Classifications for Business Graduate Admissions

Project 7

Aqdas Juya, Kiranjot Bhatia, Qwyn Petersen, Truong An Truong Lam

Outline

1. Introduction
2. Exploratory Data Analysis
3. Methodologies
    a. Support Vector Classifier
    b. Support Vector Machine with Polynomial Kernel
    c. Support Vector Machine with Radial Kernel
4. Conclusion

Introduction

Introduction

Objective: Performed classifications on business graduate admissions to help decide which applicants should be admitted to the school's graduate programs

Variables

- GPA(X1): Undergraduate grade point average
- GMAT(X2): Graduate management admission test scores
- De(group type): Response variable index for admission rulings
    - Admit - 1
    - Do not admit - 2
    - Borderline - 3

Exploratory Data Analysis

Summary Statistics

Table 1: Summary statistics of GPA and GMAT for Admit, Do not Admit, Borderline

Group(De)	Type	Min	Q1	Median	Q3	Max	Mean/SD	N	Missing
Admit(1)	GPA	2.96	3.265	3.385	3.477	3.78	3.375/ 0.193	26	0
	GMAT	431	524	558	594.75	693	561.38/ 68.869	26	0
Do not Admit(2)	GPA	2.13	2.355	2.43	2.525	2.68	2.43/ 0.144	23	0
	GMAT	321	411.5	458	504.5	542	453.56/ 61.818	23	0
Borderline (3)	GPA	2.73	2.86	3	3.12	3.5	2.99/ 0.189	21	0
	GMAT	313	419	446	485	546	447.95/ 50.716	21	0

Exploratory Data Analysis (EDA)

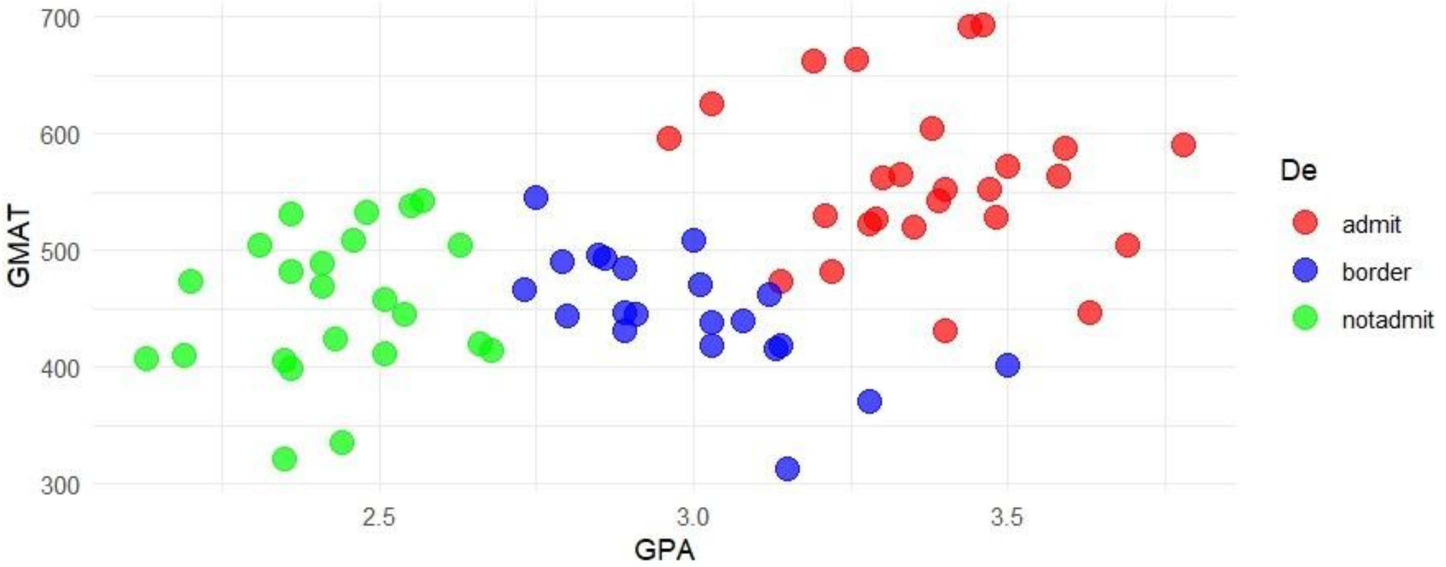


Figure 1. Relationship between GPA, GMAT scores, and the distribution of the three groups.

