Student Behavior

The Bayesian Believers team

Salma Muhammed Entsar

Ahmed Kamal Abdelrasol

Ahmed Moustafa Attia

Nesma Osama

Mazen Adel

Muhammed Ashraf

Table of Contents

Contents

Tables of Figures: 3

Introduction: 6

Descriptive statistics journey about our data: 8

A-Pie Charts: 9

Comments on that part: 9

B-Bars: 10

C-Histograms: 10

Comments on that part: 10

D-Correlations: 11

E-general statistics about our data: 11

Methodology part: 12

What is the probability of stable financial status for students who work part-time? 12

Can we predict college marks by high school marks? 13

Does a student's college mark suggest their stress level? 13

Is there a correlation between high school marks and college marks? 14

Does stress level decrease with traveling? 14

Is there a correlation between continuous variables like Social media and academic performance? 14

Numerical Analysis part: 15

1-High school marks versus college marks: 15

A-getting correlation between High school marks and college marks: 15

B-Linear Regression Model: 16

2- What is the probability of stable financial status for students who work part-time? 17

A-obtaining values for the Frequencies table for naïve Bayes: 17

B-Naïve Bayes Classification: 18

3-Does a student's college mark suggests their stress level? 18

A-Metrics: 19

Confusion metrics : 19

TABLE OF CONTENTS

Accuracy Score & Precision & Recall: 20

Random Forest Classifier Model: 20

4-Is there a correlation between continuous variables like Social media and academic performance? 20

Getting a correlation between social media hours and college marks: 20

B-Linear Regression Problem with that model: 21

5- Does stress level decrease with traveling? 22

A-Metrics: 22

Confusion metrics : 22

Accuracy Score & Precision & Recall: 23

B-Logistic Regression Model: 23

Conclusion Part: 24

Final Message: 26

References: 27

helping tools for report: 27

Appendix: 27

TABLE OF CONTENTS

Tables of Figures:

Figure 1:cell chart of dataset part 1 8
Figure 2:Cell Chart part2 8
Figure 3:Correlations 11
Figure 4:descriptive statistics values 11
Figure 5:Financial Status Categories 12
Figure 6:correlation needed values 16
Figure 7:value of correlation 16
Figure 8:Linear Regression Plotting 17
Figure 9:Categories of Financial Status 17
Figure 10:Part Time jobs categories 17
Figure 11:Frequency table for naive algorithm 18
Figure 12:Naive Bayes Model 18
Figure 13:naive Bayes model 18
Figure 14:traing set frequency 19
Figure 15:teasting set frequency 19
Figure 16:Confussion metrics 19
Figure 17:Histogram of frequencies 20
Figure 18:Random forest classifier model 20
Figure 19:Calculations needed for Correlation 21
Figure 20:correlation between social media and marks 21

Abstract

Students comprise a substantial portion of our population. They serve as the next generations' skeletons. The process of learning is not simple. Does any student ever ask themselves if learning or feeling something is their primary objective? How am I going to process? These inquiries were made by numerous students, whose conduct is influenced by various factors. Out of all the ideas we came across as students, we chose to keep an eye on a select few that affected our investigation.

We will travel in a planned manner because we are addressing uncertainty around the solution. We will use probability and statistics concepts at each stop along the way to help us discover the answers. By collecting data on a set sample size of students, we were able to identify our population. We probed them extensively to learn about several tenses, including grades, study time, and so on.

Using pies, bars, and histograms, we first start with descriptive statistics to make important inferences about the data's skewness, trends, tendencies, mean, standard deviation, and other relevant subjects. We classified our features as dependent and independent based on those. We wrap up a few connections.

Abstract

can assist us in determining which model or methodology to use in order to find our primary answers.

Artificial intelligence is a vast field with numerous subfields. It involves using a collection of algorithms to try and find solutions and then generalizing them for larger purposes. Here, we made use of numerous machine learning algorithms, particularly supervised ones. As we have previously stated, it is too difficult to survey every student on the planet; nevertheless, we can use our predictions on a small sample size before extrapolating them to larger ones. In order for the model to be able to generalize or respond to any pattern on a population other than our own, we have employed supervised learning, which is dependent on labeled data from "our dataset."

Ultimately, we will contrast our numerical analysis with our model's prediction in order to draw conclusions and extrapolate certain generalizations regarding student behavior from the numerical side of life.

Introduction

Introduction:

As students for more than three years in university, we've seen a lot – our own stories and those of our friends. We've learned about different aspects of student life and what influences it. This motivated us and got us thinking about helping new students. We wanted to figure out the key to getting good grades while still enjoying life outside of studies.

So, we did a study. We looked at various parts of a student's daily life like hobbies, money situations, how far they live from the university, and more. Then, we compared all this info with their grades, happiness levels, and other interesting things. Our goal was to find out what makes students successful..

