IŞIK UNIVERSITY
Faculty of Engineering
Department of Industrial Engineering

Final Project

INDE4185

Introduction to Data Mining

"The Impact of Social Media on College Students"

By

Feras Mohammad

19INDE1086

Mohammad Alkorom

20INDE1053

Supervised by: Dr. ISMAIL KAYAHAN

30 MAY

Contants

1) Introduction

Social media is a big part of our life. It's a place to share experience and conversation, but if it's overused, maybe impacting The focus and Grades performance, We often use platforms like Instagram, TikTok, Snapchat, WhatsApp, Facebook, LinkedIn, YouTube, and Twitter(X) to keep in touch with friends and family, share moments, and participate in different online communities. These platforms will help us access news and education materials and various forms of entertainment, making social media an essential part of our daily routines. We have noticed that using social media too much can sometimes create problems, like lower academic performance, increased distractions, stress, and a decrease in productivity.

It's easy to lose a track on time on using social media, I noticed that I almost putting things off when I used a bit of time on social media like assignment sound projects and studying . We also noticed heavy social media usage. Make sleeping harder and More difficult, especially when We stay up late chatting with someone or watching videos. This results in tiredness and difficulty concentrating during classes or exams. Also, constantly seeing carefully curated content from peers can make us feel anxious or insecure, affecting our mental wellbeing.

It's not all bad though. Social media can actually bring students together through online groups and forums and those direct messages can be a real source of support. We've also seen the positive impact of educational resources on social media like study groups and live chats which help students connect and learn from each other's experiences.

In this study, how social media impacts students grades mental well-being and overall social lives will be using surveys to gather data from students about their experiences. We analyze this information; we hope to find clear patterns that can offer practical insights. These findings are useful for students like us and teachers as well helping us develop healthier social media habits. And the objective is to discover how we can use social media effectively while minimizing it negative impacts.

2) Data Collection

To have better understanding for the impact of social media on the students academic performance and mental health and social life. We conducted a structured survey targeting university students across various academic years, The aim of this survey was to gather both demographic data and subjective insights into students' social media usage habits, perceived effects, and academic outcomes.

2.1) Survey Design

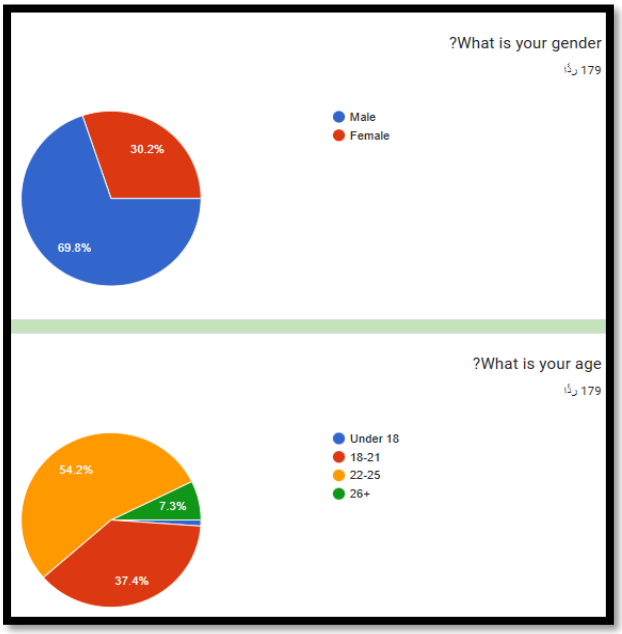
The survey were designed using a mix of multiple choice and Liker scale questions for ensure both quantitative and qualitative insights the questions was grouped in to four main sections

Demographic Information Gender Age group Year for Study Primary Field of Study you will see on the questions down

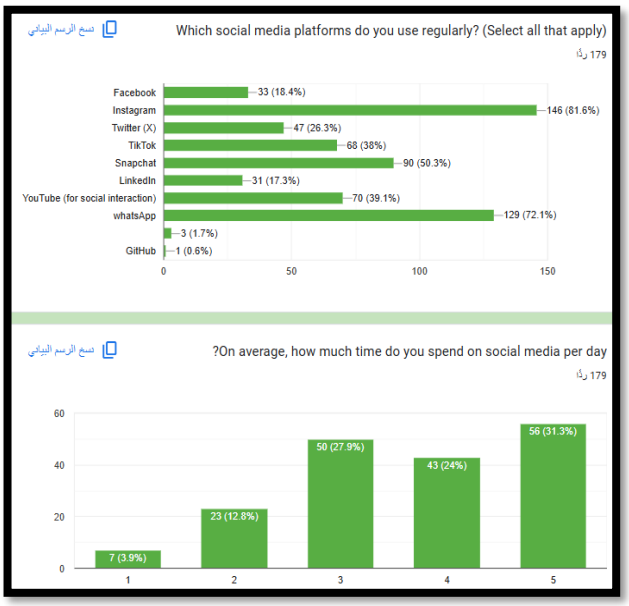
Social Media Usage Habits platform using regularly like Instagram and TikTok and WhatsApp Average daily time spent in social media frequency for checking notifications Using before bedtime As you will see on the questions down

Academic Impact Self Reported distraction levels due for social media Tendency for procrastinate in academic tasks Current GPA frequency for social media use interrupting study class tim as you will see on the questions down.

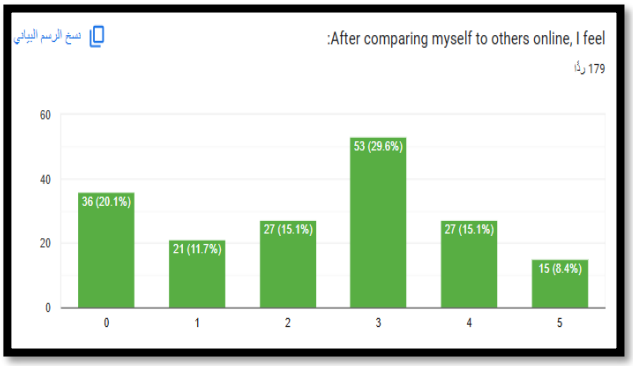
Mental Health and Social Effects Impact on sleep quality Feelings of stress or anxiety FOmO (Fear of Missing Out) levels Emotional response for peer comparisons Effects in motivation or self worth role for social media on build relationships or networks As you will see on the question down

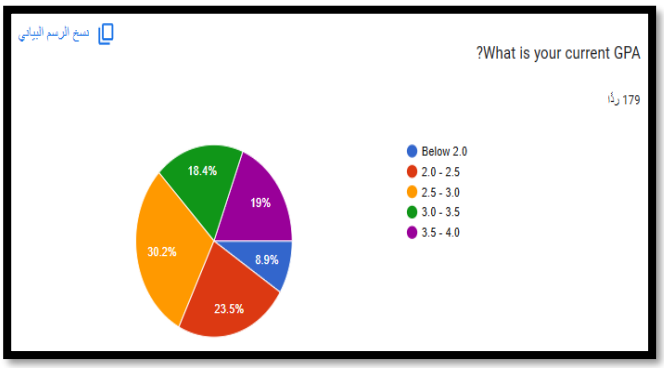

?What is your gender
179 ردًا
Male
Female
30.2%
69.8%
?What is your age
179 ردًا
Under 18
18-21
22-25
26+
54.2%
7.3%
37.4%

