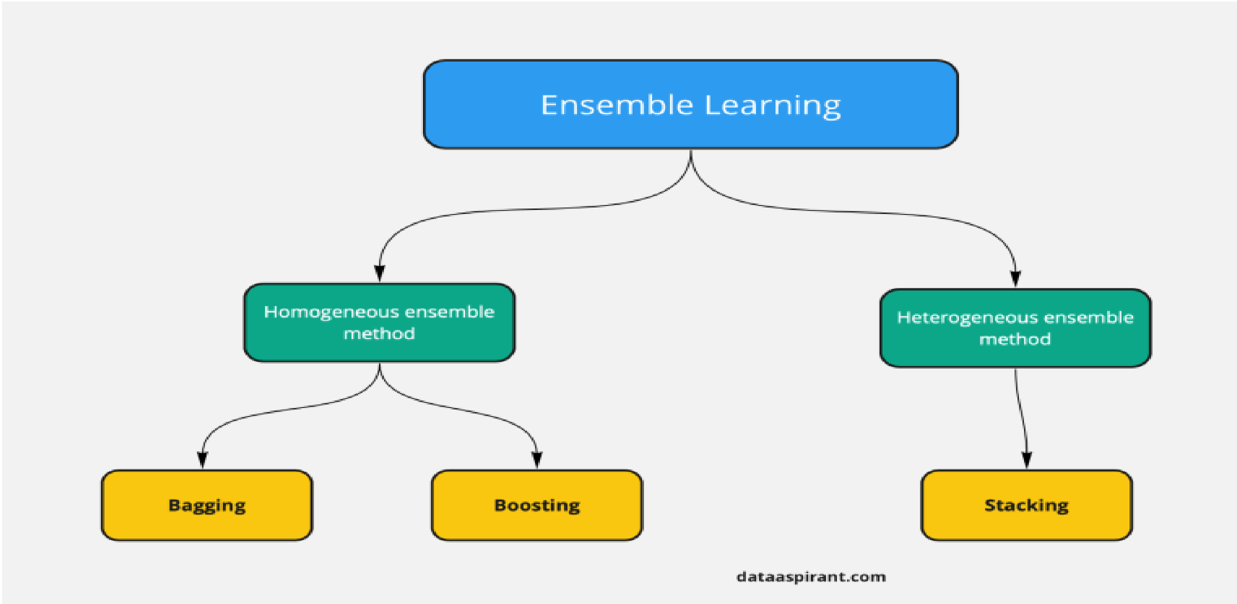

Figure 1: Ensemble Technique

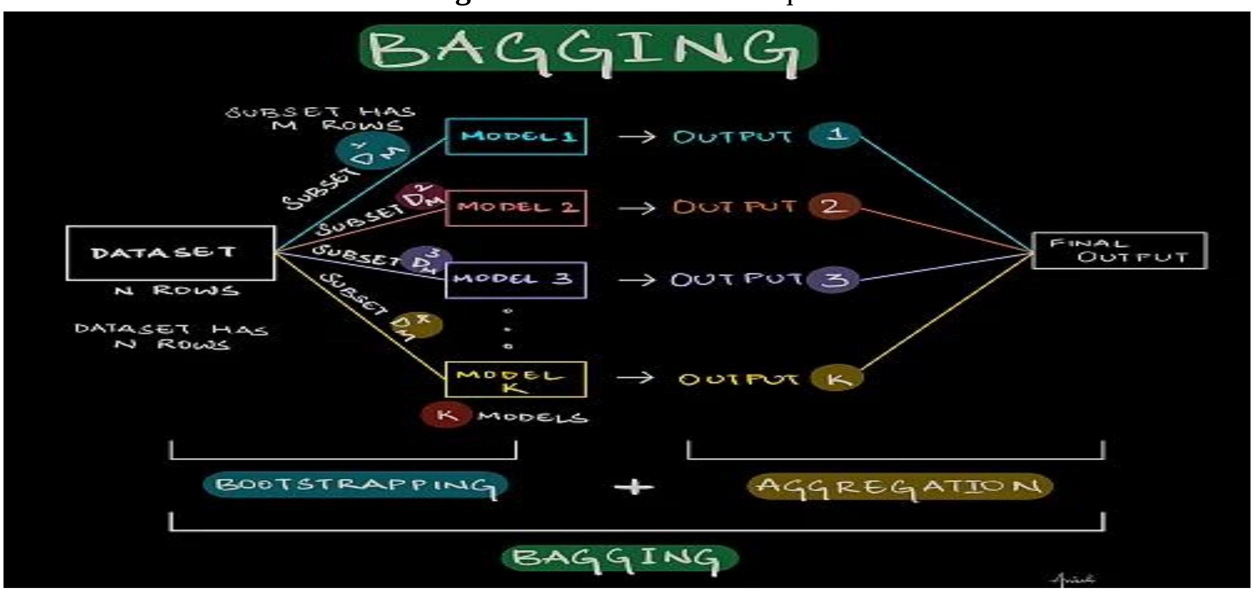

Figure 2: Bagging method in ensembling

Bagging, also known as bootstrap aggregating, is an ensemble learning technique that aims to improve the accuracy and stability of machine learning models. It works by creating multiple subsets of the training data, training a base model on each subset, and then combining the predictions of the base models to make a final prediction. The key idea behind bagging is that by training multiple models on different subsets of the data, we can reduce the variance of the overall model. This is because each base model is trained on a slightly different dataset, so it is less likely to overfit the training data.

