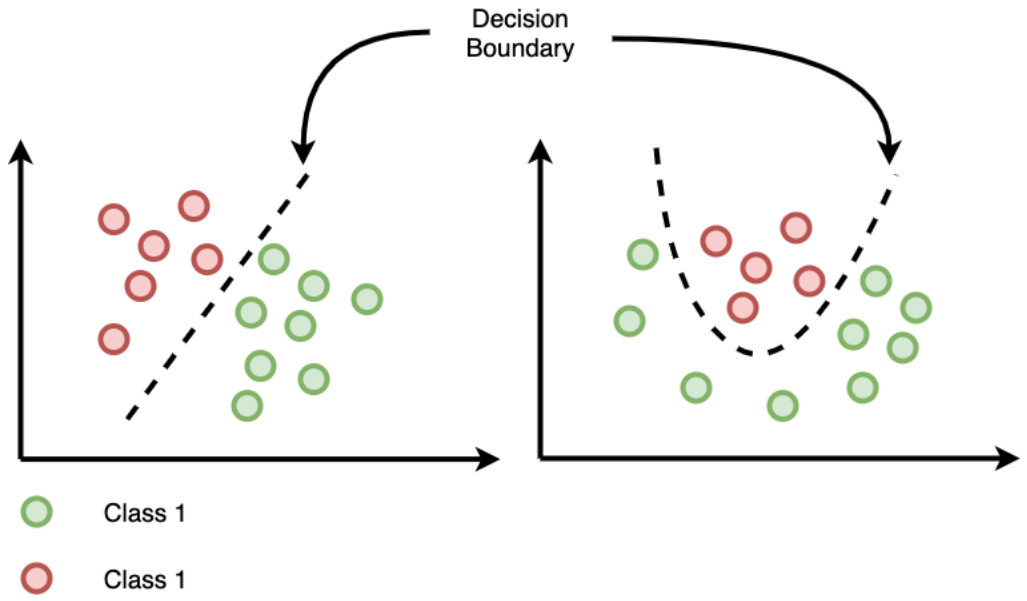
Decision Boundaries

20CP401T

Himanshu K. Gajera
Department of Computer Science & Engineering
Pandit Deendayal Energy University, Gandhinagar

What is Decision Boundary

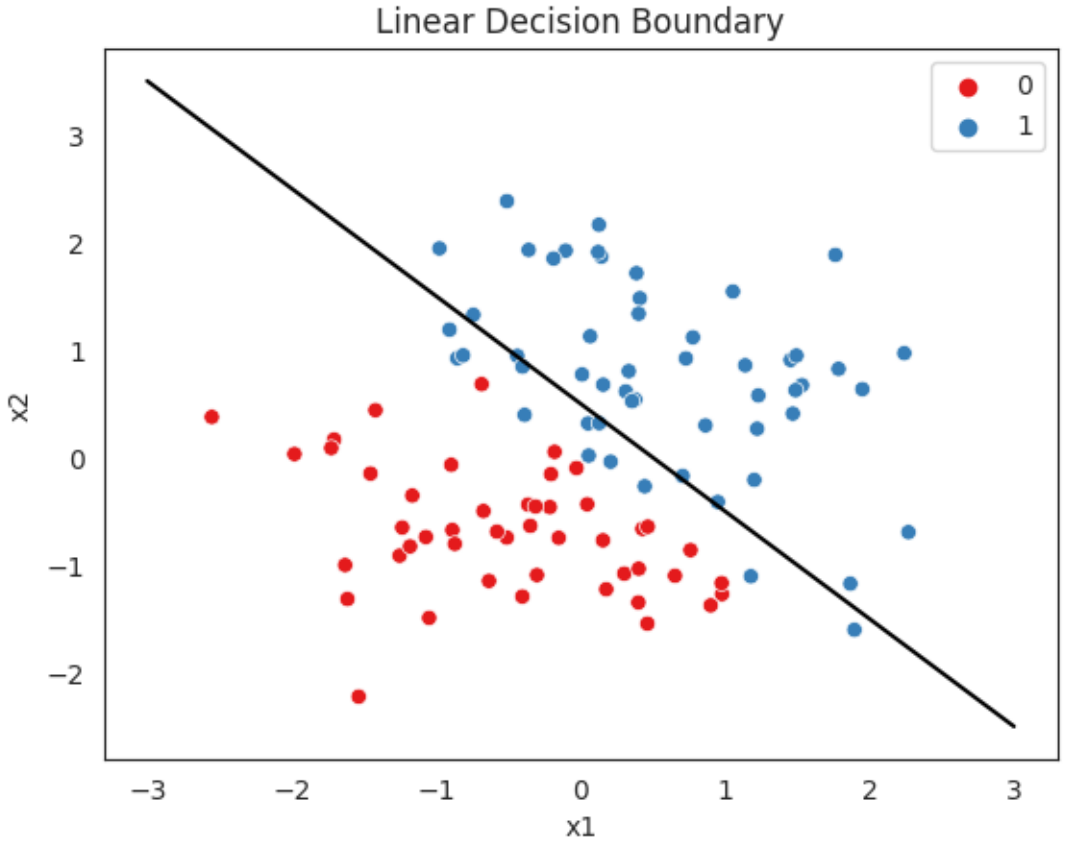
- While training a classifier on a dataset, using a specific classification algorithm, it is required to define a set of hyper-planes, called Decision Boundary.
- That separates the data points into specific classes, where the algorithm switches from one class to another.