Introduction

Our data comes from 235 students. It includes details like gender, high school, and college grades, hobbies, study time, where they like to study, if they like their degree, if they want a job related to their degree, how much they use social media, how long they travel for studies, how stressed they feel, and their financial situation. This data helps us understand how students behave and live.

Now, let's break down what some of the terms mean in our data:

Secondary Mark: This is the grade a student got in the 12th grade.

College Mark: It's the grade a student gets in their college or university.

Hobbies: This is what a student likes to do in their free time.

Daily Studying Time: It's how much time a student spends studying every day.

Prefer to Study in: This tells us where a student likes to study.

Do you like your degree? It's whether a student enjoys what they are studying.

Willingness to pursue a career based on their degree: This tells us if a student wants a job related to what they are studying.

Social Media & Video: This shows us how much a student uses social media and watches videos.

Traveling Time: It's how long a student takes to travel to their college.

Stress Level: This is how much stress a student feels.

Financial Status: It's about how much money a student has or their family's economic situation.

Introduction

Part-time Job: This tells us if a student works part-time while studying.

Each of these points helps us understand a different aspect of a student's life and behavior.

Here are some interesting visualizations using descriptive statistics that interestingly describe our dataset.

Descriptive statistics journey about our data:

First, let's identify the elements of the dataset that piqued our curiosity and motivated more research.

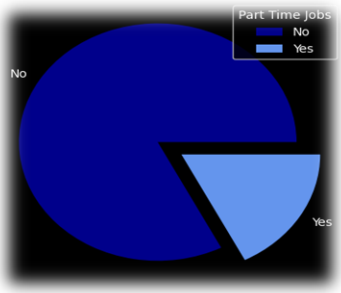
FIGURE 1:CELL CHART OF DATASET PART 1

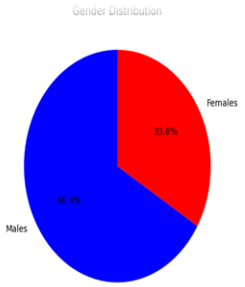
FIGURE 2:CELL CHART PART2

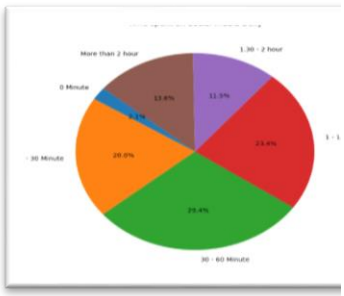
We will also provide additional details about our data and how we organized them into bars, pies, and histograms.

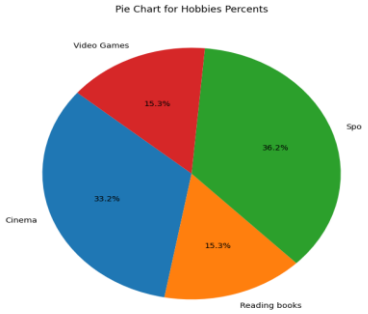
Introduction

A-Pie Charts:







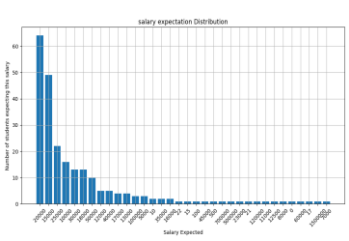


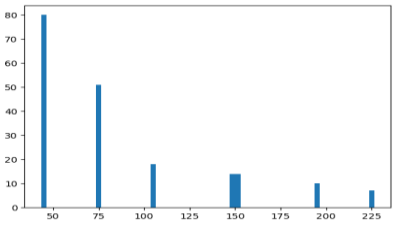
Comments on that part:

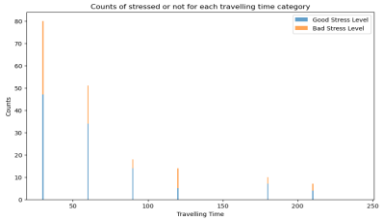
First of all, men made up the majority of our population. We will also see that the majority of students enroll in certification programs, and that while their financial situation is generally favorable, only a small percentage of them live in poverty, according to our graphical data. But since most of them work part-time jobs, we became curious about why they are required to do so. In addition, a large portion of them reported that sports were their favorite pastime, which does not imply that studying interfered with it. We also found that the majority of them spent one hour a day on social media, refuting claims that they were addicted.