Boosting is an ensemble learning technique that reduces the bias of a machine learning model by training multiple models sequentially. Each model is trained to correct the errors of the previous model. This means that each model is focused on learning the parts of the training data that the previous models were unable to learn correctly.

An example of the stack ensemble technique is in the field of fraud detection. Fraud detection is a challenging task because fraudsters are constantly developing new methods to defraud businesses and individuals. Stacking can be used to combine the predictions of multiple fraud detection models to improve the overall detection rate. For example, a stack ensemble for fraud detection could include the following base models: a logistic regression model trained on historical fraud data, a decision tree model trained on transaction data, a random forest model trained on customer data.The predictions of these base models would then be stacked into a meta-model, such as a logistic regression model, to make the final prediction. Stacking is a powerful ensemble learning technique that can be used to improve the accuracy and robustness of machine learning models. It is a relatively complex technique, but it is often worth the effort to implement, especially for problems where high accuracy is required.

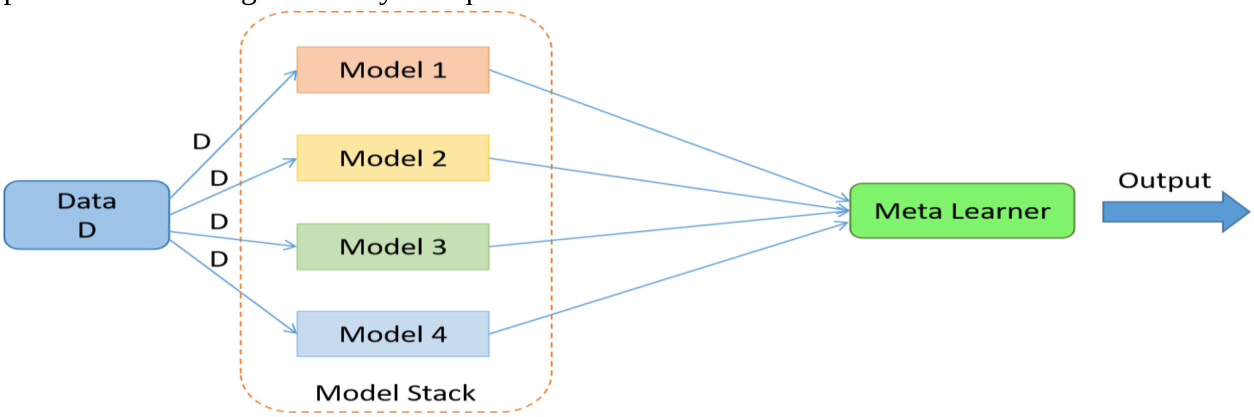

Figure 3: Stack ensemble method (Yash Khandelwal, 2021).

The steps involved in the stack ensemble technique are as follows:

i. Split the training data into folds. This is typically done using a cross-validation technique, such as k-fold cross-validation.
ii. Train a set of base models on each fold of the training data. The base models can be any type of machine learning model, but it is often recommended to use heterogeneous models, meaning that the base models are of different types.
iii. Generate predictions from the base models on the training data. This will create a new training dataset, where each data point contains the predictions of the base models as features.
iv. Train a meta-model on the new training dataset. The meta-model is responsible for combining the predictions of the base models to make the final prediction.
v. Use the meta-model to make predictions on the test data.

Stacking can combine the predictions of any type of machine learning model, including heterogeneous models (models that are of different types). This is important for mitigating and preventing gender disparity in academic achievement because there is no single model that can perfectly predict academic success. By combining the predictions of multiple models, stacking can reduce the overall error and improve the accuracy of the predictions. It can help to reduce bias in machine learning models. This is important for mitigating and preventing gender disparity in academic achievement because machine learning models can be biased against certain groups of people, such as girls and women. It can also help to reduce bias by combining the predictions of multiple models that were trained on different data sets and using different algorithms. Stacking can make machine learning models more interpretable. This is important for mitigating and preventing gender disparity in academic achievement because it allows us to understand how the model is making predictions and to identify any potential biases. Stacking models are more interpretable than boosting and bagging models because they use a meta-model to combine the predictions of the base models. The meta-model can be interpreted to understand how the different factors influence the final prediction. The above reasons are why I chose “Stacking” as the best ensemble technique for this project.

3.1 Analysis of the Proposed System

After much research, and having studied the limitations of the existing system, I have come up with the “Stack ensemble technique” as the best methodology for analyzing gender disparity in academic achievement. Stack ensemble is a powerful machine learning approach that combines the strengths of multiple base models to improve prediction accuracy and robustness. Also known as stacking, is a machine-learning technique that involves combining the predictions of multiple base models to create a meta-model that makes final predictions. The base models can be diverse, and their predictions can be weighted or averaged to improve the overall model's performance.