Types of Decision boundaries

Linear decision boundary:

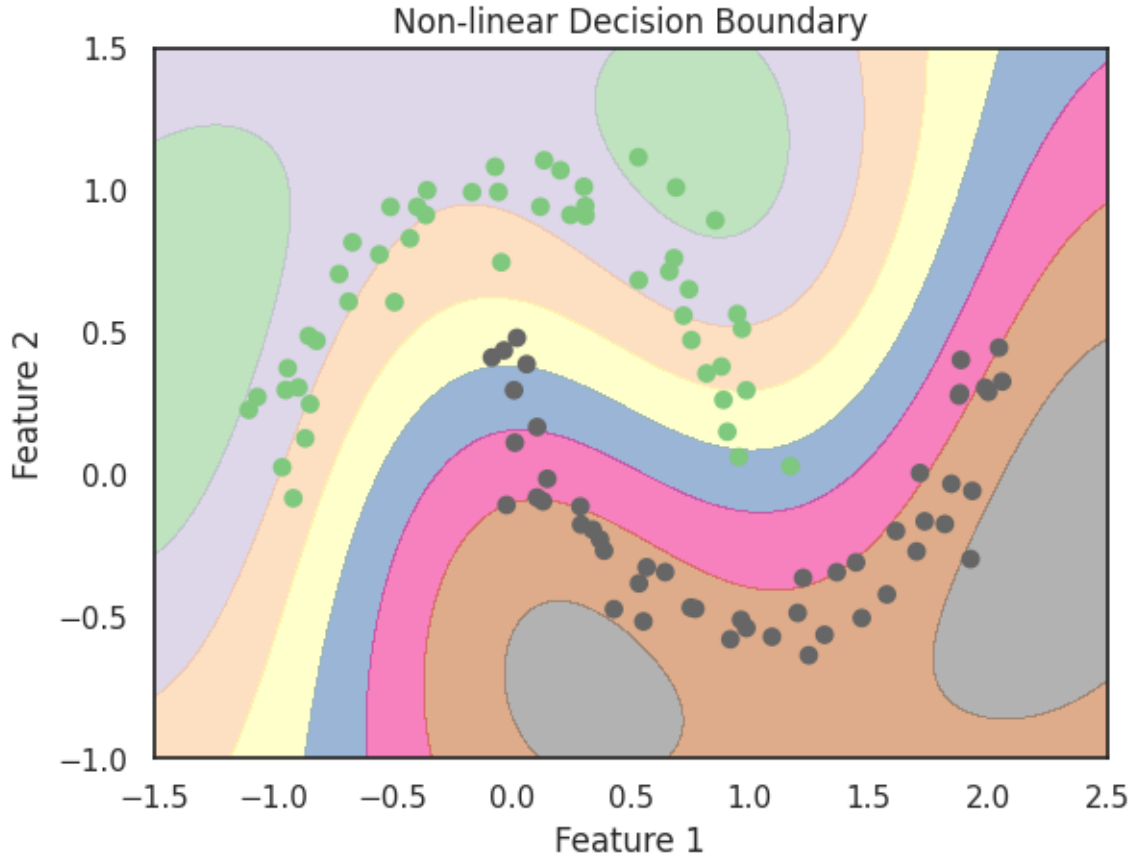
- A linear decision boundary is a straight line that separates the data into two classes.
- It is the simplest form of decision boundary and is used when problem is linearly separable.
- Expressed as a linear equation, y = mx + b, where m is the slope of the line and b is the y-intercept.



Types of Decision boundaries

Non Linear decision boundary:

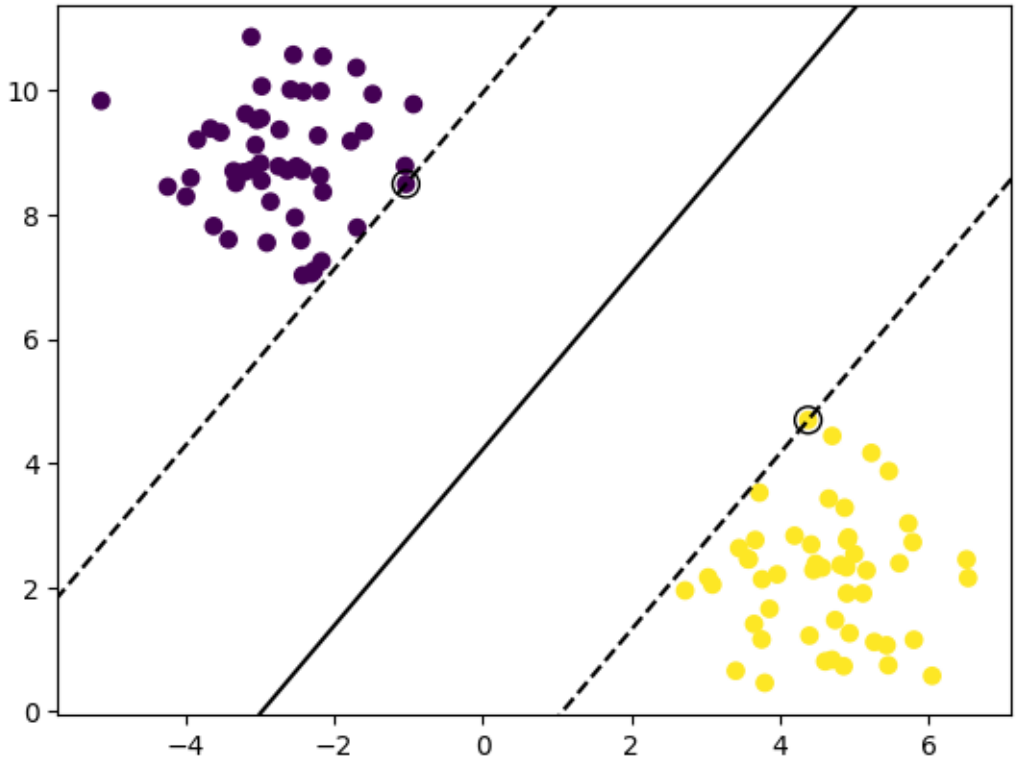
- A non-linear decision boundary is a curved line that separates the data into two or more classes.
- Used when the classification problem is not linearly separable.
- Take different forms such as parabolas, circles, ellipses, etc.



Types of Decision boundaries

Decision Boundary with Margin:

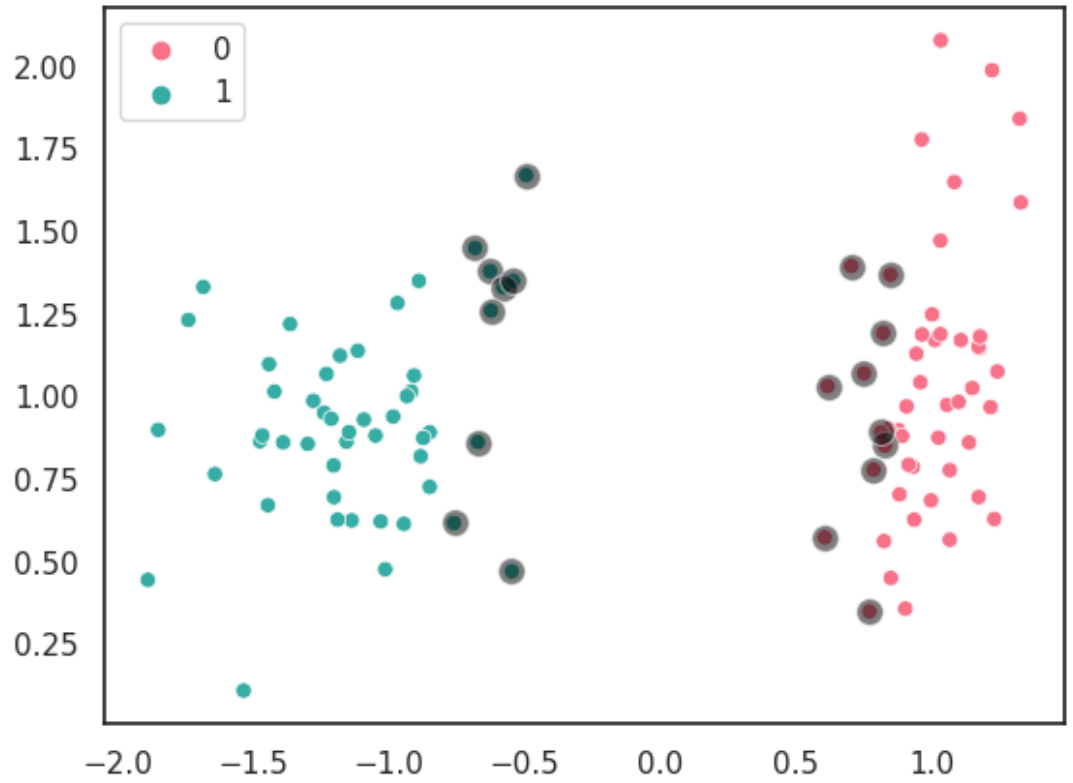
- A decision boundary with margin is a line or curve that separates the data into two classes while maximizing the distance between the boundary and the closest data points.
- Margin is defined as the distance between the decision boundary and the closest data points of each class.
- The objective is to improve the generalization performance of classifier by reducing the risk of overfitting.