Introduction

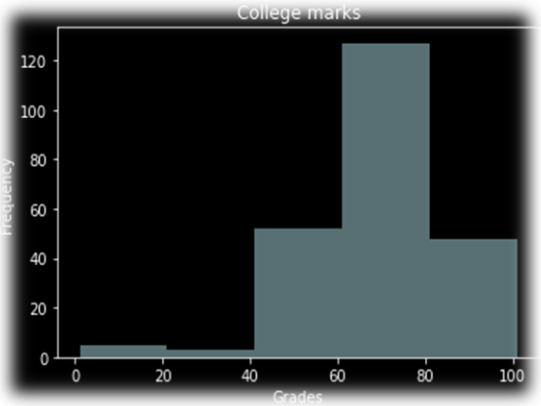
B-Bars:

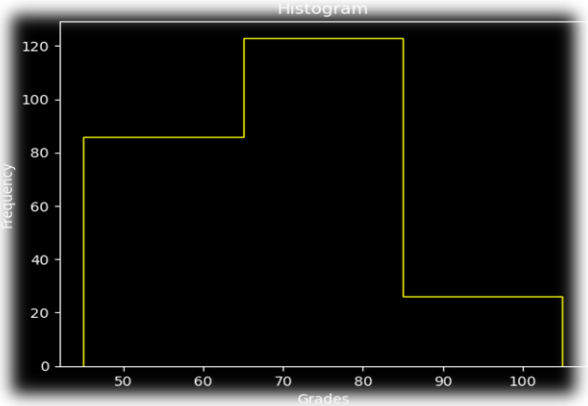






C-Histograms:





Comments on that part:

Here, we questioned students about their projected earnings in order to observe that the distribution of expected salaries is skewed appropriately in favor of higher incomes, indicating that financial considerations are taken into account. The second bar shows the disproportionate number of people who have travelled. The final graph compares the amount of time spent traveling and stress levels, and it is also true that most of them find that traveling reduces stress.

Introduction

D-Correlations:

Right now Let's examine the link between a few characteristics of our data sets after discovering some visualizations.

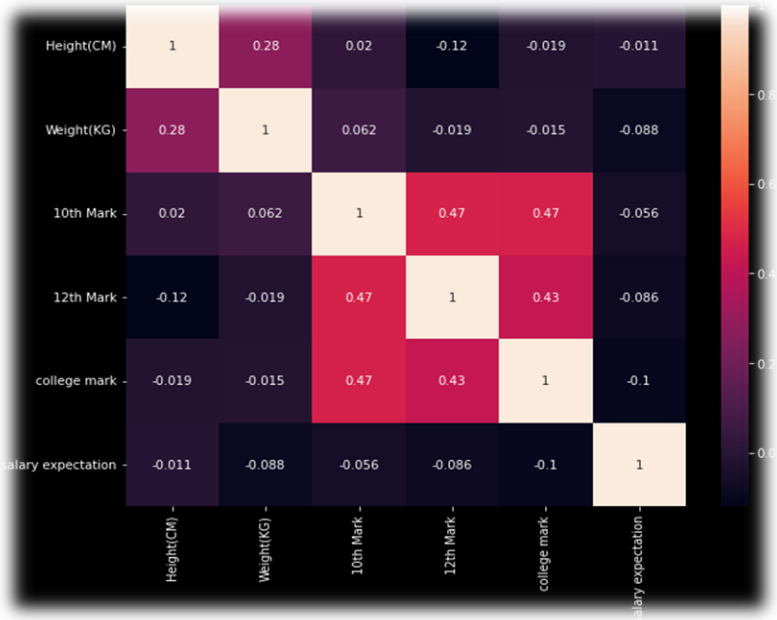


Figure 3:Correlations

This graphic will display the mapping of their fundamental relationships.

The diagonal indicates a correlation of one to each variable and to itself.

The connection's strength is shown by the colors; a positive correlation has a light color. This indicates that they have an inverse correlation as it grows darker.

E-general statistics about our data:

The generic descriptive statistics in this section regarding our dataset will be useful for certain numerical procedures.

Height(CM)		Weight(KG)		12th Mark		college mark		salary expectation		willingness to pursue d	
Mean	157.4021	Mean	60.80383	Mean	68.78013	Mean	70.66055	Mean	32481.68	Mean	0.695745
Standard I	1.40321	Standard I	0.971698	Standard I	0.718639	Standard I	1.025945	Standard I	7261.366	Standard I	0.013815
Median	160	Median	60	Median	69	Median	70	Median	20000	Median	0.75
Mode	160	Mode	60	Mode	70	Mode	70	Mode	20000	Mode	0.75
Standard I	21.51081	Standard I	14.89584	Standard I	11.01653	Standard I	15.72745	Standard I	111314.6	Standard I	0.211787
Sample Va	462.7147	Sample Va	221.8862	Sample Va	121.364	Sample Va	247.3526	Sample Va	1.24E+10	Sample Va	0.044854
Kurtosis	12.00423	Kurtosis	-0.3509	Kurtosis	-0.63836	Kurtosis	4.979634	Kurtosis	136.4474	Kurtosis	0.205987
Skewness	-2.57943	Skewness	0.248998	Skewness	0.068273	Skewness	-1.6297	Skewness	11.12667	Skewness	-0.50523
Range	187.5	Range	86	Range	49	Range	99	Range	1500000	Range	1
Minimum	4.5	Minimum	20	Minimum	45	Minimum	1	Minimum	0	Minimum	0
Maximum	192	Maximum	106	Maximum	94	Maximum	100	Maximum	1500000	Maximum	1
Sum	36989.5	Sum	14288.9	Sum	16163.33	Sum	16605.23	Sum	7633195	Sum	163.5
Count	235	Count	235	Count	235	Count	235	Count	235	Count	235
Confidenc	2.76454	Confidenc	1.914394	Confidenc	1.41583	Confidenc	2.02127	Confidenc	14306.01	Confidenc	0.027219

