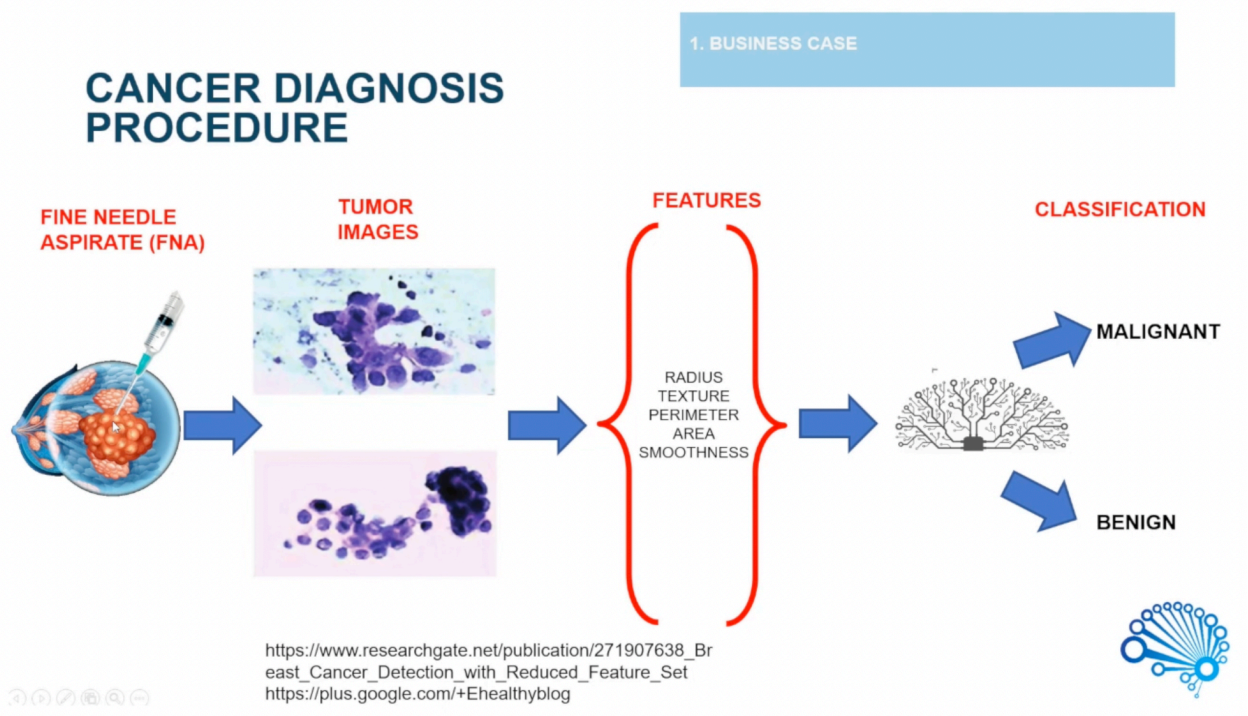
BREAST CANCER CLASSIFICATION

Utsav Acharya

BUSINESS CASE:



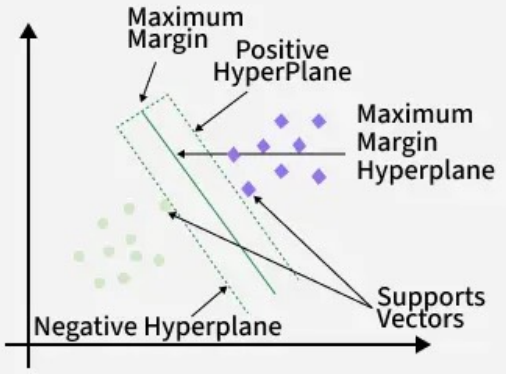
SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It tries to find the best boundary known as hyperplane that separates different classes in the data. It is useful when you want to do binary classification like spam vs. not spam or cat vs. dog.

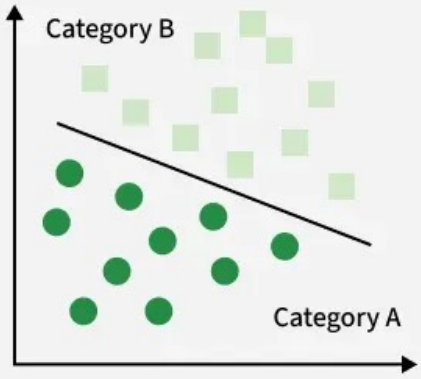
The main goal of SVM is to maximize the margin between the two classes. The larger the margin the better the model performs on new and unseen data.

Support Vectors & Hyperplane

- Support Vectors are the closest data points to the hyperplane that define the class boundary.
- A hyperplane is a plane that separates different classes.

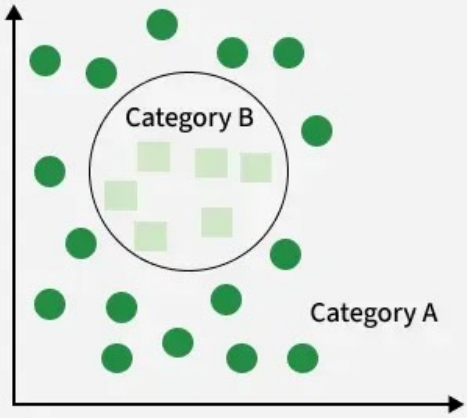


Linear SVM

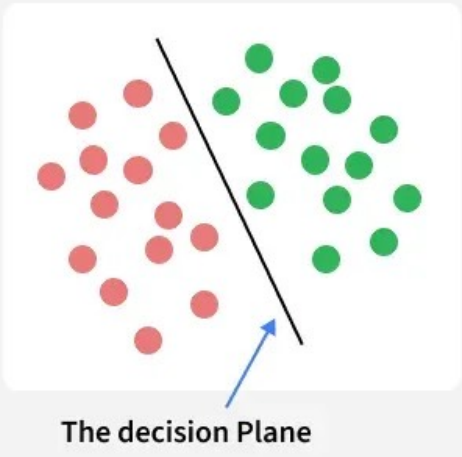


Vs

Non-Linear SVM



- Support Vector Machine is used for classification and regression.
- It works by finding decision planes that separate data into different classes.