Types of Decision boundaries

Decision Boundary with Soft Margin:

- A decision boundary with soft margin is a line or curve that separates the data into two classes while allowing some misclassifications.
- Soft margin is used when the data is not linearly separable and has some noise or outliers.
- The objective is to find a balance between the accuracy and its ability to generalize to unseen data.



How Decision Boundaries are Generated

There are different methods to generate decision boundaries depending on the type of classification problem and the algorithm used. The most common methods are:

1) Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation is a statistical method used to estimate the parameters of a probability distribution based on the observed data.
- MLE is used to estimate the parameters of a model that generates the decision boundary that best fits the observed data.

2) Support Vector Machine (SVM)

- Support vector machine is a supervised learning algorithm that is used to find the best decision boundary that separates the data into two or more classes.
- SVM uses a technique called kernel trick to transform the data into a higher-dimensional space where a linear decision boundary can be found.

How Decision Boundaries are Generated

3) Decision Trees

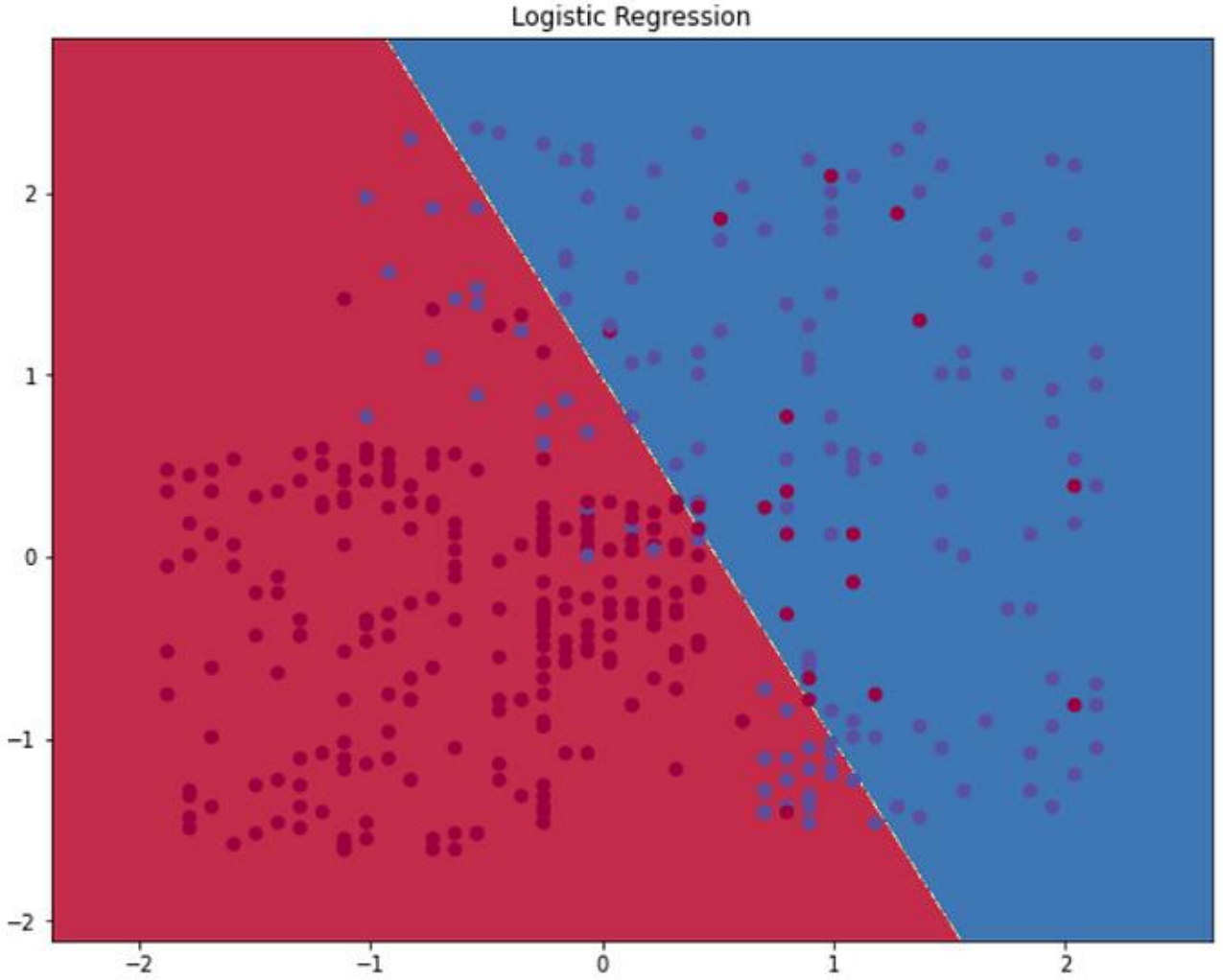
- Decision trees are a type of supervised learning algorithm that is used to generate a decision boundary by recursively partitioning the feature space into smaller subspaces.
- The decision boundary is represented by a tree-like structure where each node represents a decision based on a feature value and each leaf represents a class label.

