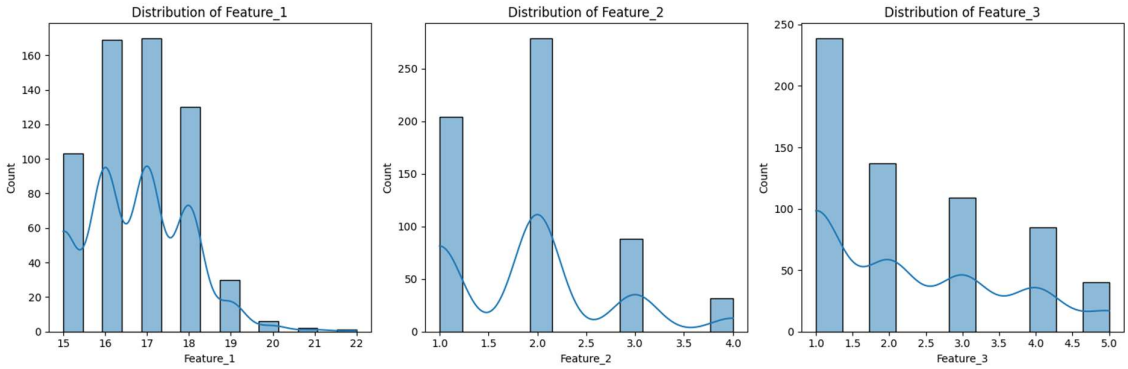
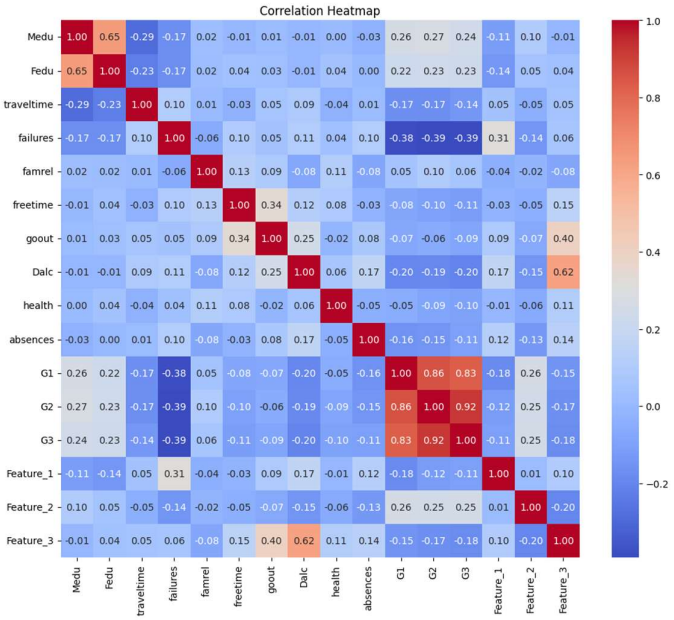
CampusPulse

- **Level 1**

  Firstly, we made histplots for Feature 1, Feature 2 and Feature 3 and had an idea of values of each.



  Feature 1 has values from 15-22 with most of the data being in the range 15-18, most likely representing Age. Then constructed a correlation matrix for numerical values.



  We see that Feature 1 has a strong correlation index with “Failures” which reflects that Feature 1 is age.

  Feature 2 has a strong correlation index with G1, G2 and G3 and also a negative index with failures, goout, Dalc. The most probable feature that it can denote is the Study hours (1 to 4) 1 being the least and 4 being the most

  Feature 3 has a VERY STRONG correlation with Dalc and goout and a negative correlation with grades. It should represent the Time a student spends hanging out with friends on a scale of 1 to 4

- **Level 2**

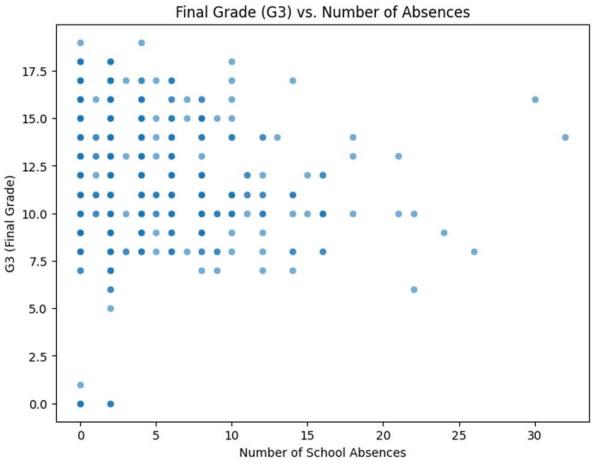
  Checked with the empty values in the dataset.