Figure 4:descriptive statistics values

Methodology

Methodology part:

Upon examining our data, we observed demographics, bars, relationships, and even genetic statistics. Based on the connection and dependency between our aim and data, we begin to develop the appropriate models to answer questions.

We will discuss a number of approaches that are employed both inside and outside of our curriculum in our second chapter in order to address a range of inquiries pertaining to our primary viewpoint. Based on our prior data, there are numerous and diverse relationships in it, from which machine learning models can infer a wide range of conclusions. Let's give them some serious thought.

What is the probability of stable financial status for students who work part-time?

As we previously mentioned, the pie chart indicates that the majority of them have strong financial status. The Financial Status feature comprises four values, each of which corresponds to a particular financial status. Thus we're curious as to why they accept part-time work as well. Thus, we shall talk about this concept of conditional probability first and foremost about financial status categories.

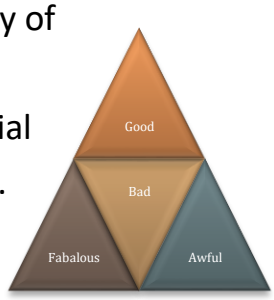


Figure 5:Financial Status Categories

As we said significant idea in probability depends on circumstances. It determines an event's probability based on a previous circumstance or occurrence.

Machine learning has a model that depends on it “Naïve Bayes”.

Methodology

Naïve Bayes classifies them based on our prior financial status pie categories in light of any part-time work we may have, we will attempt to apply it here. In the Numerical analysis section, we will go deeper into solution details. and there will be a scikit learn model in the appendix its output will be provided also in the Numerical analysis section.

Can we predict college marks by high school marks?

Three different mark kinds are included in our dataset; their histograms have already been shown. Our inquiry is to determine whether or not a student's high school grades truly indicate that they will continue to have good grades in college. Since our goal is the collage mark and we have determined that the 12th grade is an independent variable, we will employ linear regression to predict the collage mark with ease.

Linear regression is a method for predicting a variable's value based on the value of another variable to utilize linear regression analysis. Here, we'll utilize it to develop a machine-learning model that forecasts college mark values according to students' high school marks and college marks.to deduce the continuity of their progress in university grades in an effort to facilitate finding an answer.

Does a student's college mark suggest their stress level?

We always here that a student got depressed when encountering bad marks on exam here we will find out about that.

Methodology

Random forest classifier will also assist us in determining how to divide our sample into four categories based on their marks pursuit: those who have stress levels

Is there a correlation between high school marks and college marks?

The correlation will help us deduce that those students who got high marks in high school can continue with that progress even in college, or vice versa.

Does stress level decrease with traveling?

Here, logistic regression will also assist us in determining how to classify stress levels with the corresponding travel duration and we will use mean to seek the least values of our cost function.

Is there a correlation between continuous variables like Social media and academic performance?

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. Here we will use it to deduce more about the relation between study hours' ranges and the college mark of the student, and the same applies to social media.

Numerical Analysis

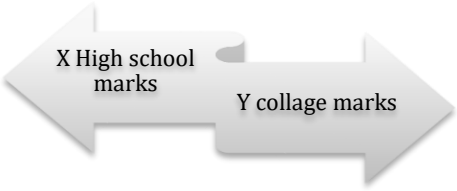
Numerical Analysis part:

After visualizing our data and analyzing its trends in light of our selected methodologies, we will have a more in-depth discussion with you regarding our models. Here, we encountered several numerical pricing and provided them with our predictive model's graphs in an attempt to find the answers to our questions.

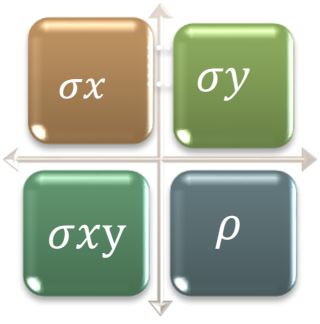
1-High school marks versus college marks:

As we previously discussed, I will employ a linear regression model of supervised machine learning to predict the college mark with ease. This mark is reliant on the high school mark, and it will allow me to ascertain if the kid will continue to perform well at college or not. Initially, we also presented the correlations among the variables in our dataset; however, we will delve more into the calculation of those correlations in this part.