In this context, I will explore how stack ensemble can be applied to address gender disparity in academic achievement, focusing on its components, benefits, and considerations. The stack ensemble system comprises the following key components: base models and a meta-model. The proposed system involves several steps, including data preprocessing, feature extraction, classification, and post-processing. In data preprocessing, the raw data is cleaned to prepare it for analysis. Feature extraction involves extracting relevant features from the preprocessed data, such as n-grams, part-of-speech tags, and syntactic dependencies. These features are used to train the machine learning model.

The classification step involves training a machine learning model, such as support vector machines, decision trees, or Random Forest, to classify text data into positive, negative, or neutral categories based on the extracted features. The machine learning model is trained on a labeled dataset, which is a collection of text data with pre-assigned sentiment labels. The accuracy of the model is evaluated using metrics such as precision, recall, and F1-score.

To build a stack ensemble, a variety of base models are trained. These base models can be diverse, including regression models, decision trees, support vector machines, neural networks, or any other machine learning algorithm. Each base model is trained on relevant data features and the target variable, which, in this case, may include academic performance data, demographic data, and other relevant factors. The meta-model is responsible for aggregating the predictions of the base models. It can be a simple model, such as linear regression or decision tree, or a more complex one, like a gradient boosting machine. The meta-model takes the predictions of the base models as input and learns to provide a final prediction.

3.2 Advantages of the Proposed System

The proposed Stack ensemble approach for gender disparity analysis offers several advantages, including:

i. **Enhanced Predictive Accuracy**: The stack ensemble technique combines the predictions of multiple base models, which often results in superior predictive accuracy compared to individual models. This accuracy is crucial for identifying and addressing gender disparities in academic achievement effectively.
ii. **Robustness and Generalization**: By leveraging a diverse set of base models, the stack ensemble system is more robust and better equipped to generalize patterns across various situations and data scenarios. It can adapt to different contexts, thereby providing more reliable predictions.
iii. **Handling Heterogeneous Data Sources**: Gender disparities in academic achievement are influenced by a multitude of factors, including socio-economic status, cultural background, and individual learning styles. Stack ensemble can effectively handle heterogeneous data sources, allowing it to capture and analyze the complex interplay of these variables.
iv. **Interpretability and Insights**: The ensemble's meta-model can provide insights into the factors contributing to gender disparities. By analyzing the ensemble's predictions and understanding how different models contribute, educators and policymakers can gain valuable insights into the root causes of disparities and devise targeted interventions.
v. **Fairness and Bias Mitigation**: The stack ensemble system can be designed to incorporate fairness-aware techniques, helping to mitigate biases in predictions. Addressing fairness concerns, it contributes to a more equitable education system and avoids perpetuating existing biases.

These advantages collectively make the stack ensemble system a valuable tool for addressing gender disparities in academic achievement by improving prediction accuracy, enhancing robustness, and promoting fairness and interpretability in the decision-making process.

3.3 High-Level Model of the Proposed System

The high-level model of the proposed stack ensemble system for mitigating and predicting gender disparity in academic achievement consists of several key components and processes. This model provides an overview of how the system operates and the flow of information within it. Here is a description of the high-level model:

i. **Data Collection and Integration**: The process begins with the collection of data from various sources. These sources can include academic records, demographic data, socioeconomic information, and any other relevant factors that influence academic achievement. The collected data is preprocessed to ensure quality and consistency. This data integration step forms the foundation for the subsequent modeling process.

ii. **Base Model Training**: Next, a set of diverse base models is trained using the integrated data. Each base model focuses on capturing different aspects of the problem. For example, one base model might emphasize the impact of socioeconomic factors, while another may consider learning styles. These base models are trained to make individual predictions related to gender disparities in academic achievement.

iii. **Meta-Model Construction**: The predictions from the base models are then used as input for the meta-model. The meta-model is designed to aggregate and combine the predictions from the base models. It can be a simple linear model or a more complex machine learning model, depending on the specific requirements of the project.

