APPLICATION OF SVM

EBB3 Team 2: Vera, Anne, Felix

TABLE OF CONTENTS

01
INTRODUCTION

Recap SVM
Research Question
Dataset Used

02
LINEAR & RADIAL SVM

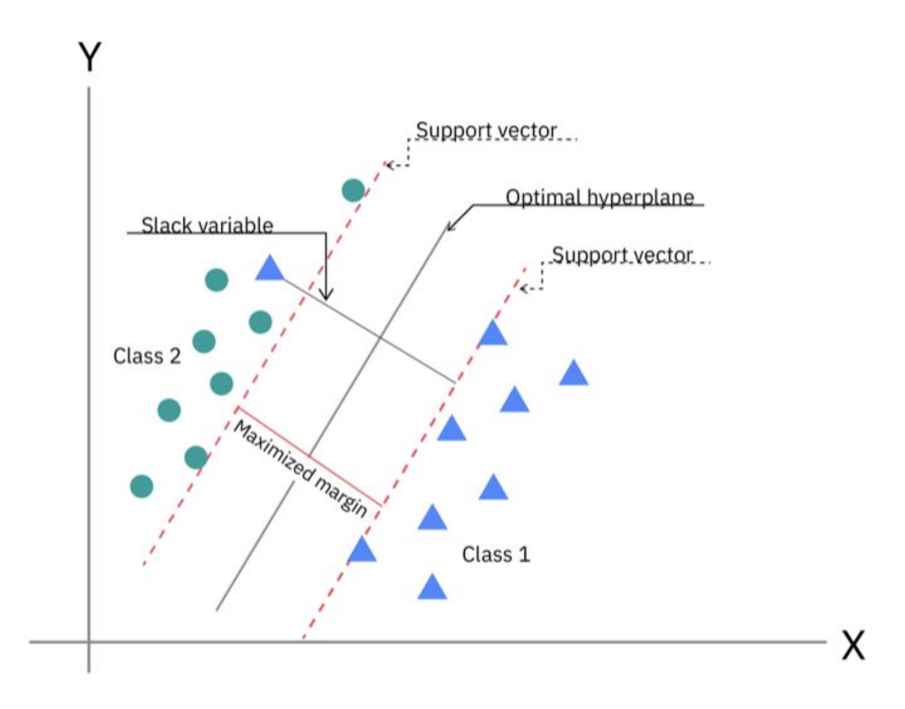
Tuning Parameters
Performance Results
Confusion Matrix

03
LIMITATIONS & ADVANTAGES

04
CONCLUSION

INTRODUCTION

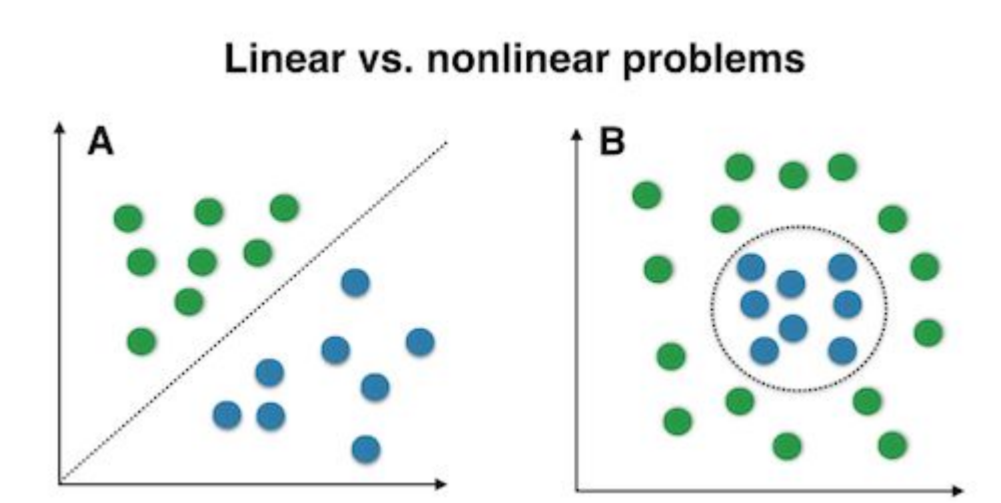
WHAT IS SVM AGAIN?