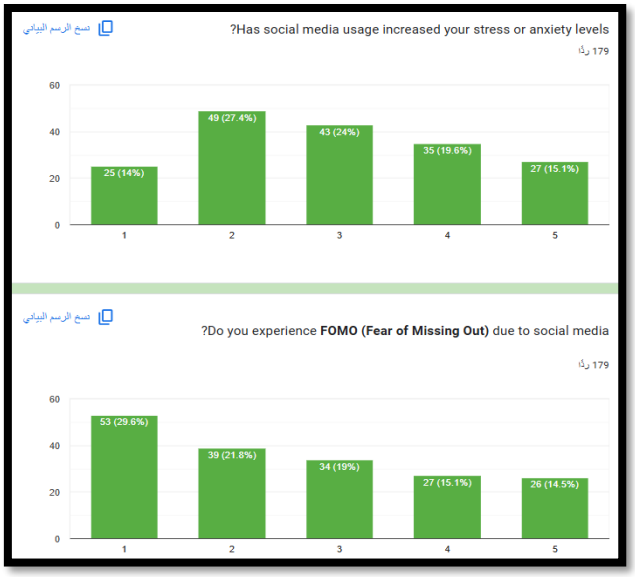

نسخ الرسم البياني
?What is your current year of study
179 ردًا
60
40
20
0
34 (19%)
21 (11.7%)
46 (25.7%)
44 (24.6%)
34 (19%)
1
2
3
4
5
نسخ الرسم البياني
What is your primary field of study/major? (Select closest match)
179 ردًا
8.9%
62.6%
STEM (Science, Technology, Engineering, Math)
Social Sciences (Psychology, Sociology, Economics, etc.)
Humanities (Literature, History, Philosophy, etc.)
Business/Finance
Arts (Visual Arts, Music, Design, etc.)
Health Sciences (Medicine, Nursing, Public Health, etc.)

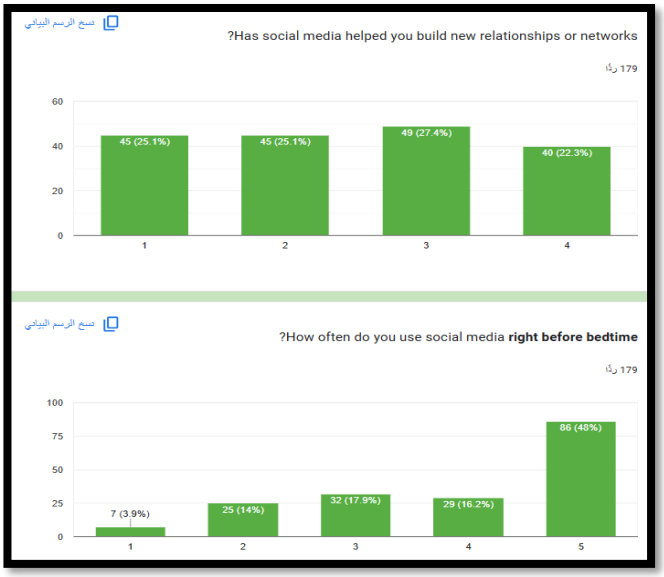

نسخ الرسم البياني
Which social media platforms do you use regularly? (Select all that apply)
179 ردًا
Facebook
Instagram
Twitter (X)
TikTok
Snapchat
LinkedIn
YouTube (for social interaction)
whatsApp
GitHub
33 (18.4%)
146 (81.6%)
47 (26.3%)
68 (38%)
90 (50.3%)
31 (17.3%)
70 (39.1%)
129 (72.1%)
3 (1.7%)
1 (0.6%)
0
50
100
150
نسخ الرسم البياني
?On average, how much time do you spend on social media per day
179 ردًا
60
40
20
0
7 (3.9%)
23 (12.8%)
50 (27.9%)
43 (24%)
56 (31.3%)
1
2
3
4
5

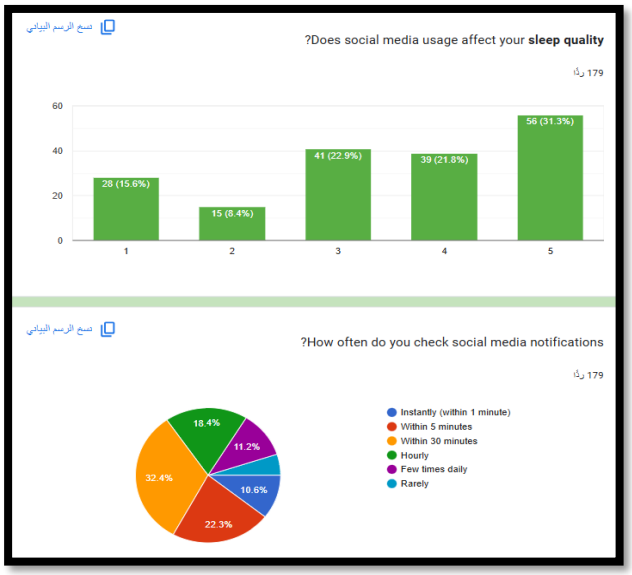

نسخ الرسم البياني
?Do you feel social media distracts you from studying
179 ردًا
80
60
40
20
0
15 (8.4%)
30 (16.8%)
37 (20.7%)
31 (17.3%)
66 (36.9%)
1
2
3
4
5
نسخ الرسم البياني
?How often do you procrastinate on academic tasks because of social media
179 ردًا
60
40
20
0
20 (11.2%)
36 (20.1%)
58 (32.4%)
35 (19.6%)
30 (16.8%)
1
2
3
4
5

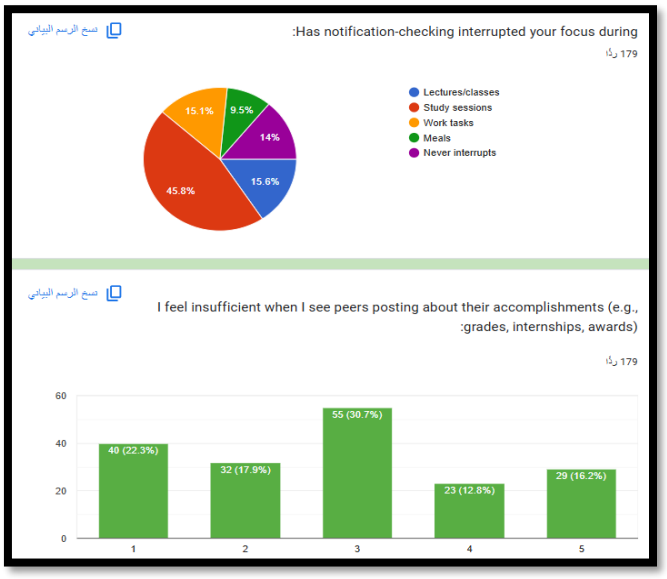

نسخ الرسم البياني
:After comparing myself to others online, I feel
179 ردًا
60
40
20
0
36 (20.1%)
21 (11.7%)
27 (15.1%)
53 (29.6%)
27 (15.1%)
15 (8.4%)
0
1
2
3
4
5


نسخ الرسم البياني
?What is your current GPA
179 ردًا
18.4%
19%
30.2%
8.9%
23.5%
Below 2.0
2.0 - 2.5
2.5 - 3.0
3.0 - 3.5
3.5 - 4.0









2.2) Converted Data

the texts based survey responses was converted in to numerical values for enable statistical analyses Liker scale answers were coded from 0 to n based on intensity or agreement level this standardize allow to easier data processing comparison also visualization ensuring consistency across the variables on dataset.