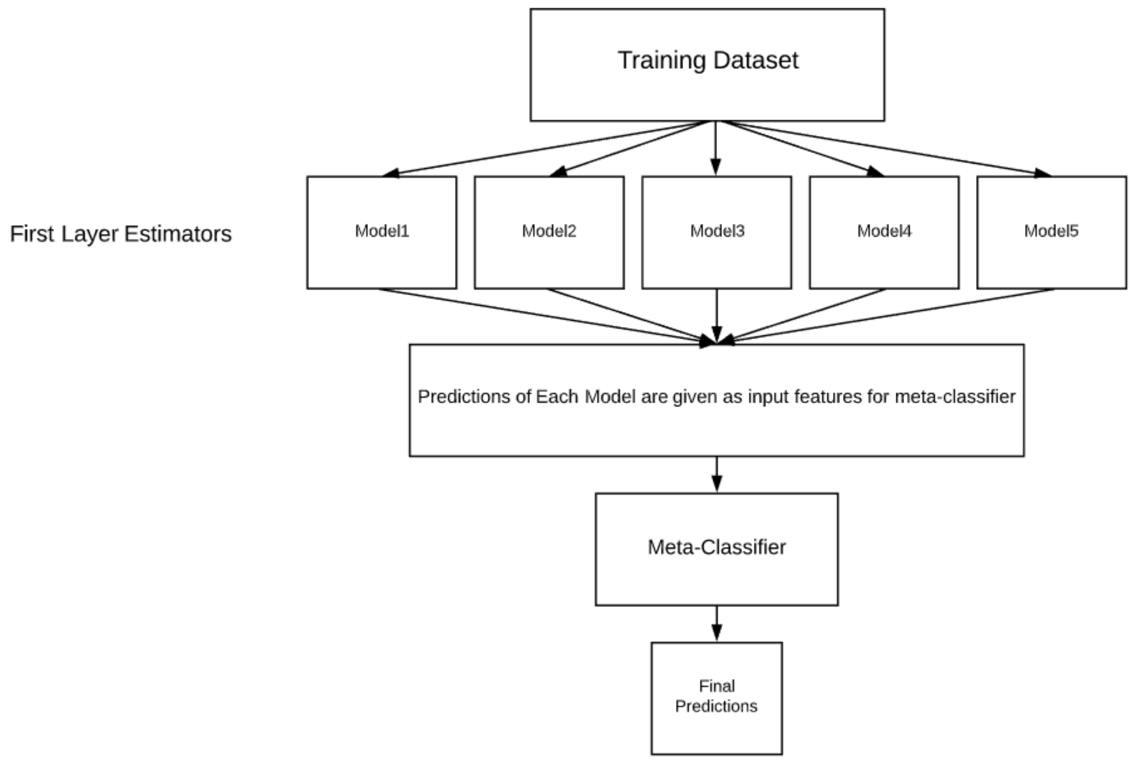

Figure 4: High level model of the proposed system

4. SYSTEM DESIGN AND IMPLEMENTATION

4.1 Objectives of the Design

The design of the project, titled "Equity Strategies in Academic Success Forecasting and Addressing Gender Disparity", encompasses a set of carefully crafted objectives that guide the methodology and implementation of the stack ensemble technique for classifying students based on various variables. These objectives are designed to address the core issues related to gender disparity in academic achievement and to leverage the predictive power of stack ensemble models for meaningful interventions. The following are the extensive objectives of the design:

i. Identifying and selecting relevant variables that significantly contribute to academic achievement, with a specific focus on factors that may contribute to gender disparity. This involves a comprehensive literature review, statistical analysis, and domain expertise to determine the critical features influencing student grades.
ii. Implementing advanced predictive modeling techniques, including the stack ensemble, to build robust classifiers for predicting student grades. This involves the construction of diverse base models, the exploration of model stacking strategies, and the optimization of hyper parameters to enhance predictive accuracy.
iii. Conduct a gender-specific analysis to understand the unique factors influencing academic achievement for male and female students. This includes identifying variables that may contribute to gender-based disparities and tailoring the modeling approach to address these specific nuances.
iv. Assess the performance of the stack ensemble models rigorously using appropriate evaluation metrics. This includes measures such as accuracy, precision, recall, and F1-score, among others. The objective is to ensure the models are not only accurate but also sensitive to gender-based disparities in academic achievement.
v. Incorporate ethical considerations in the design, implementation, and interpretation of the stack ensemble models. This involves addressing potential biases, ensuring fairness, and maintaining transparency in the model-building process to prevent unintended consequences and reinforce ethical standards.

4.2 System Implementation and Design

The dataset is formatted as a CSV file and can be downloaded from Kaggle. It includes a readme file that provides more information about the data collection process and the format of the data.

data - DataFrame

Index	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	isedhanc	edResou	ncement	Discussio	nswerinc	hoolSati	ntAbsenc	Class
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	15	16	2	20	Yes	Good	Under-7	M
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	20	20	3	25	Yes	Good	Under-7	M
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7	L
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7	L
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	40	50	12	50	No	Bad	Above-7	M
5	F	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	42	30	13	70	Yes	Bad	Above-7	M
6	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	35	12	0	17	No	Bad	Above-7	L
7	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	50	10	15	22	Yes	Good	Under-7	M
8	F	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	12	21	16	50	Yes	Good	Under-7	M
9	F	KW	KuwaIT	MiddleSchool	G-07	B	IT	F	Father	70	80	25	70	Yes	Good	Under-7	M
10	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	50	88	30	80	Yes	Good	Under-7	H
11	M	KW	KuwaIT	MiddleSchool	G-07	B	Math	F	Father	19	6	19	12	Yes	Good	Under-7	M
12	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	5	1	0	11	No	Bad	Above-7	L
13	M	lebanon	lebanon	MiddleSchool	G-08	A	Math	F	Father	20	14	12	19	No	Bad	Above-7	L
14	F	KW	KuwaIT	MiddleSchool	G-08	A	Math	F	Mum	62	70	44	60	No	Bad	Above-7	H
15	F	KW	KuwaIT	MiddleSchool	G-06	A	IT	F	Father	30	40	22	66	Yes	Good	Under-7	M
16	M	KW	KuwaIT	MiddleSchool	G-07	B	IT	F	Father	36	30	20	80	No	Bad	Above-7	M
17	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	55	13	35	90	No	Bad	Above-7	M
18	F	KW	KuwaIT	MiddleSchool	G-07	A	IT	F	Mum	69	15	36	96	Yes	Good	Under-7	M
19	M	KW	KuwaIT	MiddleSchool	G-07	B	IT	F	Mum	70	50	40	99	Yes	Good	Under-7	H
20	F	KW	KuwaIT	MiddleSchool	G-07	A	IT	F	Father	60	60	33	90	No	Bad	Above-7	M

Format Resize Background color Column min/max Save and Close Close

Figure 5: Figure showing the dataset

To analyze the data, I first utilized Pandas' read_csv() function to import the relevant information from the provided CSV file. Leveraging the power of Pandas, I was able to efficiently import the dataset stored in a CSV file using the read_csv() method.