  From the analysis we can see that there are some null values in the columns famsize, Fedu, traveltime, higher, freetime, absences, Feature_1, Feature_2, Feature_3

  Columns like famsize, Fedu, traveltime, higher, freetime, Feature_2, Feature_3 are categorical values that we'll fill with the mode values

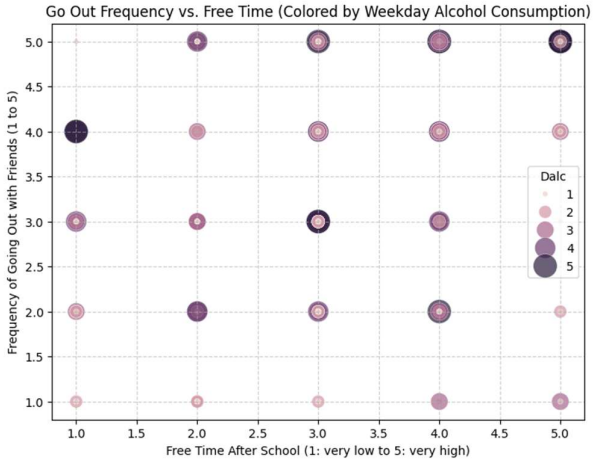
  Then we will check skew values of the remaining numerical columns if the |skew|>1, we replace the null values with median else, we replace with the mean.

- **Level 3**
    1. How does Number of Absences affect Final Grade (G3)



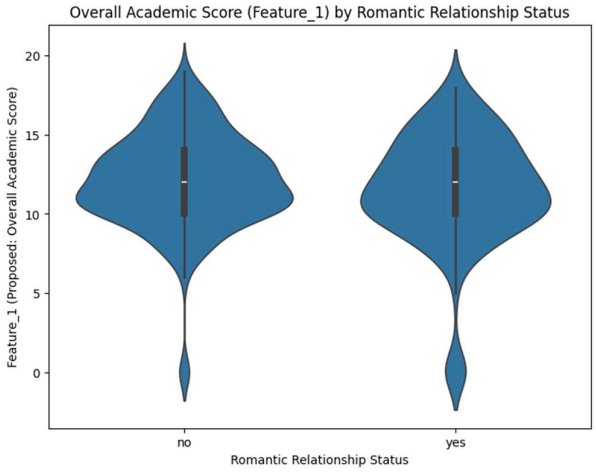
This scatter plot explores the relationship between the number of school absences and the final grade (G3). We can observe if there's a negative correlation (more absences, lower grades), indicating the impact of attendance on academic performance.

    2. How does 'freetime' vary with 'goout' frequency?



This scatter plot visualizes the relationship between free time and going out frequency, with points colored and sized by weekday alcohol consumption. It can reveal if more free time leads to more going out, and how alcohol consumption might be related to these social patterns.

    3. Overall Academic Score vs Romantic Relationship Status



This violin plot shows the distribution of the proposed 'Overall Academic Score' (G3) for students in and not in romantic relationships. It can reveal if there are noticeable differences in academic performance between these two groups.