Social Media Platform Using (Multi Select Question) Facebook (1 if select 0 ow) Instagram (1 if select 0 ow) Twitter (X) (1 if select 0 ow) TikTok (1 if select 0 ow) Snapchat (1 if select 0 ow) LinkedIn (1 if select 0 ow) YouTube (1 if select 0 ow) WhatsApp (1 if select 0 ow).

Example How the Converted already Done

Which social media platforms do you use regularly? (Select all that apply)
Instagram, TikTok, Snapchat, whatsApp
Instagram, TikTok, Snapchat, YouTube (for social interaction), whatsApp
Instagram, TikTok, Snapchat, whatsApp
Instagram, TikTok, Snapchat, YouTube (for social interaction)
Instagram, whatsApp
Instagram, Snapchat, whatsApp
Instagram, Twitter (X)
Instagram, TikTok, Snapchat, whatsApp
Instagram, TikTok
Instagram, TikTok, Snapchat
Instagram, TikTok, Snapchat, YouTube (for social interaction), whatsApp
TikTok, Snapchat, whatsApp
Instagram, TikTok
Facebook, Instagram, Twitter (X), TikTok, Snapchat, YouTube (for social inter
Instagram, Snapchat, YouTube (for social interaction), whatsApp
Instagram, Twitter (X), TikTok, Snapchat, whatsApp
Instagram, Twitter (X), YouTube (for social interaction), whatsApp
Instagram, Snapchat, whatsApp
Facebook, Instagram, TikTok, whatsApp
Instagram, whatsApp
Facebook, Instagram, Twitter (X), TikTok, Snapchat, LinkedIn, YouTube (for s
Instagram, whatsApp
Facebook, Instagram, Twitter (X), whatsApp
TikTok, Snapchat, whatsApp
Instagram, TikTok, Snapchat, whatsApp
Instagram, TikTok, Snapchat, whatsApp
Instagram, TikTok, Snapchat, YouTube (for social interaction), whatsApp
Instagram, Twitter (X), Snapchat, YouTube (for social interaction), whatsApp
TikTok, LinkedIn, whatsApp
Instagram, Twitter (X), Snapchat, whatsApp

Facebook	Instagram	LinkedIn	Snapchat	TikTok	Twitter (X)	YouTube	whatsApp
0	1	0	1	1	0	0	1
0	1	0	1	1	0	1	1
0	1	0	1	1	0	0	1
0	1	0	1	1	0	1	0
0	1	0	0	0	0	0	1
0	1	0	1	0	0	0	1
0	1	0	0	0	1	0	0
0	1	0	1	1	0	0	1
0	1	0	0	1	0	0	0
0	1	0	1	1	0	0	0
0	1	0	1	1	0	1	1
0	0	0	1	1	0	0	1
0	1	0	0	1	0	0	0
1	1	0	1	1	1	1	1
0	1	0	1	0	0	1	1
0	1	0	1	1	1	0	1
0	1	0	0	0	1	1	1
0	1	0	1	0	0	0	1
1	1	0	0	1	0	0	1
0	1	0	0	0	0	0	1
1	1	1	1	1	1	1	1
0	1	0	0	0	0	0	1
1	1	0	0	0	1	0	1
0	0	0	1	1	0	0	1
0	1	0	1	1	0	0	1
0	1	0	1	1	0	0	1
0	1	0	1	1	0	1	1
0	1	0	1	0	1	1	1
0	0	1	0	1	0	0	1

We remove platform columns for example LinkedIn if they was use with less than 15% for participants We also cleaned the dataset with removing responses by empty or unanswere questions for ensure data quality**.**

3) Methodology

To understand how social media affects college students academic performance mental health also social life we decided to use a survey we created an online questionnaire that asked students about their habits, feelings, and experiences related to social media. The survey was shared with university students from different years and majors to get a variety of opinions. The survey had different parts. First we ask basic question for exmples age gender also what they study Then we focused in how much time they spend in social media which platforms they use and how often they check notifications ather questions ask if social media makes it harder for study caus stress or effects sleep Some questions and ask how they feel when comparing themselfe in others online we use multiple choice and scale based question so we could later turn all the answers into number this help us organize the data better like if someone chose TikTok as a platform they use we wrote it as 1 which is like yes and if they didnt it was 0 which is like For the scale questions we gave numbers from 0 to n depending in how strong the answer was this method helped us clearly see patterns in how students use social media and how it connects to things like their grade stress levels or sleep.

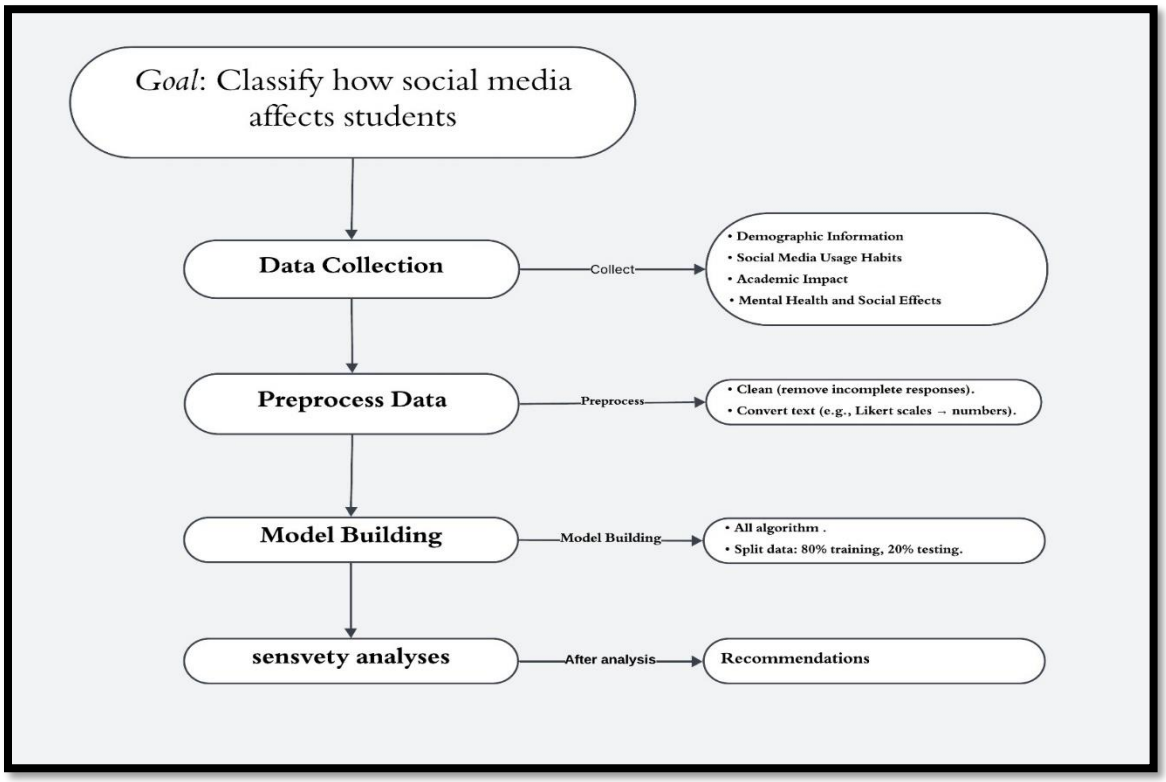


Figure-1

we used classification techniques in MATLAB to analyze the survey data and predict key outcomes like academic performance and mental health effects.