Index	gender	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	ilTedResourc	ouncementsV	Discussion	tAnsweringS	tschoolSatisf	tAbsen	Class
gender	1	-0.0177929	0.0168688	0.0549067	0.0317693	0.0491565	-0.195142	-0.149978	-0.210932	-0.0521393	-0.124703	-0.0223586	-0.0934784	-0.20…	0.123675
StageID	-0.01779…	1	-0.961835	0.296416	-0.0474931	-0.0295124	0.0342047	-0.172751	-0.0686214	-0.163666	-0.161406	-0.114025	0.0142716	-0.11…	-0.01169
GradeID	0.0168688	-0.961835	1	-0.303949	0.061389	0.0660786	-0.0336023	0.182621	0.078262	0.183033	0.168462	0.118246	-0.0184214	0.088…	0.013482
SectionID	0.0549067	0.296416	-0.303949	1	0.267445	0.0467635	0.00578344	-0.143862	-0.0809088	-0.144955	-0.102538	-0.0184487	-0.0704048	0.037…	0.017597
Topic	0.0317693	-0.0474931	0.061389	0.267445	1	-0.0359753	-0.139487	-0.0804175	-0.118144	-0.063856	0.0540639	0.00473018	-0.0640865	-0.03…	0.10361
Semester	0.0491565	-0.0295124	0.0660786	0.0467635	-0.0359753	1	0.148705	0.178358	0.173219	0.287066	0.0190826	0.0236279	-0.0252581	0.072…	-0.04328
Relation	-0.195142	0.0342047	-0.0336023	0.00578344	-0.139487	0.148705	1	0.364237	0.36024	0.339505	0.0267196	0.163811	0.287698	0.219…	-0.27211
raisedhands	-0.149978	-0.172751	0.182621	-0.143862	-0.0804175	0.178358	0.364237	1	0.691572	0.643918	0.339386	0.31657	0.297015	0.463…	-0.23101
VisITedResources	-0.210932	-0.0686214	0.078262	-0.0809088	-0.118144	0.173219	0.36024	0.691572	1	0.5945	0.243292	0.382472	0.363835	0.499…	-0.16174
AnnouncementsView	-0.05213…	-0.163666	0.183033	-0.144955	-0.063856	0.287066	0.339505	0.643918	0.5945	1	0.41729	0.396357	0.298744	0.312…	-0.14399
Discussion	-0.124703	-0.161406	0.168462	-0.102538	0.0540639	0.0190826	0.0267196	0.339386	0.243292	0.41729	1	0.232197	0.0611041	0.218…	-0.12197
ParentAnsweringSurvey	-0.02235…	-0.114025	0.118246	-0.0184487	0.00473018	0.0236279	0.163811	0.31657	0.382472	0.396357	0.232197	1	0.539875	0.261…	-0.12325
ParentschoolSatisfaction	-0.09347…	0.0142716	-0.0184214	-0.0704048	-0.0640865	-0.0252581	0.287698	0.297015	0.363835	0.298744	0.0611041	0.539875	1	0.228…	-0.14627
StudentAbsenceDays	-0.209011	-0.112536	0.088342	0.0370624	-0.0365371	0.0724619	0.219687	0.463882	0.49903	0.312134	0.218778	0.261152	0.228385	1	-0.19908
Class	0.123675	-0.0116959	0.0134826	0.0175974	0.10361	-0.0432874	-0.272111	-0.231016	-0.161748	-0.143996	-0.121971	-0.123254	-0.146277	-0.19…	1

Figure 6: Figure showing the correlation matrix of all variables in the dataset

The corr() method in the pandas library calculates the pairwise correlation coefficient between all columns in a DataFrame. It helps identify the strength and direction of linear relationships between variables. By analyzing the correlation matrix, we can identify the most relevant variables for various purposes: Feature selection, Data exploration, and Anomaly detection. The values range of the possible values are from 0 to 10. With higher correlation coefficients approaching 10 and lower correlation coefficients approaching 0.

The following are the visualizations from the dataset which describes visually the impact of the various variables under study.