A-getting correlation between High school marks and college marks:



Here there are two variables we need to get 4 important variables.



Numerical Analysis

StandardDeviation12Mark	StandardDeviationCollege	covariance	correlation coefficien
11.01653293	15.72744632	73.65759779	0.425122805

Figure 6:correlation needed values

After we calculated those values we got that value of correlation is above zero which means that they are positively proportional.

Figure 7:value of correlation

B-Linear Regression Model:

Here we used the least square method of linear regression.

Linear Regression Results

Metric	Value
Residuals	[-71.61, -5.33, 1.46, 8.6, 26.04]
Coefficient	[28.92, 0.61]
Residual Estimate Error	14.21
Multiple R-squared	0.1807
Adjusted R-squared	0.1772

S Values

Metric	Value
Sxx	28399.18
Sxy	-1.73e-10
Syy	57880.5

Comments:

These are the values required for model testing.

The training data set consisted of 235 rows, and the table below shows the coefficients of the linear line. R squared values are little and marginally good. However, let's also observe the scattering.

```
Cost Function (MSE): 201.78635184478057
```

This value of error is lightly high but let's see why this happens in the plotting section

Numerical Analysis

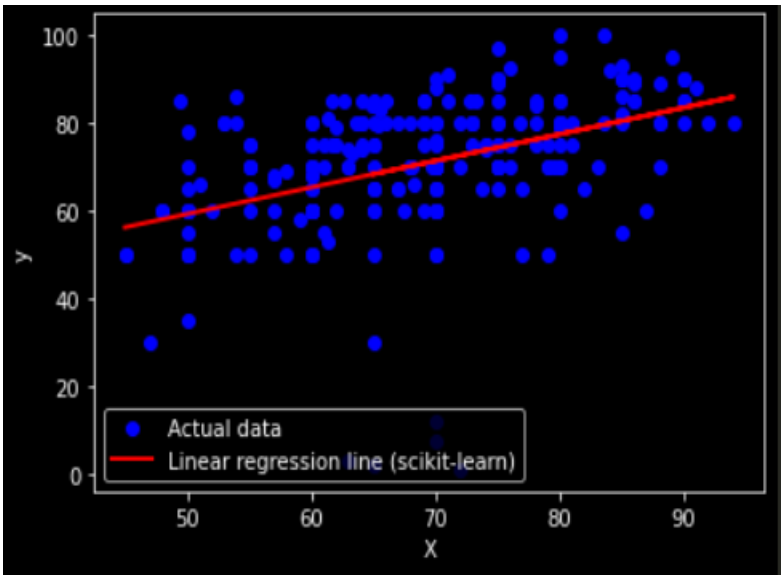


Figure 8:Linear Regression Plotting

It will be observed that there are numerous points that are somewhat distant from the line and that the scattering is more concentrated around the line, which led to a slightly elevated cost function.

This indicates that the quality of students' work has improved in collage.

2- What is the probability of stable financial status for students who work part-time?

As we previously stated, naïve Bayes will be used to classify students' financial situation despite the fact that they work part-time employment.

A-obtaining values for the Frequencies table for naïve Bayes:

Let's first recap our input and outputs to the model.

Financial Status into 4 categories

Take part-time jobs into 2 categories

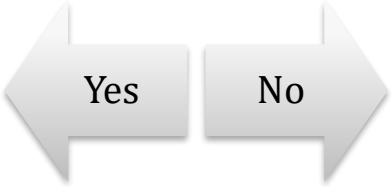


Figure 10:Part Time jobs categories

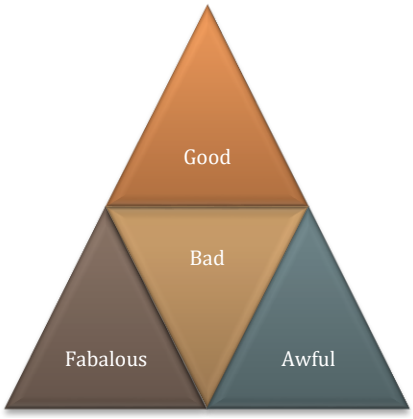


Figure 9:Categories of Financial Status

Numerical Analysis

These are the frequency table results for every probability

Finanical	Yes	NO	P(Finanical|Yes)	P(Finanical|No)
Bad	17	71	0.414634146	0.365979381
Good	18	111	0.43902439	0.572164948
Awful	4	10	0.097560976	0.051546392
Fabulous	2	2	0.048780488	0.010309278
NumOfPartTime	41	194		

Figure 11:Frequency table for naive algorithm

B-Naïve Bayes Classification:

The data show that students in both good and bad financial status work part-time jobs; those in the best financial standing are the least likely to do so. It could be a terrible situation if you work part-time jobs yet lack the funds to study for students with awful financial status.