- First, we converted all answers into numerical values so that the data could be processed properly.
- second, we randomly split the dataset into 80% training data and 20% test data using MATLAB code (shown in the figures below). This help ensure that the model would train well also test fairly in unseen data.

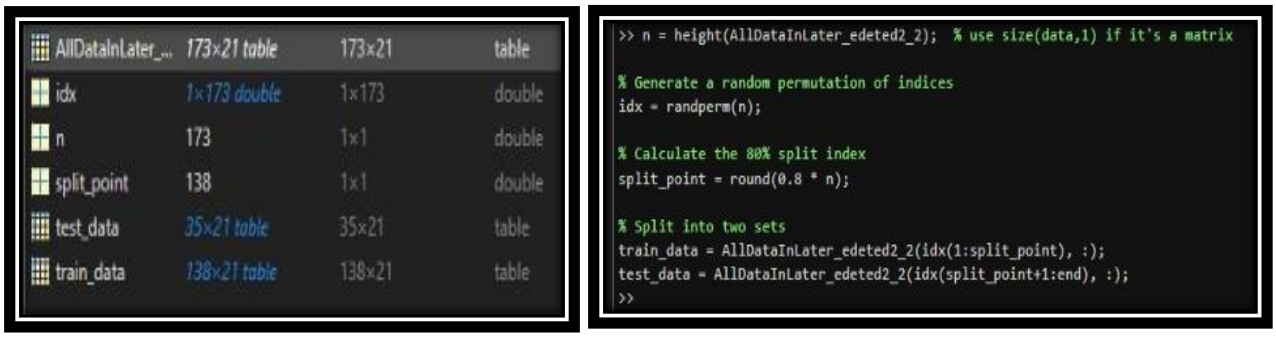


Figure-2

We appled all main classification method available on MATLAB including Decision Trees K Nearest Neighbors KNN Support Vector Machines SVM Naive Bayes and all other classification methods.

each method We will train and test the models We also used cross validation for measure the validation accuracy for each method.

After running all models, we compared both validation accuracy and test accuracy and selected the best three classification methods that performed the highest. These top models were compared in more detail to understand their strengths and how well they can predict student outcomes based on social media habits.

4) Literature Review

1. Junco R. 2012. So much face and not enough books, The relationship between multiple indices of Facebook use and academic performance.

Influence for Facebook activity in GPA and student engagement were evaluated using an variety for indices including number of friends and time spent and frequency for visits they found an statistical significant decrese link between Facebook using and academic performance using regression analyses in survey data from 1,800 students how excessive social media use might divert students from their academic obligations and lower academic success these data bolsters the projects concept.

2. Al Menayes J J 2015. Social media using engagement , addiction is predictors for academic performance.

Using (SEM) and a structured survey discovered that excessive use of social media and behavioural addiction both has an detrimental impact in students academic performance our categorization models psychological and behavioural characteristics like procrastination bedtime using closely match the emotional and compulsive aspects found in this study like anxiety and also as a compulsive checking.

3. Kirschner P A and Karpinski A C 2010. Facebook and an academic performance.

For their findings Facebook user report spending low time studying each week in addition to having noticeably worse GPA in univertstiy this study validates the inclusion of GPA as a dependent variable driven by certain social media use patterns which makes it directly relevant to thier model.

4. Lau W W F 2017. Effects of social media usage and social media multitasking on the academic performance of university students.

using path analyses on conjunction by the time use diary method to asses the frequency of students use social media concurrently by academic assignments The findings indicated thet multi tasking have an significant detrimental effect in academic focus and performance This

provides empirical backing of us survey designs inclusion for features like usage before bed time and the frequency for checking messages.

5. Kuss, D J and Griffiths M. D. 2015. Social networking sites and addiction: Ten lessons learned.

10 major behavre and psychologi effects for excessive using was discovered with the scientists includ decreased motivation disturbed sleep and elevated stress this are on line with our models focus in mental health markers as predictors of academic performance including anxiety emotional comparison with peers also poor sleep quality.

6. Azizi S M et al 2019. The relationship in internet addiction and academic performance among medical students.

In this researchers performed an correlational analyses usage the (IAT) and GPA data and thay discovered an strong negative link Academic achievement was consistently poorer among those with greater internet addiction ratings This study supports our data mining classification processs inclusion of stress screen time also digital habit like characteristics.

7. Zhao N and Zhou G 2021. Social media using also academic performance among college students A meta analyses.

findings supported a persistently negative correlation acros many technikues and demographics Ther results statistically validate the noteon that unmoderated social media use can negatively effect student success measures like focus and GPA providing strong theoretical support for our approach.

5) Result

After preparing our data, we applied all available classification methods in MATLAB using the Classification Learner app. Each model was trained and tested using an 80% training and 20% testing split. We then compared their performance based on Validation Accuracy and Test Accuracy to identify the most reliable models.

we found that the top three classification methods were:

- Fine Tree (Model 1) Figure-3 Validation Accuracy: 54.35%, Test Accuracy: 84.06% This model performed very well on the test set despite lower validation accuracy.

- Medium Gaussian SVM (Model 2.15) Figure-4 Validation Accuracy: 63.04%, Test Accuracy: 83.33% This model showed a strong balance between validation and test accuracy.

- SVM Kernel (Model 2.33) Figure-5 Validation Accuracy: 64.49%, Test Accuracy: 87.68% This model gave the highest combined performance across both validation and test sets.