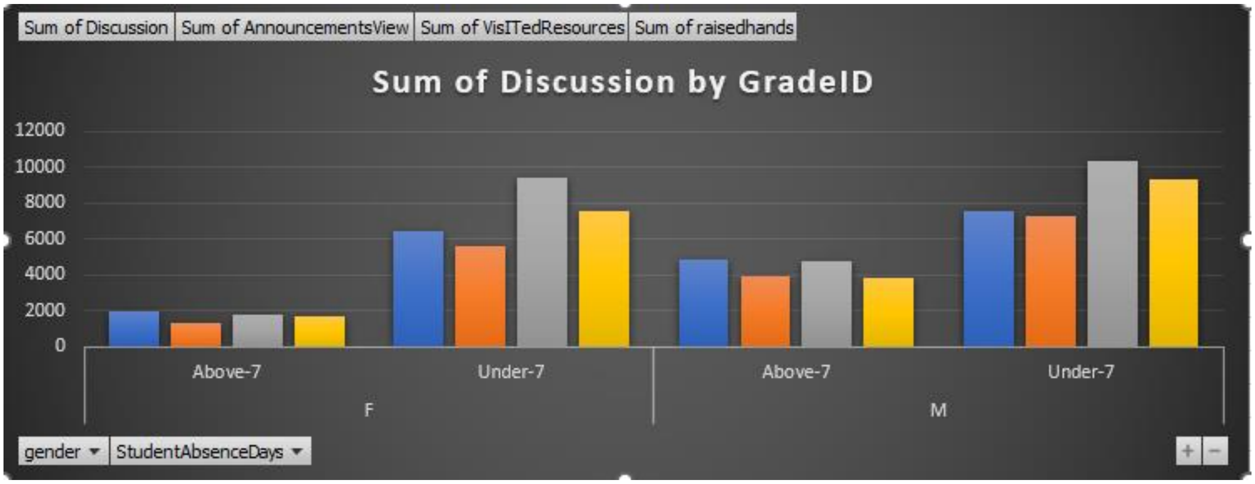

Figure 7: Bar chart showing the relationship between (sum of academic discussions had, sum of announcements read, sum of resources read, sum of times the students raised hands for questions) and (gender, rate of absence).

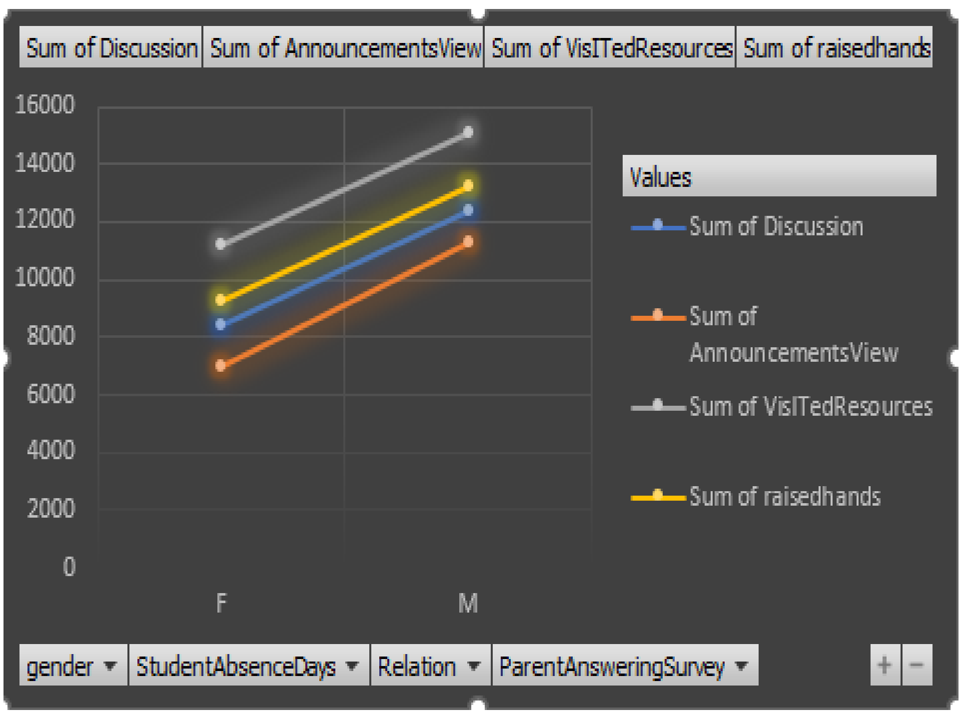

Figure 8: Line graph showing that male students are most likely to show a positive attitude towards their studies.