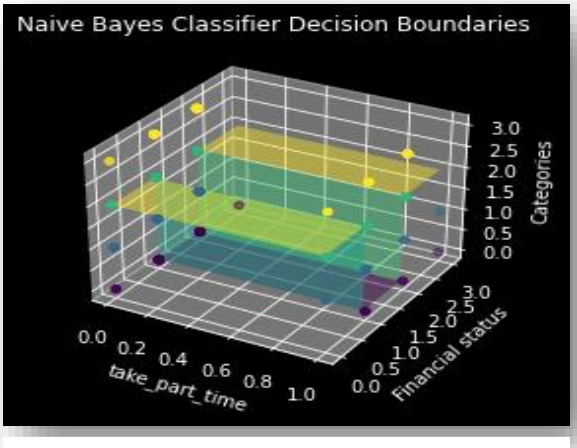
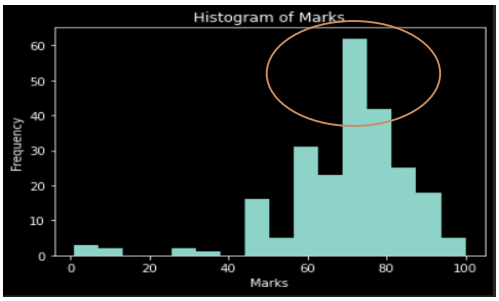


Figure 12:Naive Bayes Model

3-Does a student's college mark suggests their stress level?



Let's recap the histogram of college marks first to show the most frequent marks category.

This shows the category of marks between [60-80]is so high.

Numerical Analysis

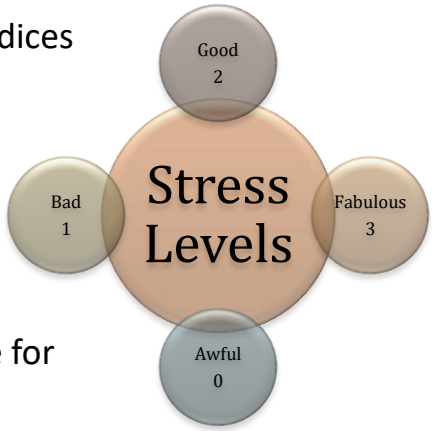
A-Metrics:

Let's talk about the metrics I used to determine the quality of my model.

Our stress levels as measured by the corresponding indices

Additionally, this table displays their frequency in our training data.

After that, I used the split method to split my training data into 2 categories one for the prediction and the other for testing the model the freauncy table for testing sample.




```
Class Distribution in Original Data:
2    137
1     68
0     19
3     11
dtype: int64
```

FIGURE 14:TRAING SET FREQUENCY

```
Predicted Class Distribution:
1    29
2     9
0     6
3     3
dtype: int64
```

FIGURE 15:TEASTING SET FREQUENCY

Confusion metrics :

It is used to compare the predicted values from our models and the original ones.

These metrics suggests that, according to our model, most stress levels fall between the good and terrible categories. This could be because some students aren't content with 60 points, while others think it's sufficient, but it still shows that they're striving for the highest degree.

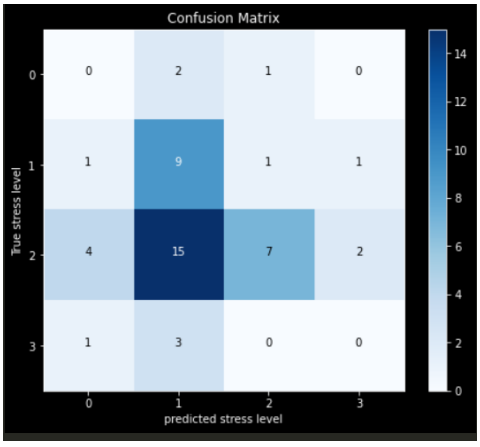


Figure 16:Confussion metrics

Numerical Analysis

Accuracy Score & Precision & Recall:

```
Accuracy: 0.3404255319148936
Precision: 0.5425939512513247
Recall: 0.3404255319148936
```

These values indicate that our model of random forest classifier works slightly well.

Random Forest Classifier Model:

Lastly, I will display the decision boundaries of our models to you.