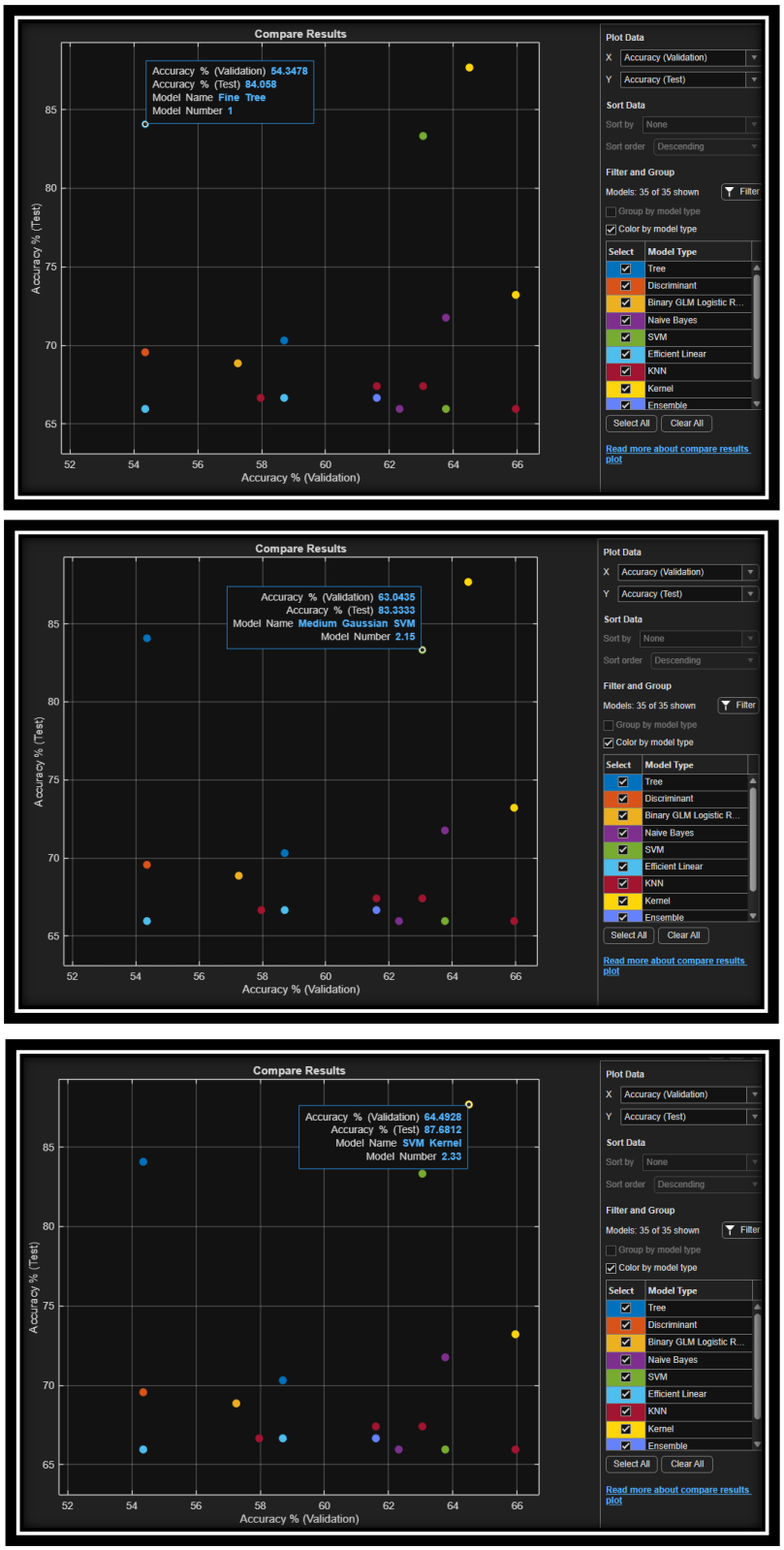


Figure-3-4-5

Confusion Matrix Comparing (Validation Results)

The Fine Tree, While the Fine Tree model had high validation accuracy 84.06% It misclassified a large number of both TP and TN samples which suggests it may be overfitting the training data.

The Medium Gaussian SVM, the model is biased towards predicting which it predicts very accurately so it fails to recognize correctly in most cases indicating class imbalance sensitivity.

The SVM Kernel, the model performs well in TPbut has difficulty predictingTN Still gives the highest and strong validation accuracy 87.68% making it the most balance and reliable of the three on overall performance.

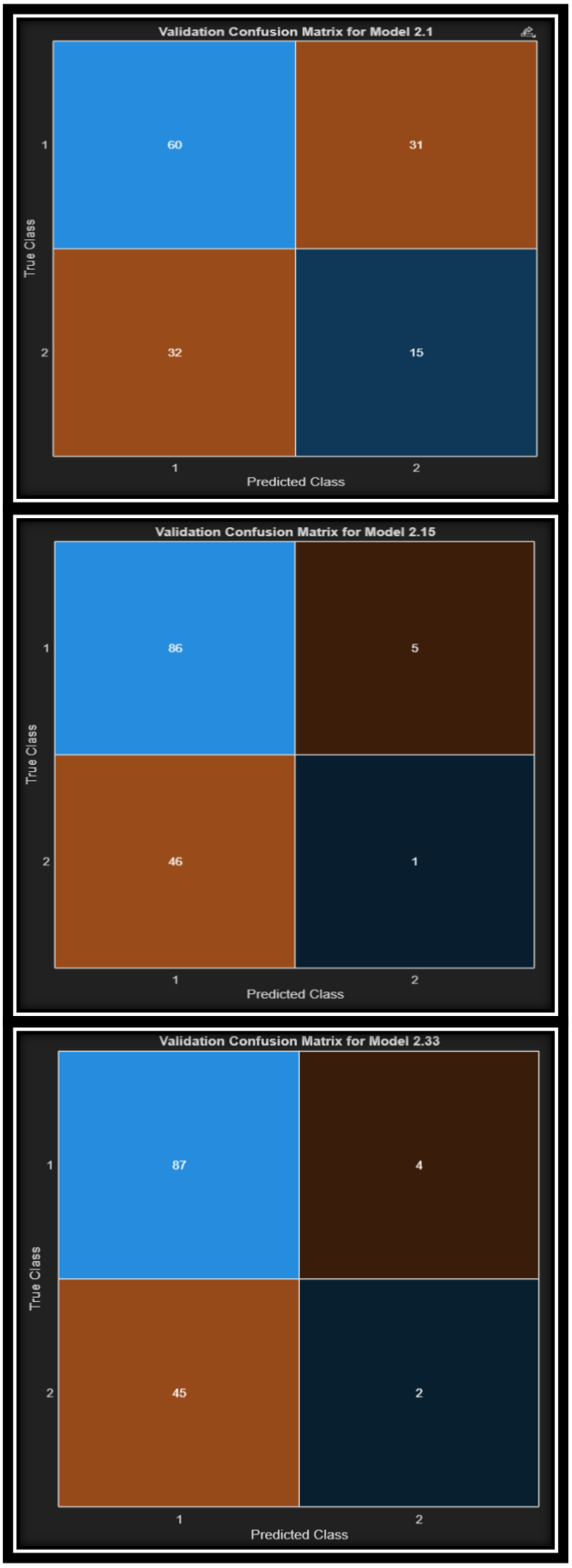


Figure-6-7-8

For better understand how each model performs, We compared the validation confusion matrices of the top three classifiers

1) Fine Tree
2) Medium Gaussian SVM
3) SVM Kernel.

These matrices show how well each model predicted the two classes based on the validation data.

Model Name	TP	TN	FP	FN	Notes
Fine Tree	60	15	31	32	Struggles in validation, risk of overfitting
Medium Gaussian SVM	86	1	5	46	Based toward TP, poor at detecting TN
SVM Kernel	87	2	4	45	Best test accuracy, more balanced, still weak in TN

The three models show good potential during validation the SVM Kernel model stands out as the most consistent in terms of both accuracy prediction balance all models show a clear weakness in correctly predictingTN as seen in the low number for true positives for that class this suggests that there may be a class imbalance in the dataset or that some features may not provide enough discriminatory power. Addressing these issues either by resampling the data or refining feature selection, could improve model performance on future work.

Confusion Matrix Comparing (Test Results)

The Fine Tree, the model showed weak performance in the new test set It misclassified a large portion for class 1 samples and struggled with class 2 resulting in high false positives also false negatives Its predictions were unbalanced and less reliable compared for other models.

The Medium Gaussian SVM, this model performed well in predicting class 1 but severely underperformed on class 2 correctly identifying only 1 true negative. The high number of false negatives 11 highlights its an strong bias toward the majority class also poor generalization to minority class examples.

The SVM Kernel, the model continues to show the most balanced performance It correctly predicted most class 1 instances and achieved the highest number of true negatives by an moderate false negative rate it remains the strongest model in terms for overall accuracy also generalization.