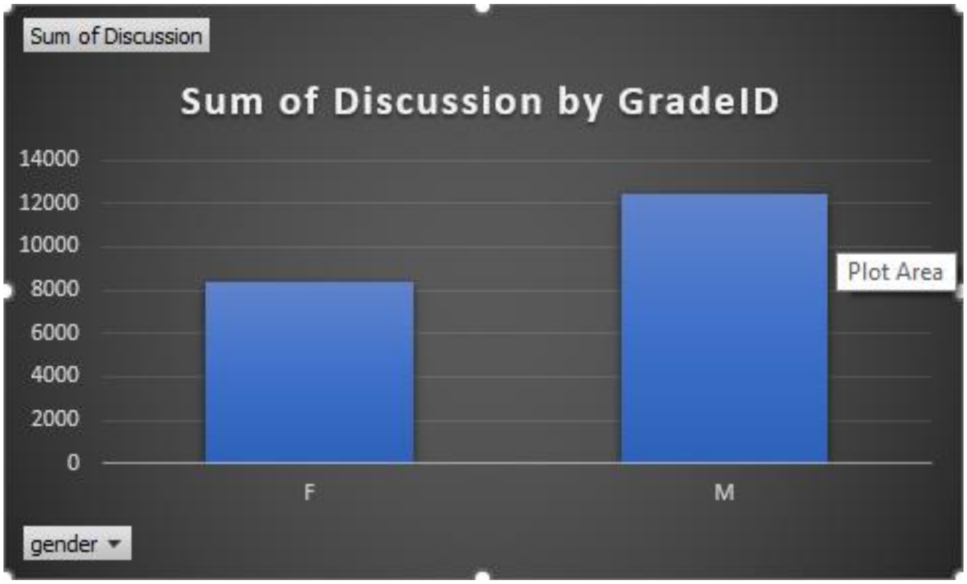

Figure 9: Bar chart showing that male students will easily engage in academic discussions more than their female counterparts.

The following table summarizes the general evaluation of the models used in the analysis.

Models	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.7152	0.72601	0.71396	0.71959
SVM	0.7569	0.75993	0.75880	0.75912
Logistic Regression	0.7707	0.77083	0.77083	0.77083
KNN	0.6944	0.68768	0.72927	0.70071
Random Forest	0.7708	0.77037	0.77966	0.77204

Figure 10: Bar chart showing comparison of the accuracies of the models.

From the above analysis, the Random Forest model is the highest in all of the metrics used for evaluation.

4.3 Control Centre/Main Menu

The Control Centre or Main Menu for the project "Equity Strategies in Academic Success Forecasting and Addressing Gender Disparity" serves as the central hub for managing, accessing, and interpreting the results derived from the stack ensemble models. This conceptual center doesn't involve a user interface but represents a cohesive structure for organizing functionalities and insights.The Control Centre organizes functionalities logically, ensuring a seamless navigation experience. It follows an intuitive flow, allowing stakeholders to systematically explore different aspects of the project without feeling overwhelmed. Within the Control Centre, key results and insights from the stack ensemble models are aggregated and presented. This provides stakeholders with a comprehensive overview of overarching patterns, trends, and gender-based analyses. The Control Centre facilitates granular access to detailed analyses, allowing stakeholders to delve into specific aspects of the project. While providing an overview, it allows for more in-depth exploration and breakdowns of data.

Contextual information and educational insights are provided alongside the presented data within the Control Centre. This enhances the interpretability of the results, ensuring stakeholders can derive meaningful implications from the findings. Mechanisms within the Control Centre proactively identify and highlight potential gender-based disparities. Alerts or indicators draw attention to specific areas or patterns that require further investigation or intervention. Functionalities and insights presented in the Control Centre align with broader educational goals. Emphasis is placed on actionable insights that contribute to the project's overarching objective of preventing and mitigating gender disparity in academic achievement. The Control Centre prioritizes secure data handling to maintain the integrity and confidentiality of sensitive information. Robust security measures are implemented to protect the data and insights presented.

4.4 The Submenus/Subsystems

These are menus or headings that provide a structured interface for navigating through key aspects of the project.

4.4.1 Gender-Based Analysis

At this phase of the project, several critical components have been accomplished, laying the foundation for the comprehensive exploration and mitigation of gender disparity in academic achievement. The achieved milestones encompass both the development of predictive models using stack ensemble techniques and the conceptualization of a user-centric Control Centre design. Extensive work has been undertaken to identify and select relevant variables that significantly contribute to academic achievement, with a particular focus on those potentially contributing to gender disparity. This involved a thorough literature review, statistical analysis, and domain expertise. Advanced predictive modeling techniques, specifically stack ensemble models, have been implemented. This process involved the creation of a diverse set of base models, experimentation with stacking strategies, and the optimization of hyperparameters to ensure robust and accurate predictions. A gender-specific analysis has been conducted to understand the unique factors influencing academic achievement for male and female students. This analysis identified variables that may contribute to gender-based disparities, laying the groundwork for tailored modeling approaches. Rigorous performance evaluation metrics have been applied to assess the effectiveness of the stack ensemble models. Measures such as accuracy, precision, recall, and F1-score have been employed to ensure not only accuracy but also sensitivity to gender-based disparities. The models have been examined for their explanatory power, providing insights into the underlying factors influencing the classification of students into different grades. Feature importance, model interpretability, and actionable insights have been derived.

4.4.2 Performance and Insights

In the "Performance and Insights" section, the primary objective is to create a focused and informative submenu within the Control Centre, designed to offer stakeholders a comprehensive understanding of the performance of the stack ensemble models and derive actionable insights. This section serves as a crucial hub for stakeholders to gauge the effectiveness of the models and uncover key trends related to academic achievement and gender disparities. One of the major components of the performance and insights is providing stakeholders with a concise yet detailed overview of the performance of the stack ensemble models. This includes accuracy metrics, precision, recall, F1-score, and other relevant performance indicators. Visualizations, such as confusion matrices, may be employed to offer a clear snapshot of the model's predictive abilities. Also, to incorporate visualizations and summary statistics to present data-driven insights in an accessible manner. Visual representations, such as charts, graphs, and heatmaps, offer stakeholders an intuitive understanding of academic achievement patterns and gender-based disparities. Summary statistics complement these visuals, providing a numerical context to the presented insights.