- Finds optimal separating hyperplane
- Maximizes margin
- Support vectors define the boundary
- Handles imperfect separation (slack variables)

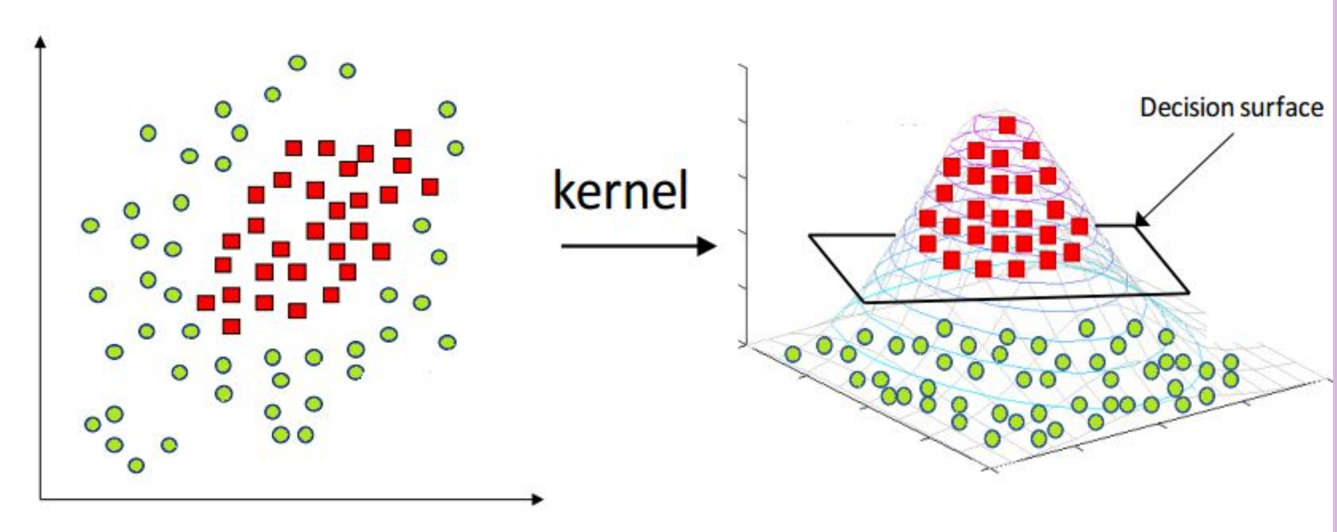
WHAT IS SVM AGAIN?

- Linear separation works for simple data
- Nonlinear patterns are common in real-world datasets
- SVM can handle both



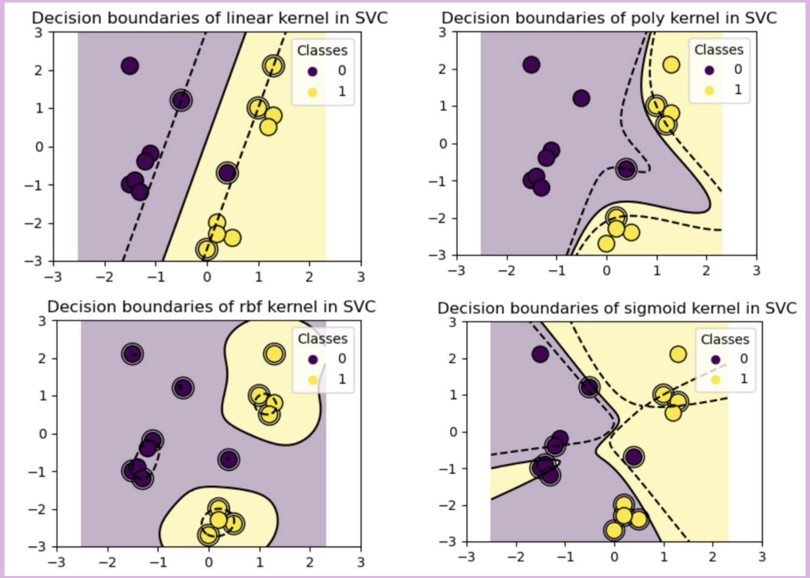
WHAT IS SVM AGAIN?

- Kernel -> maps data to higher dimensions
- Makes nonlinear data linearly separable
- SVM can model very complex shapes while still relying on simple margin maximizing principle



WHAT IS SVM AGAIN?