4) Neural Networks

- Neural networks are a type of supervised learning algorithm that is used to generate decision boundaries by learning a non-linear mapping between the input features and the output classes.
- Neural networks consist of multiple layers of interconnected nodes that perform non-linear transformations on the input data to generate the output classes.

Decision Boundary for Logistic Regression.

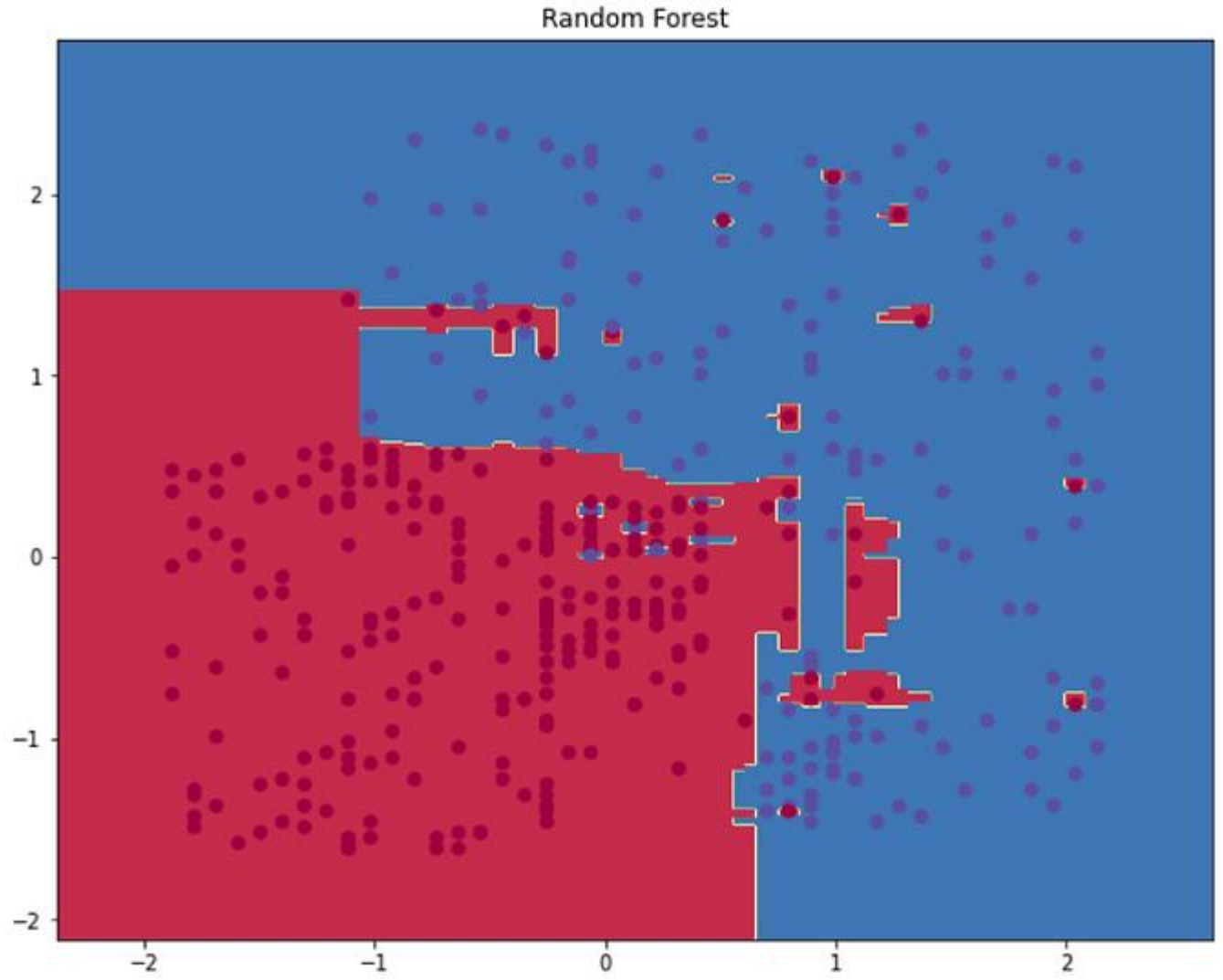
In Logistic Regression, Decision Boundary is a linear line