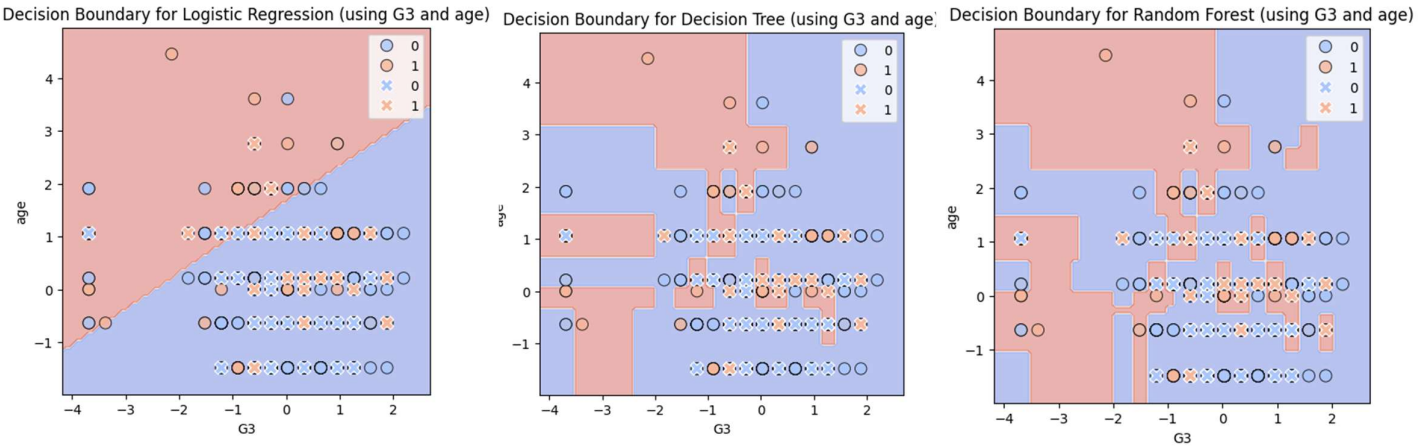
- **Level 4**

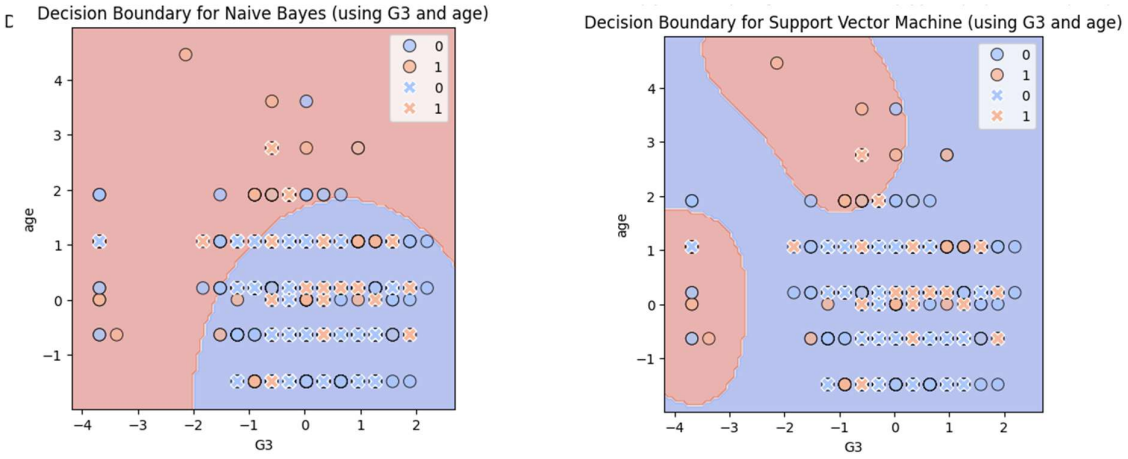
Model	Accuracy	Precision	Recall	F1-Score	ROC AUC	Confusion Matrix
Logistic Regression	0.569231	0.300000	0.125000	0.176471	0.581301	[[34, 7], [21, 3]]
Decision Tree	0.676923	0.560000	0.583333	0.571429	0.657520	[[30, 11], [10, 14]]
Random Forest	0.600000	0.375000	0.125000	0.187500	0.537602	[[36, 5], [21, 3]]
Gradient Boosting	0.553846	0.272727	0.125000	0.171429	0.485772	[[33, 8], [21, 3]]
Support Vector Machine	0.600000	0.333333	0.083333	0.133333	0.601626	[[37, 4], [22, 2]]
K-Nearest Neighbors	0.553846	0.380952	0.333333	0.355556	0.538618	[[28, 13], [16, 8]]
Naive Bayes	0.507692	0.250000	0.166667	0.200000	0.444106	[[29, 12], [20, 4]]