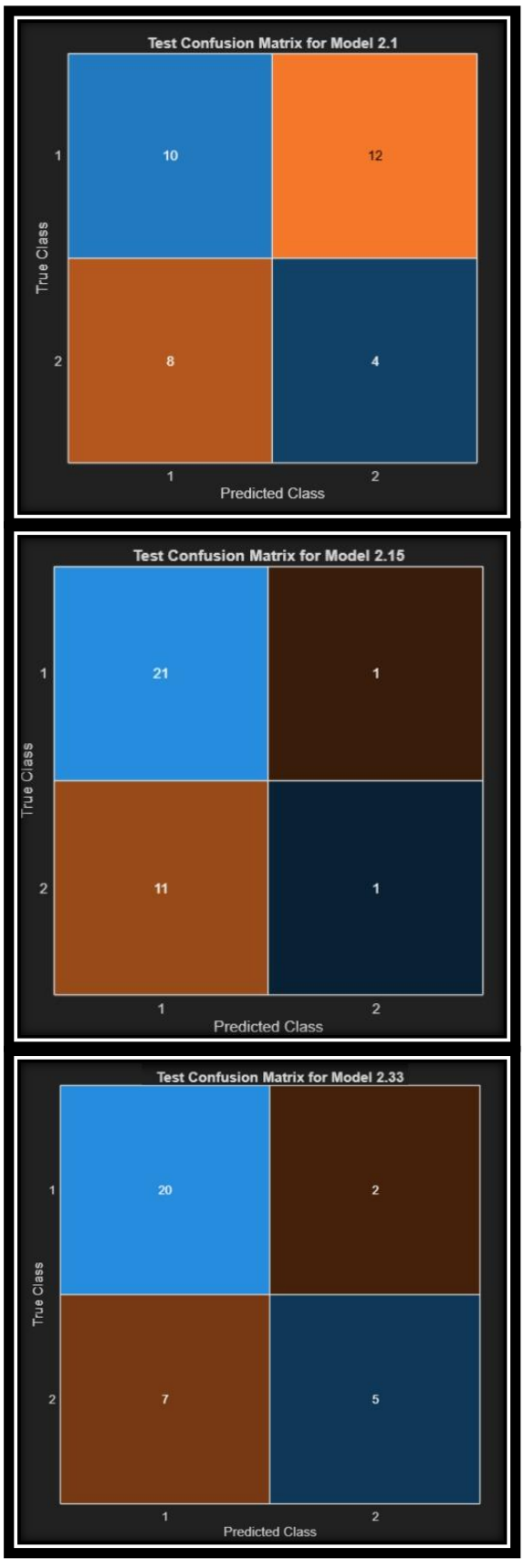


Figure-9-10-11

For evaluate the real world performance for our model in new unseen data we compare the test confusion matriceses of the same three classifiers

1. Fine Tree,
2. Medium Gaussian SVM,
3. SVM Kernel,

This matrices show how accurately each model predict the 2 response classes during this new round for testing.

Model Name	TP	TN	FP	FN	Notes
Fine Tree	10	4	12	8	Weak in overall also high miss classification in both class
Medium Gaussian SVM	21	1	1	11	Very strong in class 1 but still fails in class 2
SVM Kernel	20	5	2	7	Most balance also accurate acros both class

In the new test data the SVM Kernel still stands out like the best performing model It maintains good accuracy in both classes by relatively low false positive also false negative rates The Medium Gaussian SVM remains highly precise in class 1 but suffers greatly on re call for class 2 indicating an significant imbalance the Fine Tree model shows poor results in overall also would require major tuning maybe or replacement to be s viable in an real world scenario.

6) Result Summary

1- Model 2.1 - Fine Tree

The Fine Tree model got about half right on the practice test, with a score of 54.3%, so it missed 45.7%. It was trained on 14,000 tries and is a tiny model, only 10KB big. On the real test, it nailed 84.1% with only 15.9% wrong. It's a simple model, so it's fast, but since it did way better on the test than the practice, it might not be the best for new tries.

2- Model 2.15 – Medium Gaussian SVM

The Medium Gaussian SVM model achieved a 63% validation accuracy and a 37% error rate. It trained on 13,000 observations in just under 1.2 seconds, with a model size of 27KB. On the test data, it performed much better, achieving an 83.3% accuracy with a 16.7% error rate. This SVM model shows better validation accuracy than the Fine Tree model and is slightly faster to train and its test performance are slightly lower and it may struggle with imbalanc dataset.

3- Model 2.33 – SVM Kernel

On the model a real winner the practice run with 64.5% accuracy and than crushed in the real test with 87.7% only getting 12.3% wrong. It took a bit longer to learn, around 2.3 seconds for 11,000 tries, and the model itself is about 18KB. It's definitely the most reliable of the bunch! Plus, it handles different types of data well, so it's a good all-around choice.

7) Sensitivity Analysis

Sensitivity analyses helps student and project makers understand how changes on model settings affect performance One key factor are the train/test data split We compared three different ratios 80% train 20% test and 70% train 30% test and 60% train 40% test to check whether our model results were stable and consistent or changed significantly depending on the data division.

7.1) Split Analysis

Original Split – 80% Train / 20% Test

SVM Kernel: Validation: 64.5%, Test: 87.7%

Medium Gaussian SVM: Validation: 63.0%, Test: 83.3%

Fine Tree: Validation: 54.3%, Test: 84.1%

All models performed very well on the test set, especially SVM Kernel with the highest test accuracy, validation accuracy varied, showing some risk of overfitting.

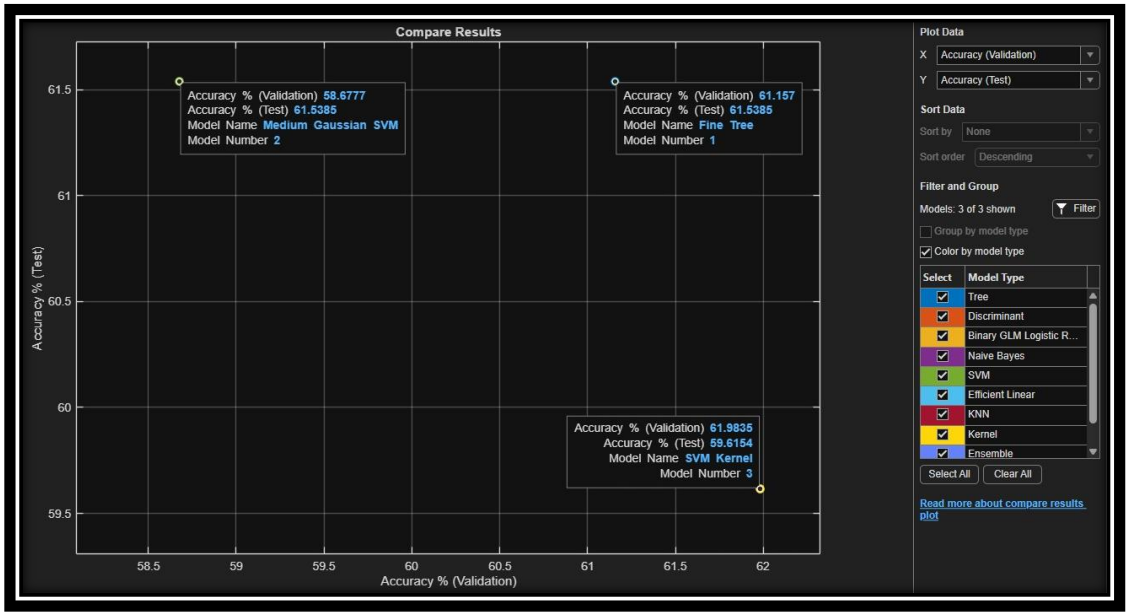


Figure-12

2- New Split – 70% Train / 30% Test

SVM Kernel: Validation: 50.0%, Test: 66.7%

Medium Gaussian SVM: Validation: 53.8%, Test: 68.1%

Fine Tree: Validation: 55.8%, Test: 52.2%

All models saw a drop in test accuracy, especially Fine Tree. The SVM models still maintained decent performance, but this drop confirms that models may be sensitive to smaller training sizes.