- Kernels → shape + flexibility
- Linear: simple, fast
- Polynomial: **degree** controls curvature
- RBF: **gamma** controls smoothness
- Sigmoid: S-shaped boundary
- Parameter C (cost): strictness of model
- Kernel choice + tuning matter



OUR RESEARCH

Angelina Jolie

Christina Applegate

Olivia Newton-John

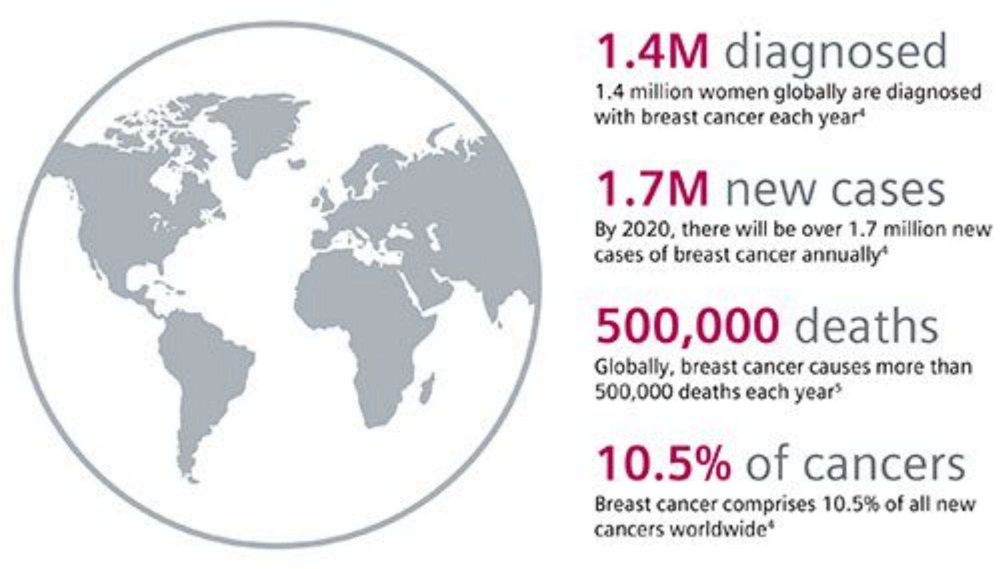
What do all these inspiring women have in common?

OUR RESEARCH QUESTION

“Is Support Vector Machines an accurate method predicting whether a cell is malignant or benign?”

IMPORTANCE

Societal relevance:



1.4M diagnosed

1.4 million women globally are diagnosed with breast cancer each year[4]

1.7M new cases

By 2020, there will be over 1.7 million new cases of breast cancer annually[4]

500,000 deaths

Globally, breast cancer causes more than 500,000 deaths each year[5]

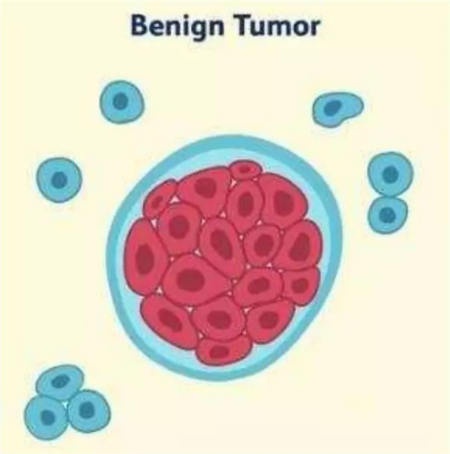
10.5% of cancers

Breast cancer comprises 10.5% of all new cancers worldwide[4]

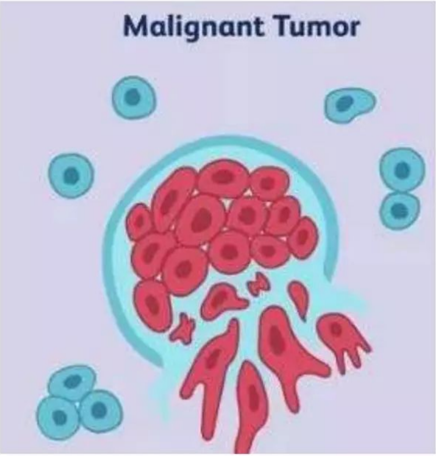
Scientific Relevance:

- Speeds up the investigation process
- Instead of busy doctors looking at the cells, a simple ML model can do the prediction.

Cancer ≠Tumor: Abnormal growth of cells causing a mass of tissue



- Benign tumors stay in their primary location.