Exploratory Data Analysis (EDA) - Density Plots

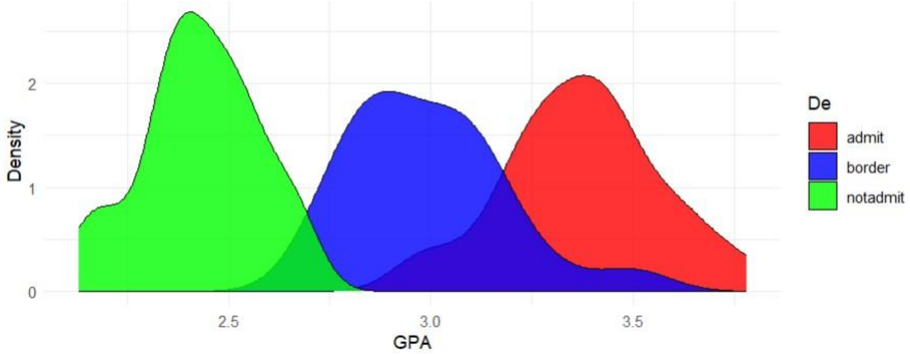


Figure 1.1 GPA values for three different groups

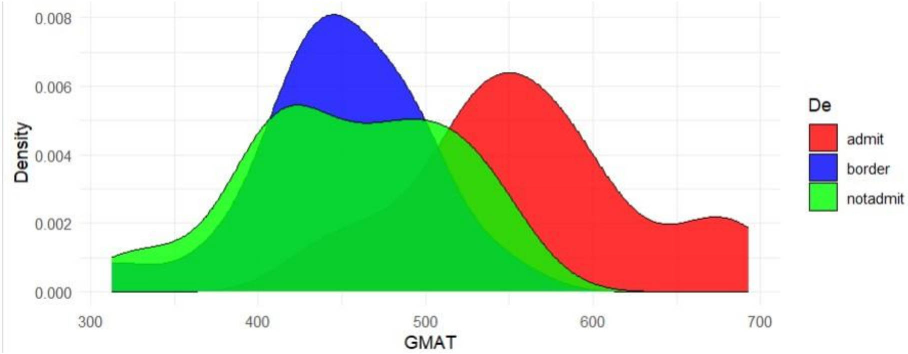


Figure 1.2 GMAT scores for three different groups

Methodologies

Data Analysis for Train Support Vector Classifier (Linear Kernel) with 10-fold CV

- Effective when data is **linearly separable**.
- Prevents overfitting
- **10-Fold Cross-Validation** helps in selecting the best hyperparameters.

Model training with SVC with 10-fold CV

SVC Confusion Matrix:

	Admit	Border	Not Admit
Admit	4	0	0
Border	1	5	2
Not Admit	0	0	3

Accuracy = 80%

- **Confusion Matrix Summary:**
  - *Admit*: 4 correct predictions, 0 misclassifications.
  - *Border*: 1 misclassification to Admit, 5 correct, 2 misclassifications to Not Admit.
  - *Not Admit*: 3 correct predictions, no misclassifications.