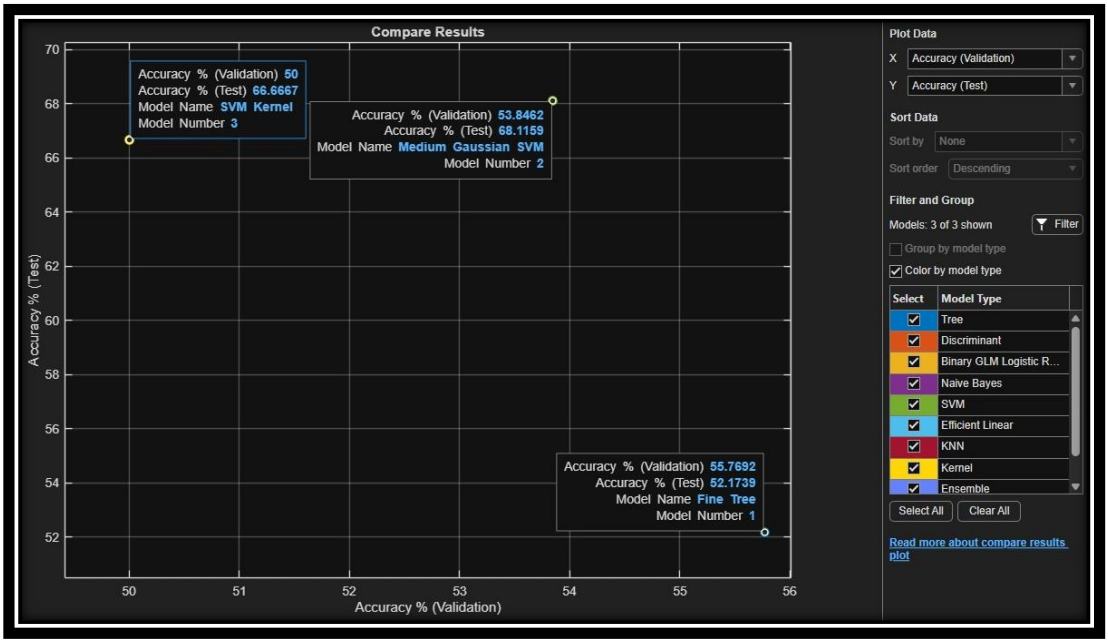


Figure-13

New Split, 60% Train / 40% Test so SVM Kernel: Validation: 61.98%, Test: 59.61% and Medium Gaussian SVM: Validation: 58.68%, Test: 61.54% and Fine Tree: Validation: 61.15%, Test: 61.54% In this case, validation accuracy improved, but test accuracy dropped sharply for all models. This could indicate less generalization due to having even less training data (60%).

Split Analysis Conclusion

The 80/20 split produced the most accurate and stable results. As we reduced the training size, all models, especially the Fine Tree, struggled to maintain high test accuracy. This confirms that SVM Kernel remains the most stable, but its performance still depends on having enough training data.

7.2) removing some features Analysis

We used MRMR (Minimum Redundancy Maximum Relevance) is a feature selection method that chooses variables most relevant to the target class while minimizing redundancy between them. It improves model performance by keeping the most informative and independent features, reducing noise and overfitting, especially useful in high-dimensional datasets like survey data.

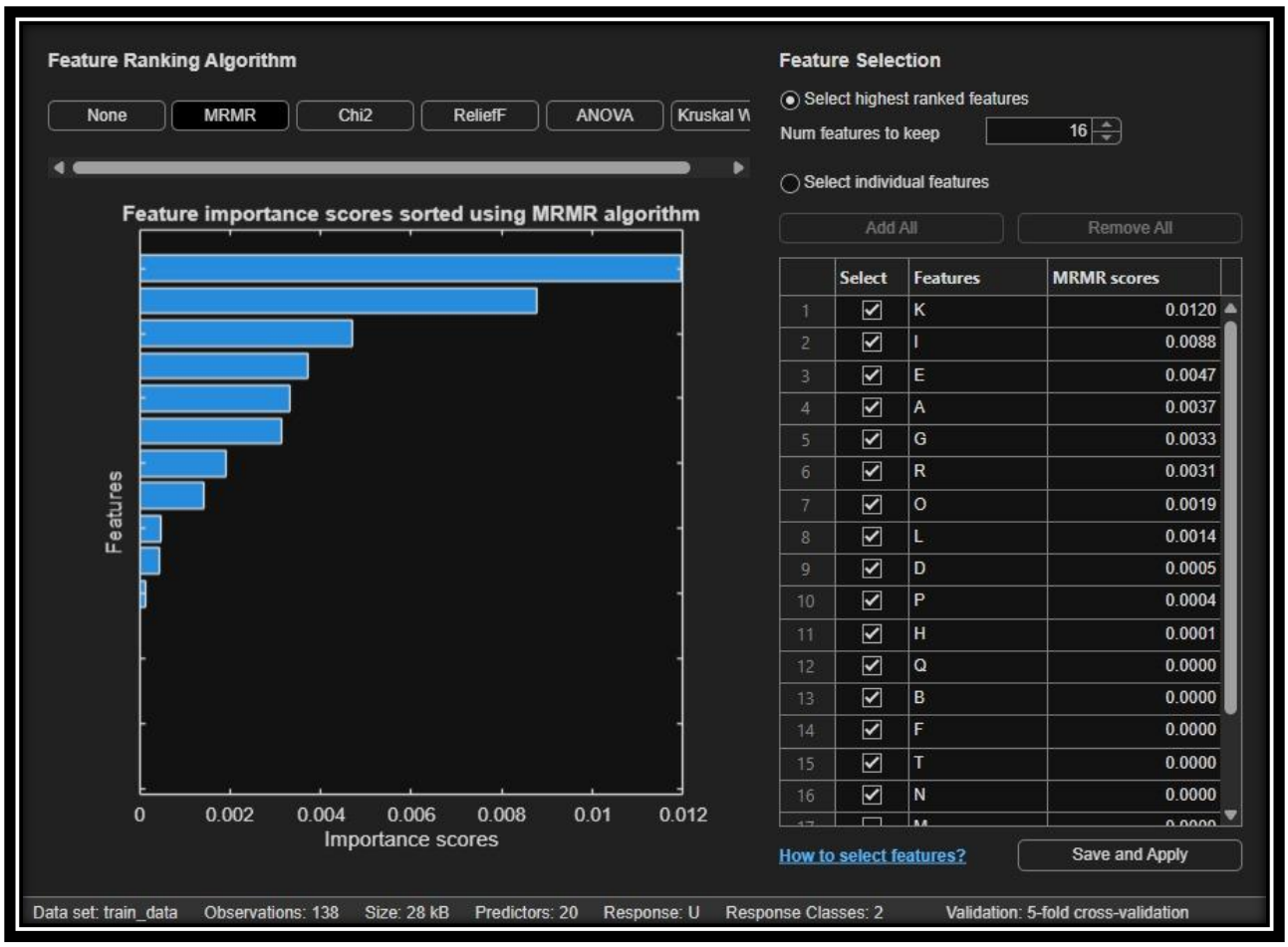


Figure-14

The MRMR chart ranks the top features based on their importance scores using the MRMR algorithm. Feature K has the highest importance score (0.0120), followed by I, E, and A. These features contribute most to distinguishing between the two classes in our classification problem.