The Accuracy and Precisions for different classification models are listed above
Here we can see that Decision Tree Algorithm has the maximum accuracy among the others and hence can be considered as the best model for this classification

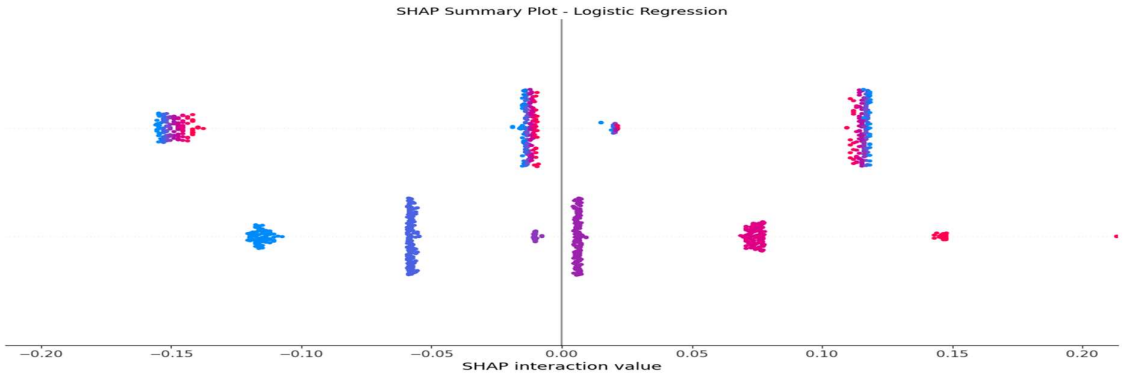
- **Level 5**

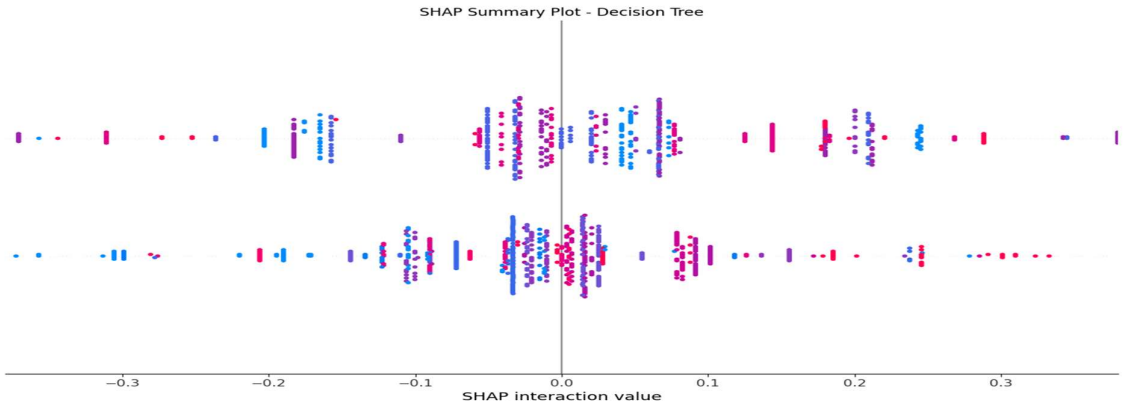
Choosing G3 and age as features, decision boundaries for different models have been plotted and shown below (1 represents 'in a relationship' while 0 represents 'not in a relationship':