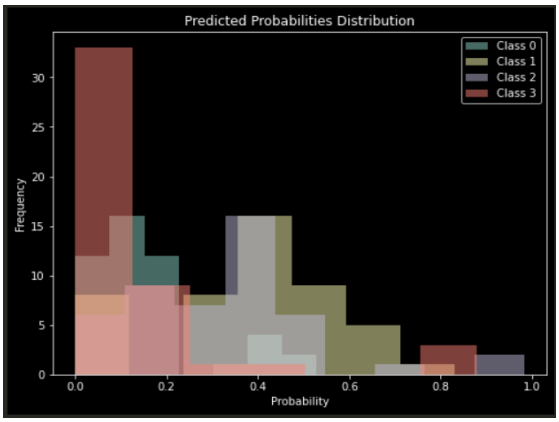


FIGURE 17:HISTOGRAM OF FREQUENCIES

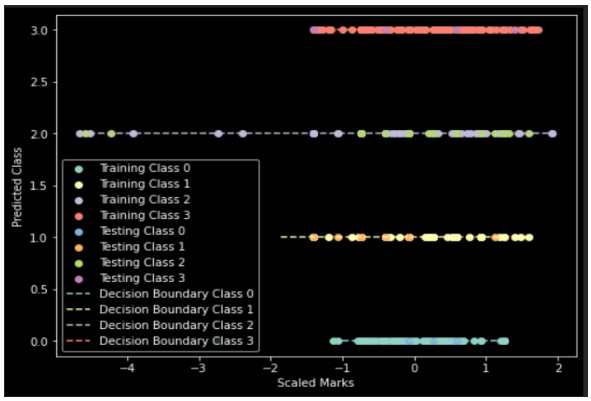


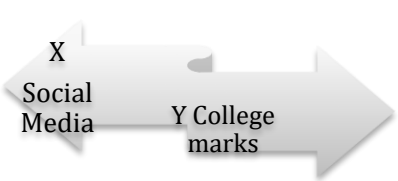
FIGURE 18:RANDOM FOREST CLASSIFIER MODEL

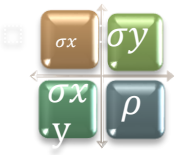
4-Is there a correlation between continuous variables like Social media and academic performance?

I would like to look into the assertion that increasing the amount of time spent on social media will result in worse marks. before, we will apply a correlation between those two variables, our response, and target, but before, let's distinguish them once more.

Getting a correlation between social media hours and college marks:

Those are our response x and target y that were sent to our model The other 4 variables are what we need to calculate.





Numerical Analysis

```
Mean of college grades :  70.66055319148937  standard deviation of college grades :  15.693947990348585
Mean of Social media time :  85.70212765957447  standard deviation of Social media time :  53.75116619956715
E(XY) =  6061.061276595744   Cov(XY) =  5.301526482569898
Correlation :  0.0062846462171480206
```

Figure 19:Calculations needed for Correlation

After calculating those numbers, we discovered that the correlation value was almost zero, indicating that the students are independent of one another and that using social media does not necessarily result in a good or bad grade.

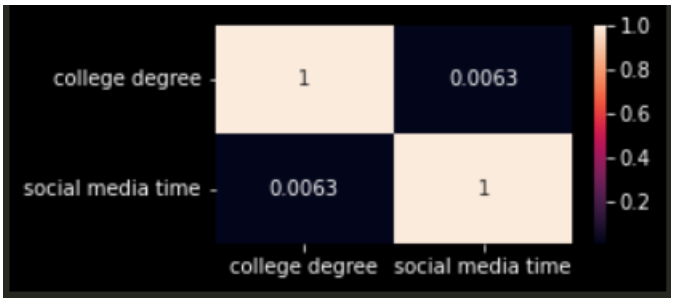


Figure 20:correlation between social media and marks

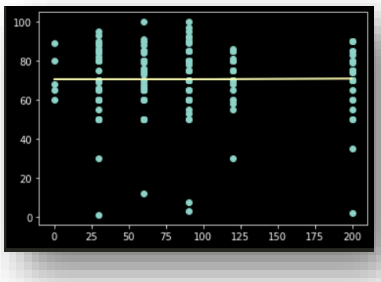
B-Linear Regression Problem with that model:

I am utilizing this model solely to demonstrate the independence of those two variables through the use of plotting, scattering, and correlation analysis of significant variables required for linear regression. We can observe that the MSE values are extremely high, indicating that there isn't a linear relationship between these variables.

```
Coefficient:
 [7.05032937e+01 1.83495388e-03]

Multiple R-squared: 3.9496778074687 99e-05

Adjusted R-squared: -0.004252179201418738
Sxx: 678959.1489361703
Syy: 57880.5008280851
Sxy: 1245.8587234042543
Sum of Squared Residuals (SSE): 57878.21473478904
cost: 246.29027546718743
```



The claim is rejected since, as we can see, the scattering does not indicate any linearity between them and the line is unable to fill all of the scattered spots.

5- Does stress level decrease with traveling?

A-Metrics:

Let's talk about the metrics I used to determine the quality of my model.

Our stress levels as measured by the corresponding indices

Additionally, this table displays their frequency in our training data.

After that, I used the split method to split my training data into 2 categories one for the prediction and the other for testing the model the freauncy table for testing sample.