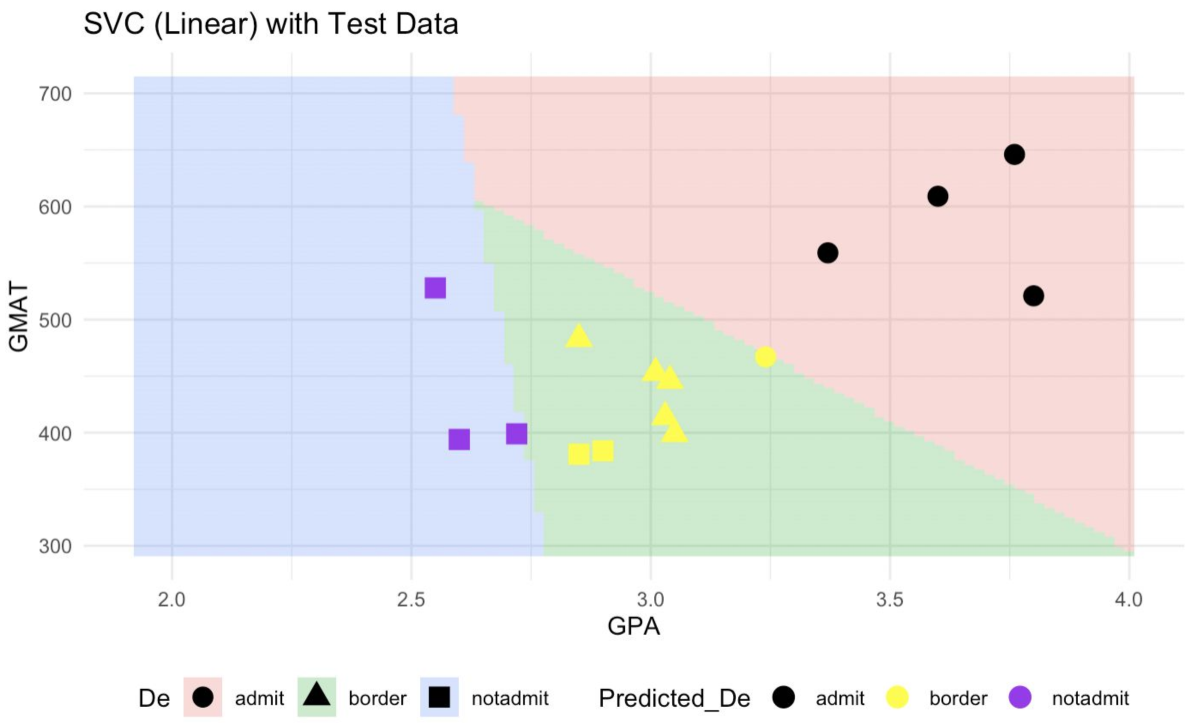


Figure 1.3 Relationship between GPA, GMAT scores using the SVC (linear) with the Testing Data.

Data Analysis for Support Vector Machines with Polynomial Kernel (degree 2)

- When data isn't **linearly separable** → map into a higher dimensional space.

- We can separate the data using a **hyperplane**.
- **SVM** can only handle two classes → OvO or OvA strategies allow for multiple classes.
- **Kernels** enable decision boundaries without expensive transformations.
- **Polynomial kernel**: $$K(x_1, x_2) = (x_1^T x_2 + c)^d$$

Implementation: A Step-by-Step Approach

1. With **Tune** (library e1071) → supply the equation, kernel, the degree, cost parameter value(s) and run a 10-fold CV on each **C** value using the **training set**.
2. Using the optimal **C** (4) → run the best model to get predictions.
3. Compare predicted & actual values to get the **accuracy rate** & **confusion matrix**.

Accuracy rate = 46.67%

	Admit	Border	Not
Sensitivity	40%	100%	0%
Specificity	50%	90%	80%

Table 3: Statistics by class for SVM polynomial kernel (degree 2)

Actual

Predicted

	Admit	Border	Not
Admit	2	0	5
Border	1	5	0
Not	2	0	0

Table 4: Confusion matrix for SVM polynomial kernel (degree 2)