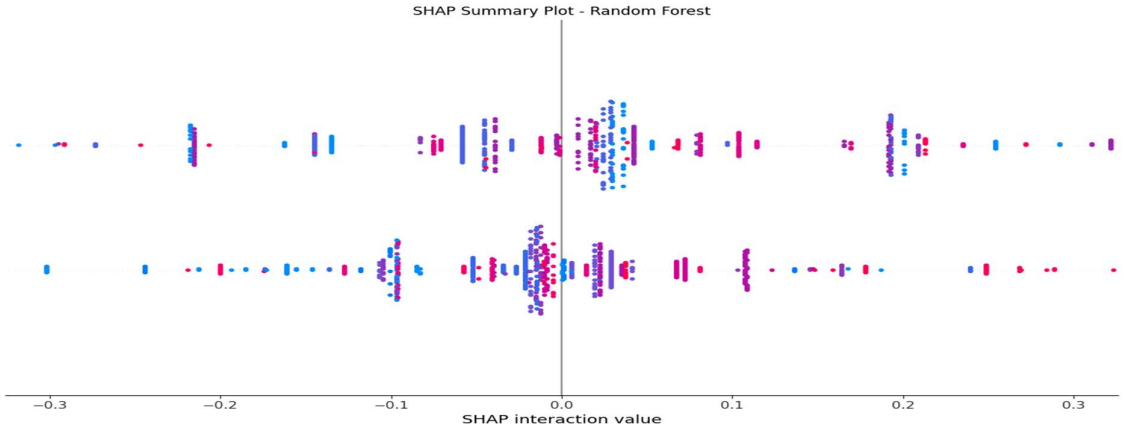


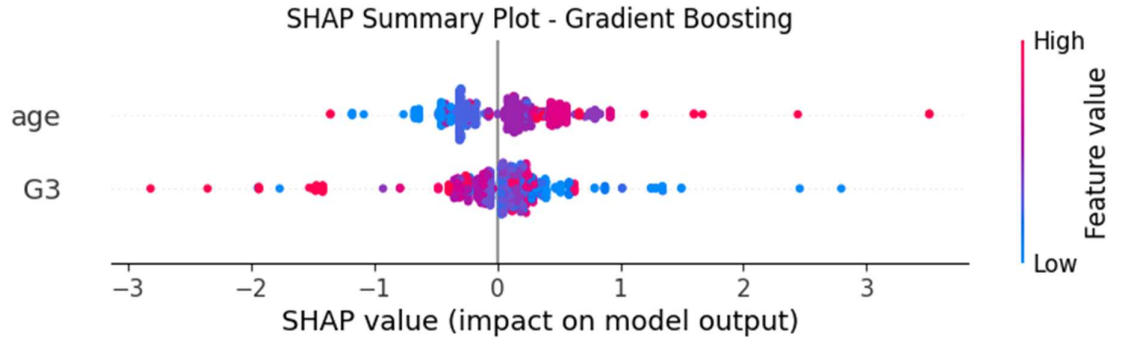


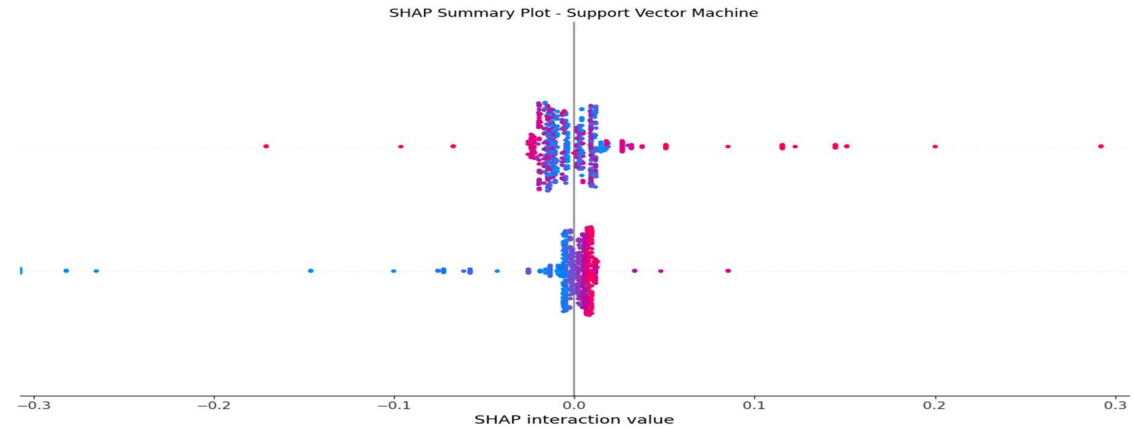
SHAP Plots:

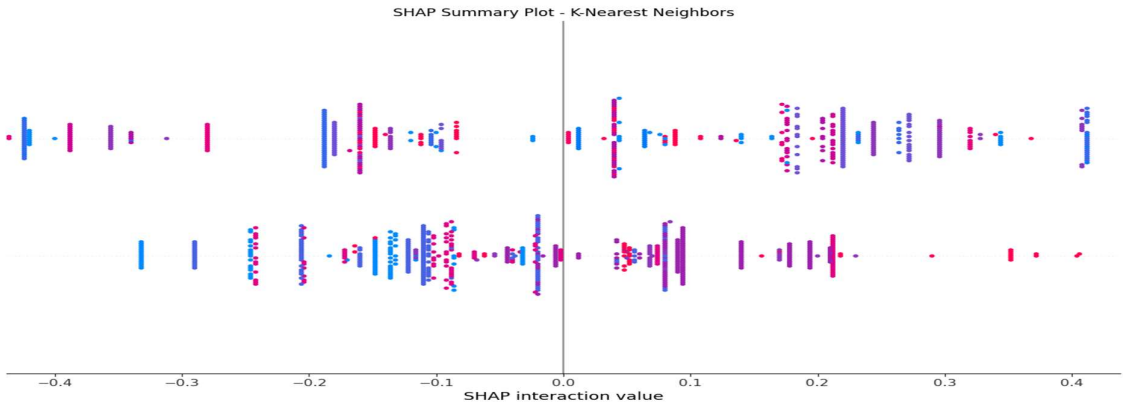

SHAP Summary Plot - Logistic Regression
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
SHAP interaction value

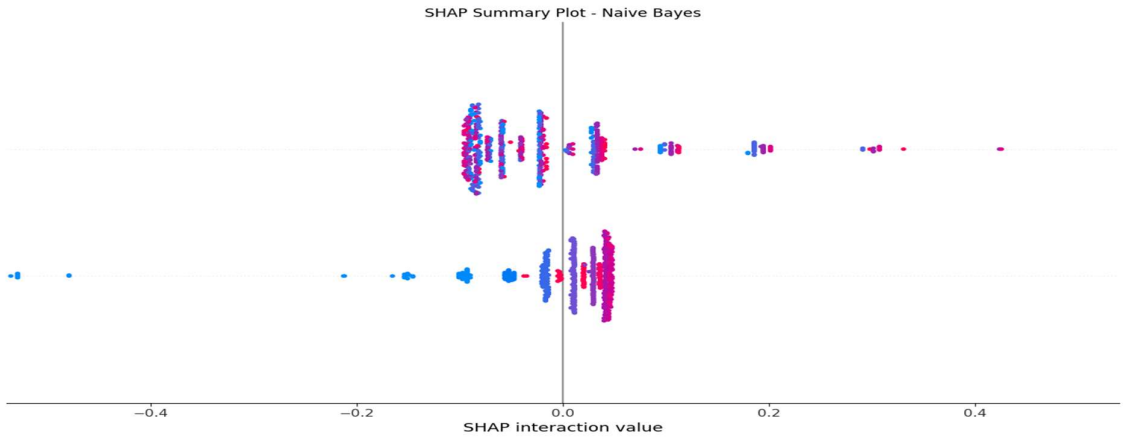

SHAP Summary Plot - Decision Tree
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP interaction value


SHAP Summary Plot - Random Forest
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP interaction value


SHAP Summary Plot - Gradient Boosting
age
G3
−3
−2
−1
0
1
2
3
SHAP value (impact on model output)
High
Feature value
Low

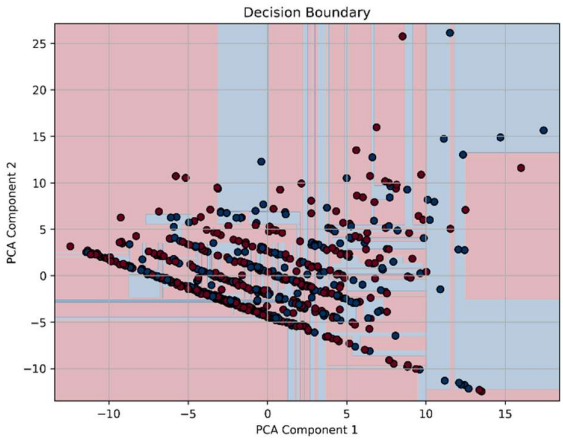

SHAP Summary Plot - Support Vector Machine
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP interaction value


SHAP Summary Plot - K-Nearest Neighbors
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
SHAP interaction value