5. CONCLUSION AND FUTURE DIRECTION

In the intricate landscape of educational data science, this research project stands as a guiding light, offering profound insights and transformative strategies to untangle the complexities surrounding gender disparity in academic achievement. By harnessing the power of cutting-edge data science techniques, the project not only seeks to uncover the root causes of this persistent gap but also endeavors to pave the way for meaningful interventions that resonate with real-world impact. The journey of understanding and addressing gender disparities in academic achievement begins with a meticulous exploration of multifaceted datasets, intricately weaving together student demographics, performance metrics, and diverse influencing factors. This wealth of information undergoes a rigorous analytical process, employing sophisticated predictive models and advanced statistical methods. The aim is not merely to predict outcomes but to discern the intricate interplay of variables, dissecting the unique challenges faced by different genders in the educational landscape. As the project advances, the strategic deployment of targeted interventions emerges as a cornerstone. Informed by the nuanced findings of the predictive models, these interventions are tailored to address specific factors contributing to gender disparity. Personalized learning programs, finely tuned to individual needs and learning styles, take shape, offering a dynamic approach that accommodates the diverse ways in which students absorb knowledge.

5.1 Future Direction

The research presented in this paper opens avenues for further exploration in the realm of gender disparity in academic achievement. Several promising directions for future research include:

i. *Intersectionality Analysis*: Investigate the intersectionality of gender disparities with other factors such as socioeconomic status, ethnicity, or geographical location to uncover more nuanced insights.
ii. *Longitudinal Studies*: Conduct longitudinal studies to track the long-term impact of implemented interventions on academic performance and career outcomes, providing a more comprehensive understanding of the efficacy of strategies over time.
iii. *Comparative Studies*: Extend research to compare gender disparities across different educational systems, regions, or cultural contexts, shedding light on the influence of diverse socio-cultural factors.
iv. *Technology Integration*: Explore the integration of emerging technologies, such as artificial intelligence and virtual reality, to enhance the effectiveness of personalized learning programs and interventions.
v. *Institutional Adoption*: Investigate the challenges and facilitators of institutional adoption of data-driven interventions, examining factors influencing the successful implementation of strategies in diverse educational settings.

These suggested avenues for further research aim to deepen our understanding of gender disparities and contribute to the ongoing development of effective interventions in the educational landscape.

6. REFERENCES

Adegoke, P. S., & Oladele, B. K. (2022). Gender variation, internet accessibility, and students' academic performance. *International Journal of Academic Multidisciplinary Research, 6*(7), 6–10.

https://www.researchgate.net/publication/362404952_Gender_Variation_Internet_Accessibility_and_Students_Academic_Performance

Khandelwal, Y. (2021). Ensemble stacking for machine learning and deep learning. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/

Kumara, A., Patel, S., & Smith, T. (2020). Gender disparity in education: A persistent issue. *Journal of Education Studies*, *15*(2), 25–35. https://digitalcommons.unl.edu/libphilprac/4521/

Lee, J. H., & Chang, T. S. (2023). Gender disparities among vitreoretinal specialists in the United States about scholarly impact and academic rank. *Ophthalmology*, *130*(10), 1386–1392. https://pubmed.ncbi.nlm.nih.gov/37073287/

Miao, S. (2023). *Gender disparity in STEM: Current trends and future directions* (1st ed.). Academic Press. https://www.elsevier.com/books/gender-disparity-in-stem/miao/9780323859782

Nirmal, S. A., Anand, S., Kumawat, M., Sharma, R. S., & Wadhwa, P. (2023). Female labor utilization in India. *Directorate General of Employment*. https://dge.gov.in/dge/sites/default/files/2023-05/Female_Labour_Utilization_in_India_April_2023_final__1_-pages-1-2-merged__1_.pdf

Patel, A., Kumar, S., & Rajan, N. (2022). Machine learning in predicting educational outcomes and addressing gender disparities. *Machine Learning Applications in Education*, *9*(4), 120–134. https://link.springer.com/journal/42140

Ritz, R., James, M., & O'Brien, P. (2023). Economic disparities and education: A global perspective. *World Development Journal*, *35*(6), 153–168. https://www.sciencedirect.com/journal/world-development

Smith, T., Johnson, D., & Anderson, E. (2022). Gender-based violence and its impact on education. *International Review of Educational Development*, *46*, 89–103. https://www.sciencedirect.com/journal/international-journal-of-educational-development

Stiso, J., Oudyk, K., Bertolero, M. A., Zhou, D., Teich, E. G., Lydon-Staley, D. M., Zurn, P., & Bassett, D. S. (2023). Modeling observed gender imbalances in academic citation practices. *arXiv Preprint*. https://www.researchgate.net/publication/360233083_Modeling_observed_gender_imbalances_in_academic_citation_practices

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2021). *Global education monitoring report 2021/2: Gender report: Building back more equal*. https://en.unesco.org/gem-report/report/2021/20212-gender-report-building-back-more-equal

World Bank. (2022). *The World Bank and girls' education*. https://www.worldbank.org/en/topic/girlseducation