Dataset: https://www.kaggle.com/datasets/rakeshrau/social-network-ads

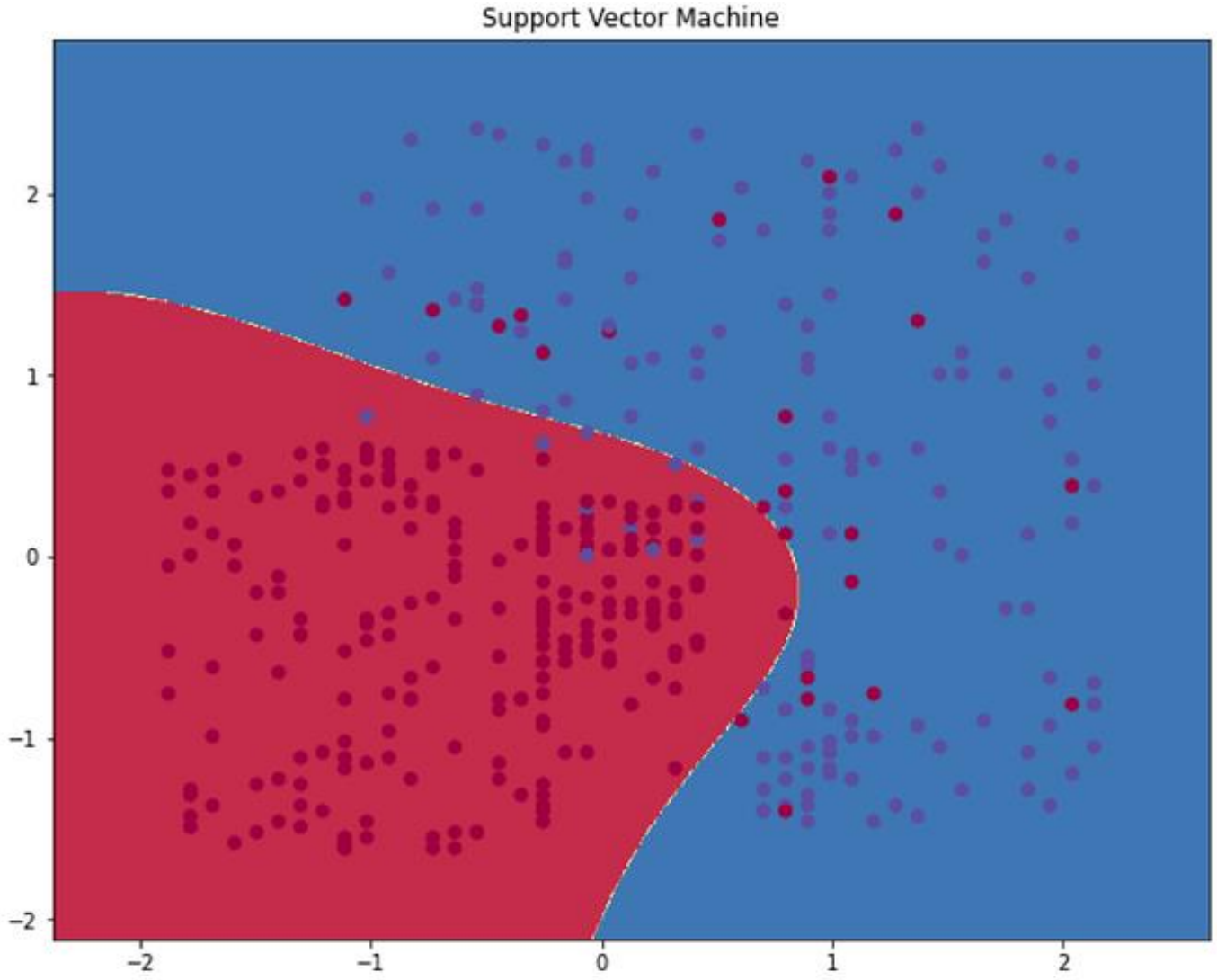
Decision Boundary for Random Forest.

The decision surfaces for the Random Forest is very complex.