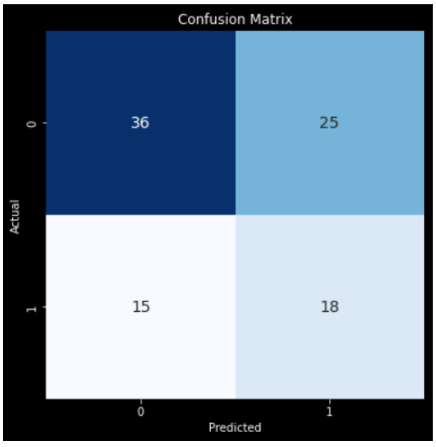

```
Class Distribution in Training Dat
0    87
1    87
dtype: int64
```

```
Predicted Class Distribution in Testing Dat
0    51
1    43
dtype: int64
```

Confusion metrics :

It is employed to contrast the initial values with the forecasted values from our models.

These measurements imply that our model expected terrible stress levels to be more accurate than good stress levels and that good stress level forecasts were likewise very accurate. this speaks to the caliber of our model.



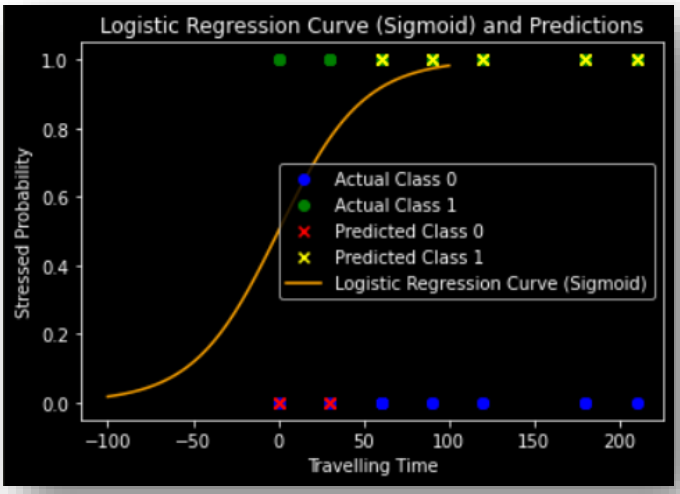
Numerical Analysis

Accuracy Score & Precision & Recall:

```
Log Loss (Cross-Entropy Loss): 0.6911815298495703
Accuracy: 0.574468085106383
Precision: 0.4186046511627907
Recall: 0.5454545454545454
```

These values indicate that our model of random forest classifier works slightly well.

B-Logistic Regression Model:



This is a plot of our sigmoid function, which was utilized to determine the expected values.

The labels display the actual and anticipated values, and based on our model, we may infer that students' stress levels are impacted by travel.

Now that we have completed all necessary numerical analysis, let us summarize the conclusions drawn from our findings in the following parts.

References

Conclusion Part:

We were able to find some solutions by using the previously indicated procedures of data collection, characterization, numerical analysis, and prediction model development. so it makes sense to claim that we have identified a few trends.

When we used a linear regression model to reach our conclusion, we discovered a strong positive correlation between high school and college grades, which supported our hypothesis. This indicates that most students would continue to make the same progress even when they faced more challenging circumstances and curricula.

References

Additionally, naive Bayes helps us to address the issue of whether or not students with strong financial standing can work part-time jobs. In reality,

students with high financial standing can work part-time jobs, too. Maybe they wish to take responsibility early, but our approach only identifies those categories with a somewhat high accuracy.

We also talked about the fact that while many people claim that kids are addicted to social media, the pie charts initially indicate a moderate usage rate, so we tried to refute this assertion and look for a correlation between those variables Students may spend more time studying than using social media between those variables, and our linear regression model ensures the independence between those, therefore the claim was also denied.

We also talked about how the kids' marks affected them and increased their

stress levels for the next tests. As the bulk of their grades are below average, some students are unaffected by this development, while others may be struggling to improve their performance in any case, which raises their stress levels and ultimately results in poor grades.

References

Additionally, compared to students who are unable to take vacations, we found that those who traveled and spent summer vacation experienced lower levels of stress.

In fact, we can say that our effective method has helped pupils identify new patterns in their life.

Final Message:

We have finally arrived at the conclusion of our voyage, having shared all the joys and sufferings of the kids. We have given a lot of thought to what is happening, how it affects their mental health, what they want, and everything else.

Perhaps we can look into them more and more, and in fact, the research was excellent in helping us as students understand what to look for.

References:

Machine learning models helping concepts:

https://en.wikipedia.org/wiki/Machine_learning

https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.

https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.

https://www.kaggle.com/datasets/gunapro/student-behavior

https://www.datacamp.com/tutorial/naive-bayes-scikit-learn#:~:text=Naive%20Bayes%20is%20a%20statistical,and%20speed%20on%20large%20datasets.

helping tools for report:

https://quillbot.com/

Appendix:

This the link of our whole python notebook with all codes

https://drive.google.com/file/d/18bZazIaBD9xoy5g35mZR1Mh0nRUPa2vB/view?usp=drive_link