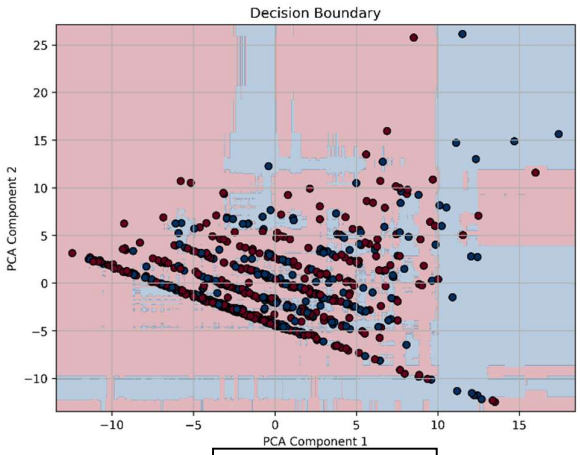

SHAP Summary Plot - Naive Bayes
−0.4
−0.2
0.0
0.2
0.4
SHAP interaction value

- **Bonus Task**

  In plot 1 and 2, we can see axis-parallel decision boundary lines that reflect that it is either Decision Tree or Random Forest. Random forest is made up of many decision trees that means it will have more number of lines, or more smoothness. Hence plot 2 will be a random forest and therefore, plot 1 will be a decision tree.

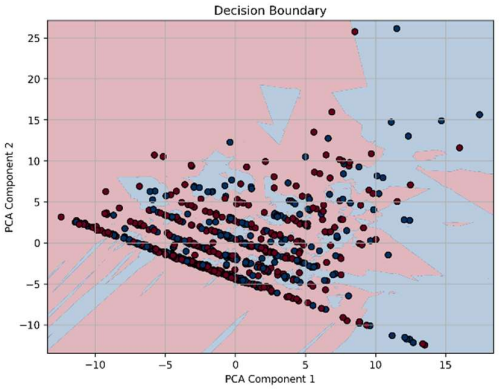


Decision Tree



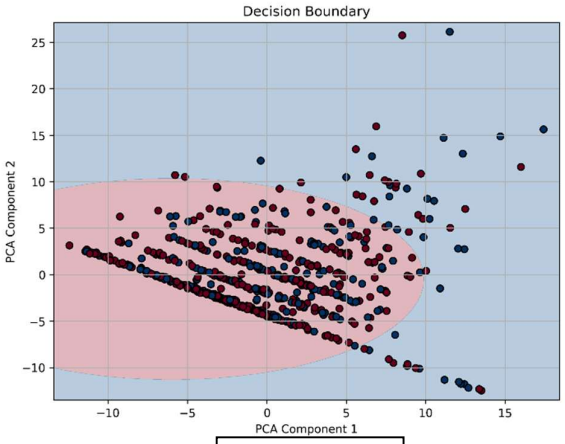
Random Forest

Plot 5 will be K Nearest Neighbour as it is very jaggy and contains many “patches”

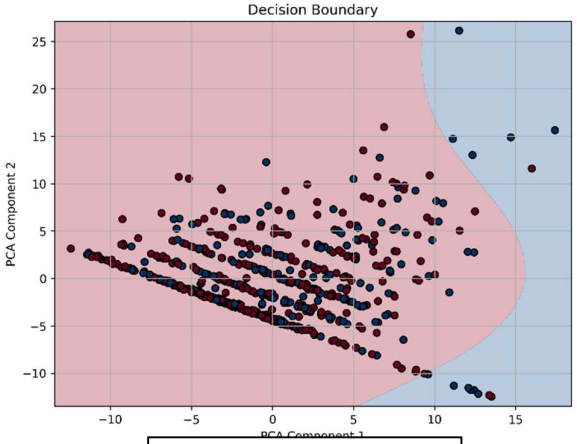


K-Nearest Neighbours

In plot 3 and 4, we can see a smooth curve with some complex functions. It can be SVM or Naïve Bayes Algorithm. Moreover, Naïve Bayes form elliptical or parabolic curves. Therefore, plot 4 will be Naïve Bayes and plot 3 will be SVM



Naïve Bayes



Support Vector Machine