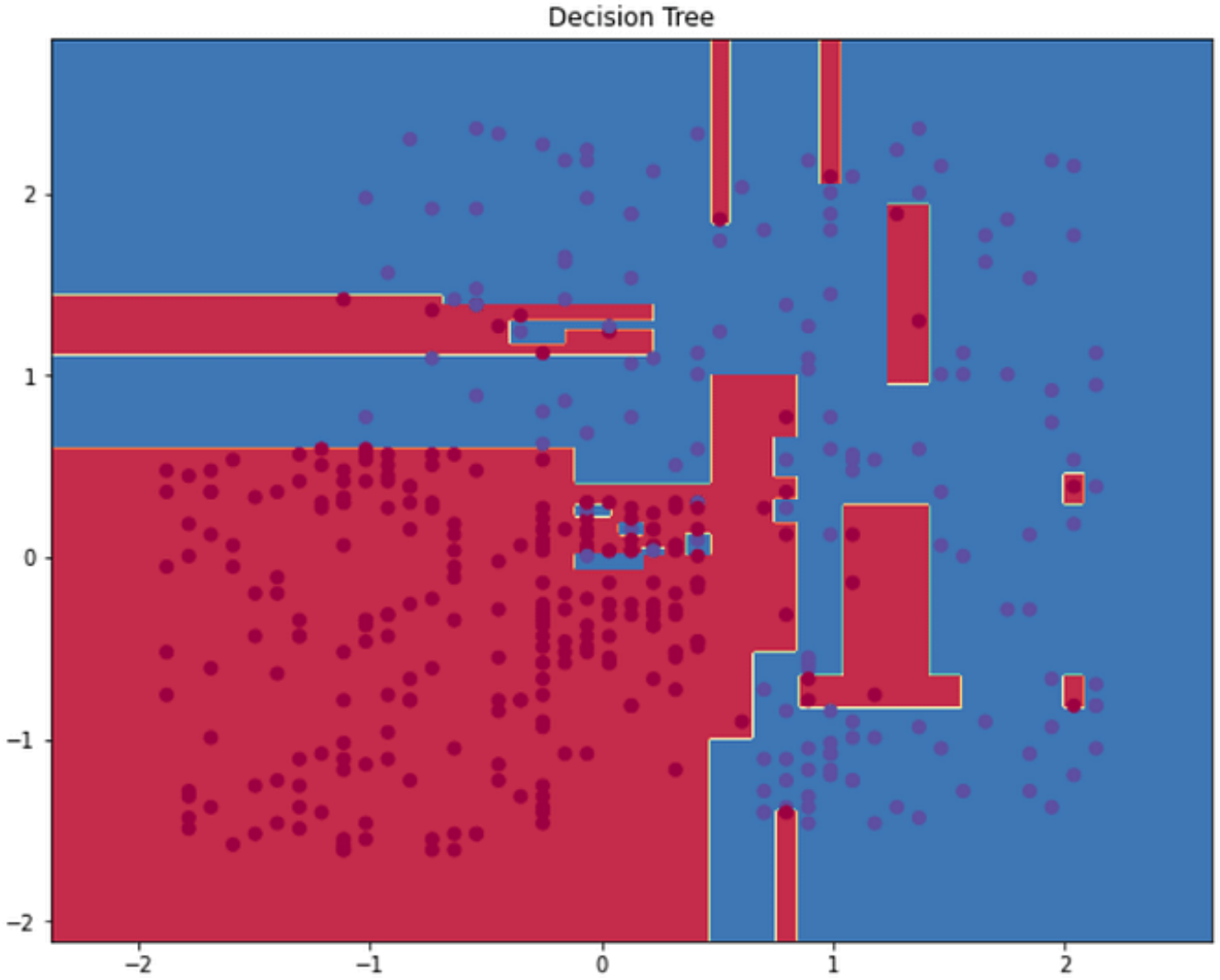
Decision Boundary for Support Vector Machine.

Support Vector Machine find a hyperplane that separates the feature space into two classes with the maximum margin.