CONFUSION MATRIX

Confusion matrix is a simple table used to measure how well a classification model is performing.
True Positive (TP): The model correctly predicted a positive outcome i.e. the actual outcome was positive
True Negative (TN): The model correctly predicted a negative outcome i.e. the actual outcome was negative.
False Positive (FP): The model incorrectly predicted a positive outcome i.e. the actual outcome was negative. It is also known as a **Type I error**.
False Negative (FN): The model incorrectly predicted a negative outcome i.e. the actual outcome was positive. It is also known as **Type II error**.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Metrics based on Confusion Matrix Data

1. Accuracy

Accuracy shows how many predictions the model got right out of all the predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2. Precision
It tells us how many of the “positive” predictions were actually correct.

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. Recall
Recall measures how good the model is at predicting positives.

$$\text{Recall} = \frac{TP}{TP+FN}$$

4. F1-Score
F1-score combines precision and recall into a single metric to balance their trade-off. It provides a better sense of a model’s overall performance particularly for imbalanced datasets.

$$\text{F1-Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

5. Specificity
It measures the ability of a model to correctly identify negative instances. Specificity is also known as the True Negative Rate.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

DATA NORMALIZATION:

Data normalization is a preprocessing method that resizes the range of feature values to a specific scale, usually between 0 and 1. It is a feature scaling technique used to transform data into a standard range. Normalization ensures that features with different scales or units contribute equally to the model and improves the performance of many machine learning algorithms.

1. Min-Max Normalization

Min-Max normalization rescales a feature to a specific range, typically [0, 1]:

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

- The minimum value maps to 0
- The maximum value maps to 1
- Other values are scaled proportionally

C parameter in SVM (The “strictness” Level)

The Regularization parameter. It tells the SVM how much it should care about avoiding misclassifying each training example.

Low C (Soft Margin)

Behavior: The “chill” parent. It allows some errors (misclassifications) in the training data to keep the decision boundary simple and the margin wide.
Result: A smoother, straighter line.
Risk: Underfitting (too simple)

High C (Hard Margin)

Behavior: The “strict” parent. It penalizes mistakes heavily. It tries hard to correctly classify every single data point.
Result: A complex wiggly line that hugs the datapoints.
Risk: Overfitting (memorizing the noise)

Gamma parameter in SVM (The “Reach”)

Used specifically in non-linear kernels (like RBF). It defines how far the influence of a single training example reaches.

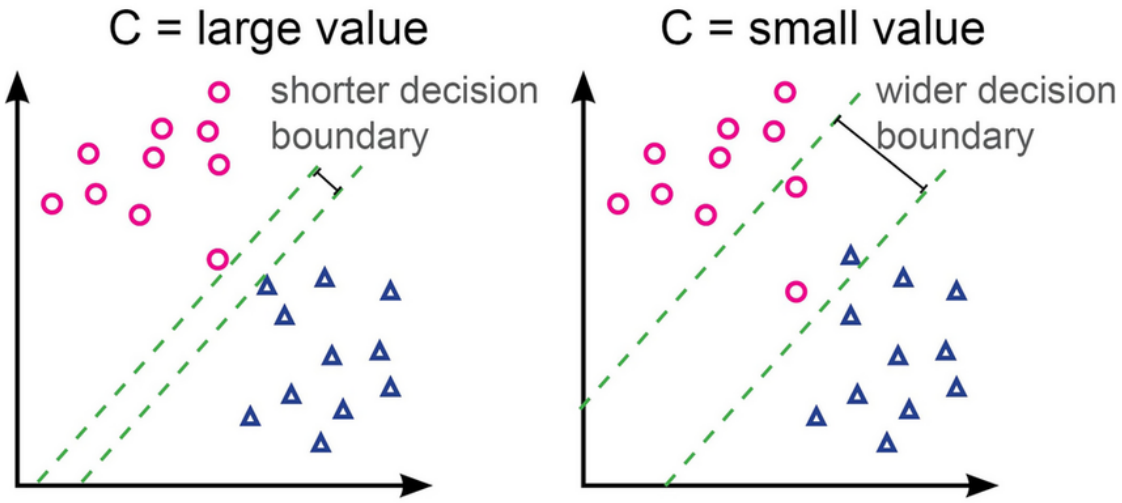
Low Gamma (γ)

Behavior: “Far” reach. Even points far away from the decision line influence where the line is drawn.
Result: The decision boundary changes slowly; it looks like a broad, smooth curve.
Risk: Underfitting (too smooth)

High Gamma (γ)

Behavior: “Close” reach. Only points very close to the decision line matter.
Result: The boundary creates tight “islands” or bubbles around individual data points.
Risk: Overfitting (captures noise)

C parameter



Gamma (γ) parameter