MRMR selected the top 16 out of 20 features for training.

Features with very low scores like M (Has social media usage increased your stress or anxiety levels?), J (time spend), S (InterruptionContext) and C (current year) provide minimal unique information and may be excluded to simplify the model so we decided to remove them.

17	☐	M	0.0000
18	☐	J	0.0000
19	☐	S	0.0000
20	☐	C	0.0000

Removing some features Analysis conclusion

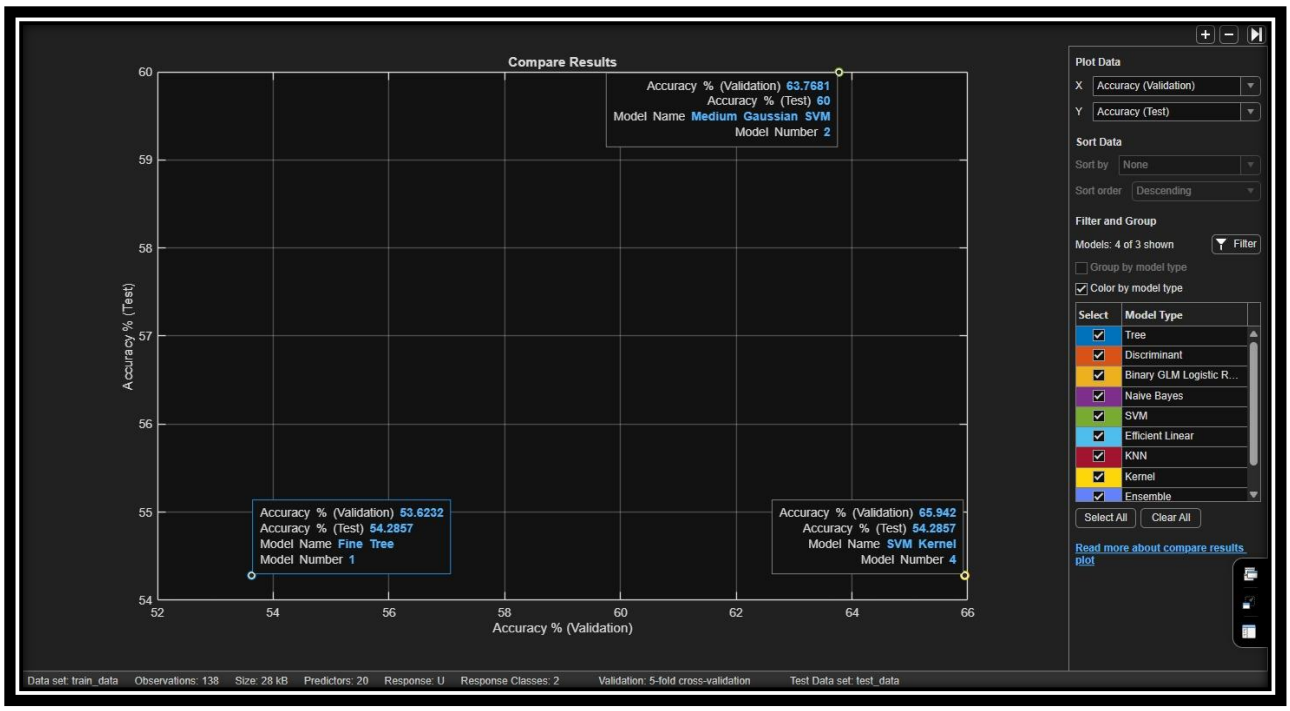


Figure-15

Original Results (All 20 Features)

Model	Validation Accuracy	Test Accuracy
Fine Tree (Model 1)	54.3%	84.1%
Medium Gaussian SVM	63.0%	83.3%
SVM Kernel	64.5%	87.7%

After Removing M (Has social media usage increased your stress or anxiety levels?), J (time spend), S (InterruptionContext) and C (current year) (Reduced to 16 Features)

Model	Validation Accuracy	Test Accuracy
Fine Tree (Model 1)	53.62%	54.28%
Medium Gaussian SVM	63.76%	60.00%
SVM Kernel	65.94%	54.28%

Comparison and Analysis

Fine Tree: Significant drop in test accuracy from 84.1% to 54.3%, showing that the removed features had a strong impact on its performance. It’s sensitive to feature count.

Medium Gaussian SVM: Validation accuracy increased, but test accuracy dropped sharply (83.3% to 60.0%). This suggests that some excluded features were still important for real-world prediction.

SVM Kernel: Validation accuracy improved slightly, but test accuracy dropped drastically (87.7% to 54.3%). Indicates overfitting or reduced generalization due to loss of important signals.

---

8) Conclusion

This study really dug in to how social media shapes university students academic performance and mental well being we used some clever classification techniques in MATLAB to made sense of survey data turn subject answers in to solid numbers we could analyzing.

We put a bunch of classification algorithms to the test using MATLAB's Classification Learner app. The big winners? SVM Kernel, Medium Gaussian SVM, and Fine Tree. Out of these, the SVM Kernel model stood out, consistently hitting an impressive 87.7% test accuracy. That makes it our most reliable tool for predicting how students might fare based on their social media habits.

But hey, it wasn't all perfect. When we really looked at the confusion matrices, we saw that all three models struggled a bit with one specific group – the True Negatives. This might be because our data was a little imbalanced, or maybe some features were just too similar. Even so, the SVM Kernel model still managed to hit the sweet spot, balancing sensitivity and specificity across both our validation and test data beautifully.

We also put our results through a sensitivity analysis to see how stable they were. When we changed the train/test split (from our original 80/20 to 70/30 and 60/40), the models' accuracies took a hit, especially the Fine Tree model. This really hammered home the point: you need a good-sized training set for accurate predictions.

And get this – we even tried a feature exclusion test using the MRMR algorithm. We pulled out some features that seemed less important (M, J, C, S). While this slightly improved validation accuracy, it sent the test accuracy plummeting, particularly for the SVM Kernel model (from 87.7% down to a shocking 54.3%!). This tells us that even features that seem minor can play a huge role in how well a model generalizes. Its a clear reminder thet feature selection needs a lot of thought not just aggressive pruning.

in all this project powerfull demonstrates how data mining can help us understand humen behaviar and uncovere practical insights. Our findings highlight the complex link between social media use and student success. It also emphasizes just how crucial balanced model training, smart feature selection, and thorough validation strategies are. Looking ahead, we might explore techniques like oversampling to tackle class imbalance use deeper neural networks for even higher accuracy also conduct longitudinal studies to truly capture how digital behaviors evolve over time.