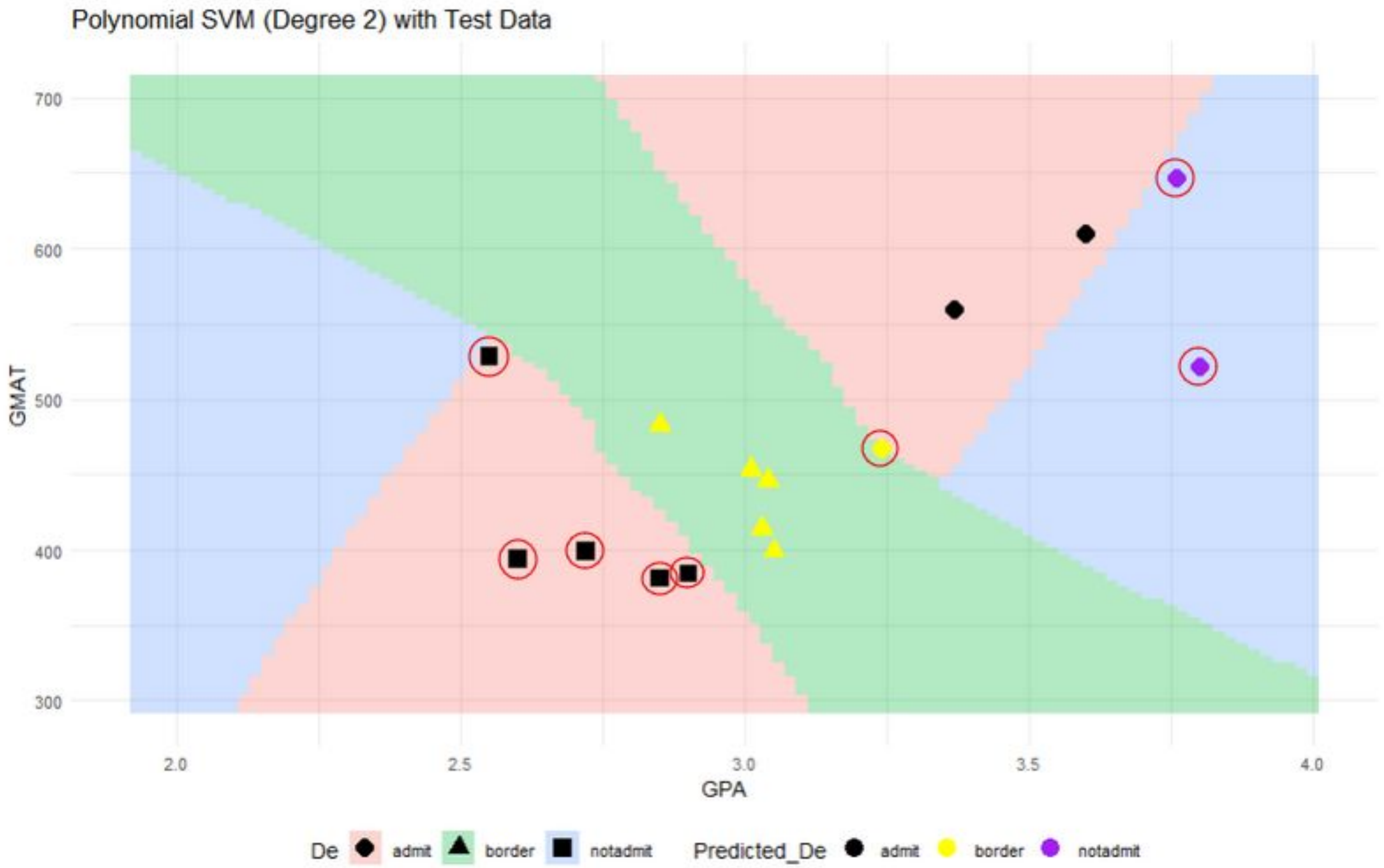


Figure 1.4: Relationship between GPA, GMAT scores using the Polynomial (degree 2) for SVM with the testing data.

Data Analysis for Train Support Vector Machines with Radial Kernel (optimizing both gamma and cost)

- RBF Kernel: Non-linear kernel, maps data to higher-dimensional space for complex boundaries.
- RBF Kernel Formula: $K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$
- Key Parameters:
  - C (Cost): Balances margin size and misclassification.
  - γ: Controls kernel spread and decision boundary flexibility.