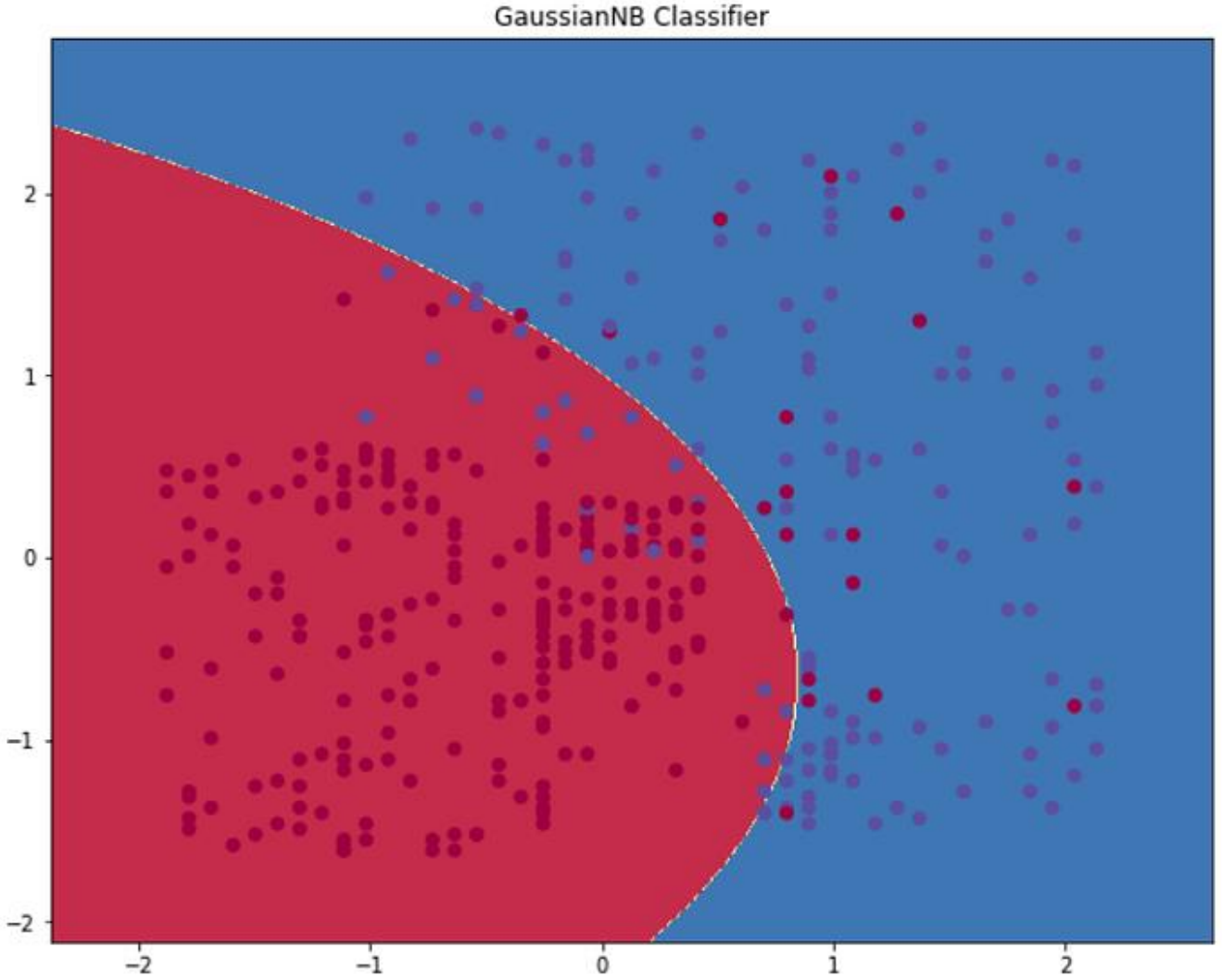
Decision Boundary for Decision Tree.

The decision surfaces for the Decision Tree is very complex.



Decision Boundary for Gaussian NaiveBayes.

Gaussian Naive Bayes has also performed well, having a smooth curve boundary line



Applications of Decision Boundaries

Decision boundaries are used in a wide range of applications in machine learning and pattern recognition. Some of the most common applications are:

Image Classification:

- Image classification is the process of assigning a label to an image based on its features. Decision boundaries are used to separate the different classes of images such as cats, dogs, flowers, etc. The decision boundary is learned from a training dataset of labeled images and is used to classify new unseen images.

Speech Recognition:

- Speech recognition is the process of transcribing spoken words into written text. Decision boundaries are used to identify the different phonemes and words in the speech signal. The decision boundary is learned from a training dataset of labeled speech signals and is used to recognize new unseen speech signals.

Fraud Detection

- Fraud detection is the process of identifying fraudulent transactions in a financial system. Decision boundaries are used to distinguish between legitimate and fraudulent transactions based on their features such as amount, location, and time of the transaction. The decision boundary is learned from a training dataset of labeled transactions and is used to detect new unseen fraudulent transactions.

Medical Diagnosis

- Medical diagnosis is the process of identifying the presence or absence of a disease based on the symptoms and medical tests. Decision boundaries are used to distinguish between different diseases based on their symptoms and test results. The decision boundary is learned from a training dataset of labeled medical cases and is used to diagnose new unseen medical cases.

Challenges of Decision Boundaries

Although decision boundaries are a powerful tool in machine learning and pattern recognition, they face several challenges, including:

Overfitting

- Overfitting occurs when the decision boundary is too complex and fits the training data too well, but fails to generalize to new unseen data. Overfitting can be addressed by using regularization techniques or by using simpler models.

Underfitting

- Underfitting occurs when the decision boundary is too simple and fails to capture the underlying patterns in the data. Underfitting can be addressed by using more complex models or by adding more features to the dataset.

Class Imbalance

- Class imbalance occurs when the dataset has an unequal distribution of samples across the different classes. Class imbalance can lead to biased decision boundaries that favor the majority class. Class imbalance can be addressed by using sampling techniques or by using different performance metrics.

Conclusion

- Decision boundary is a fundamental concept in machine learning and pattern recognition.
- It is used to separate the data into different classes based on their features.
- There are different types of decision boundaries, including linear, non-linear, decision boundary with margin, and decision boundary with soft margin.
- Decision Boundary in determining a classifier model, built several classifier models and plotted their respective decision boundaries to select the best model and also knew that plotting a Decision Boundary for higher dimensional data is a complex task.
- Can be plotted, using tSNE, where the dimensions of the data can be reduced in several steps.