Model Training with RBF Kernel

Table 5: Confusion matrix for RBF kernel for SVM

	Admit	Border	Not Admit
Admit	4	0	0
Border	1	5	2
Not Admit	0	0	3

Accuracy = 80%

- **Used RBF kernel for SVM.**
  - Tuned C (from 0.1 to 1000) and gamma (from 0.1 to 4).
  - Performed 10-fold cross-validation to find the best parameters.
  - Best cost: 1
  - Best gamma: 0.1
- **Confusion Matrix Summary:**
  - *Admit*: 4 correct predictions, 0 misclassifications.
  - *Border*: 1 misclassification to Admit, 5 correct, 2 misclassifications to Not Admit.
  - *Not Admit*: 3 correct predictions, no misclassifications.

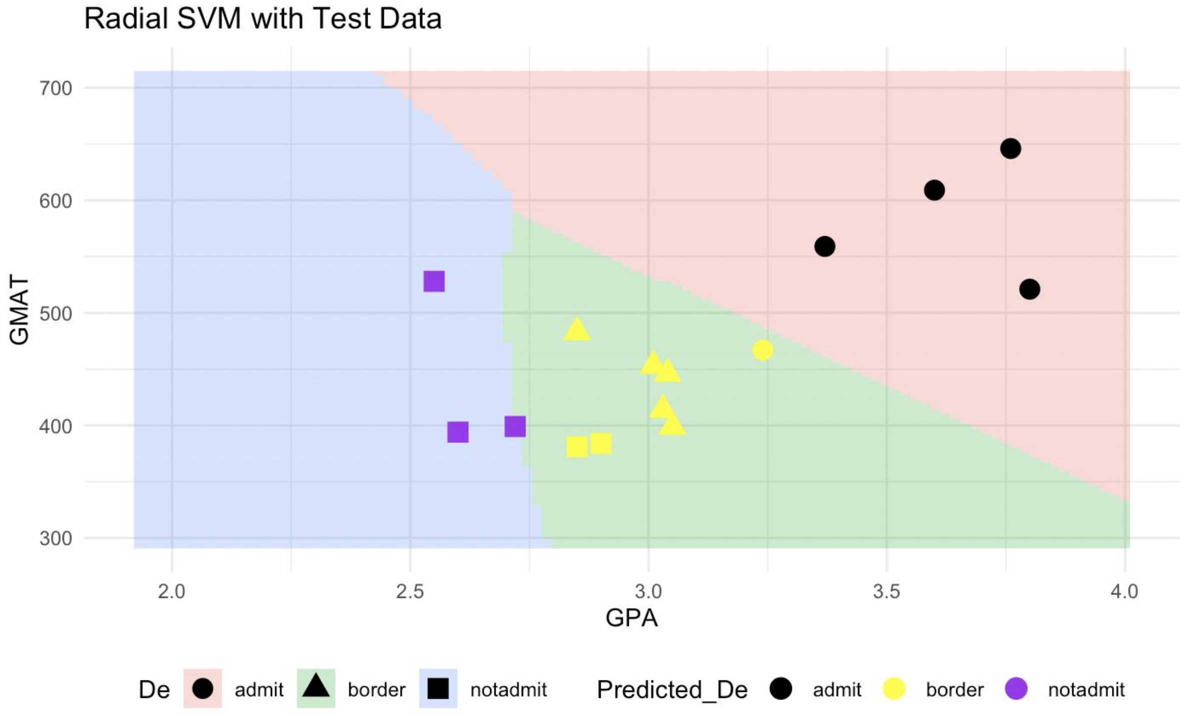


Figure 1.5 Relationship between GPA, GMAT scores using the RBF kernel for SVM with the testing data.

Conclusion :

Comparison of SVM Classifiers: Confusion Matrices and Accuracy

- Linear SVM and RBF SVM both had 80% accuracy, meaning the data is mostly linear.
- Polynomial SVM had 46.67% accuracy, so it didn't fit the data well

SVC:

	Admit	Border	Not Admit
Admit	4	0	0
Border	1	5	2
Not Admit	0	0	3

Accuracy = 80.00%

Polynomial (d = 2) SVM:

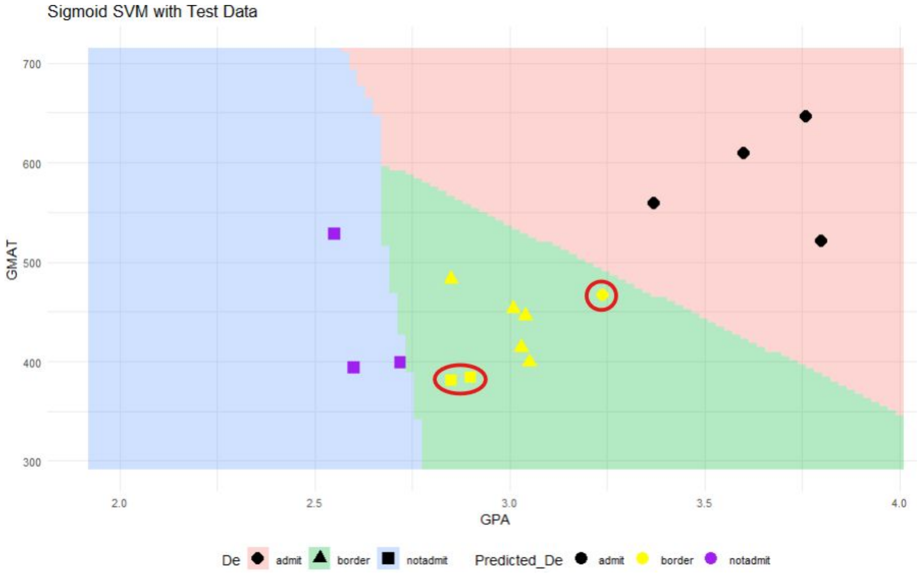
	Admit	Border	Not Admit
Admit	2	0	5
Border	1	5	0
Not Admit	2	0	0

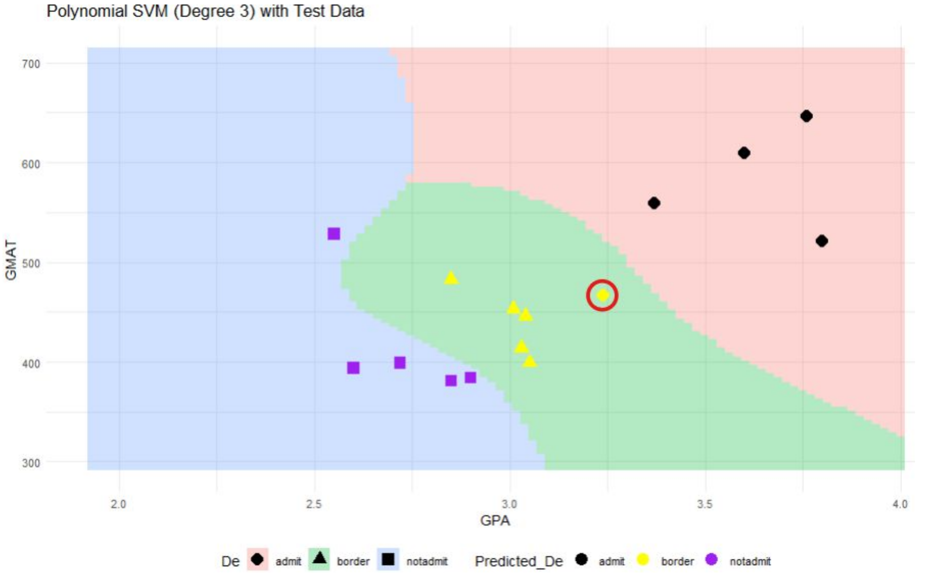
Accuracy = 46.67%

RBF SVM:

	Admit	Border	Not Admit
Admit	4	0	0
Border	1	5	2
Not Admit	0	0	3

Accuracy = 80.00%





Confusion matrices, accuracy rates and decision boundaries for other kernels

Sigmoid:

	Admit	Border	Not
Admit	4	0	0
Border	1	5	2
Not	0	0	3

Accuracy = 80.00%

Polynomial (d = 3):

	Admit	Border	Not
Admit	4	0	0
Border	1	5	0
Not	0	0	5

Accuracy = 93.33%

Questions?