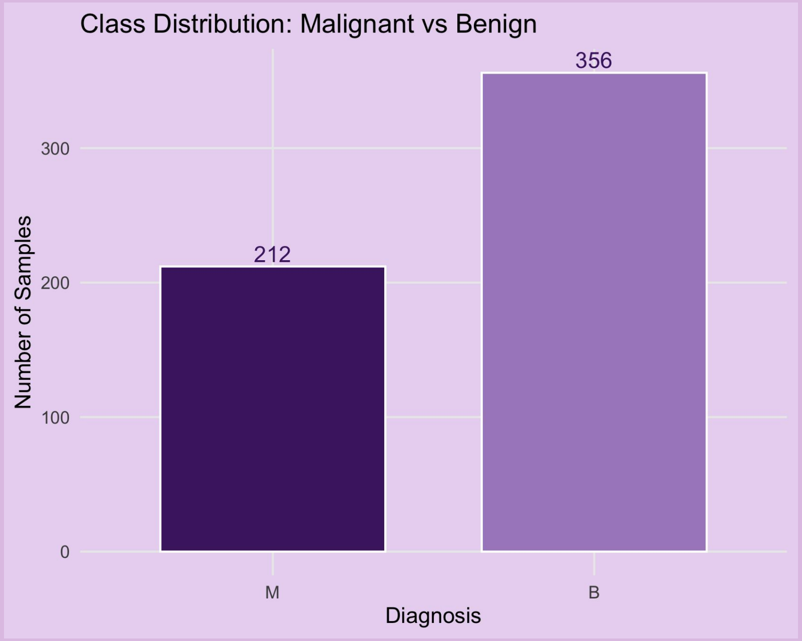


- Malignant tumors are cancerous and invade other sites.

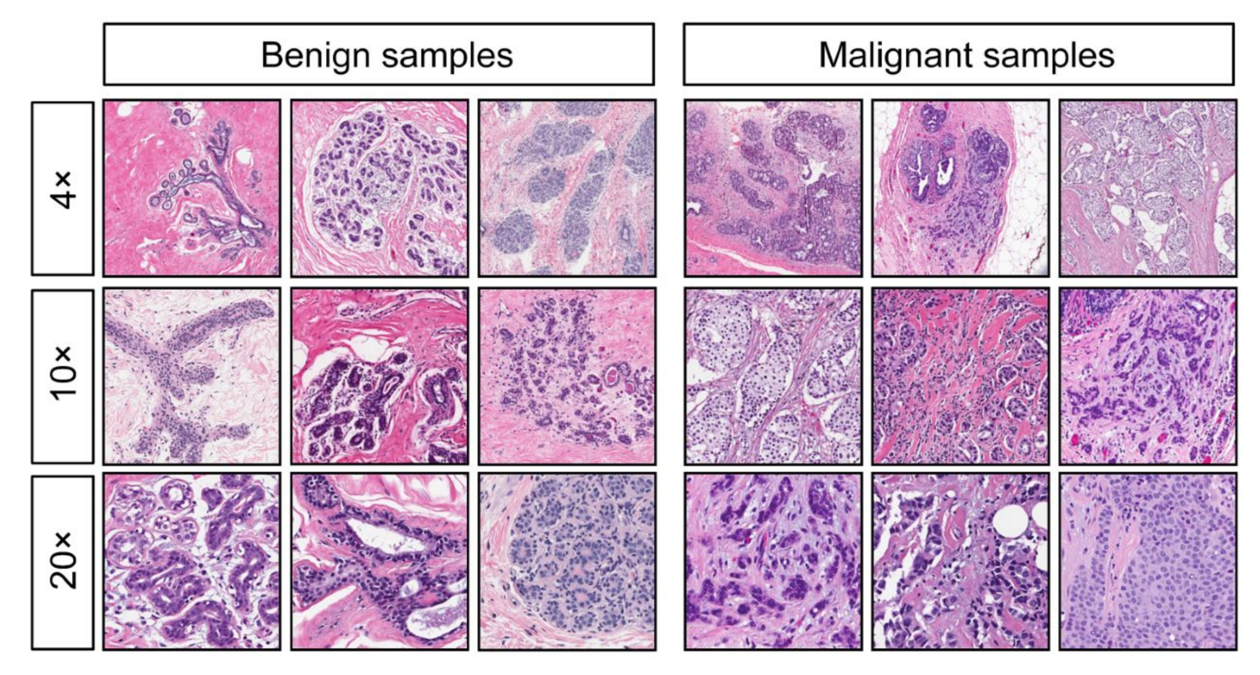
DATA SET USED

DATA SET USED

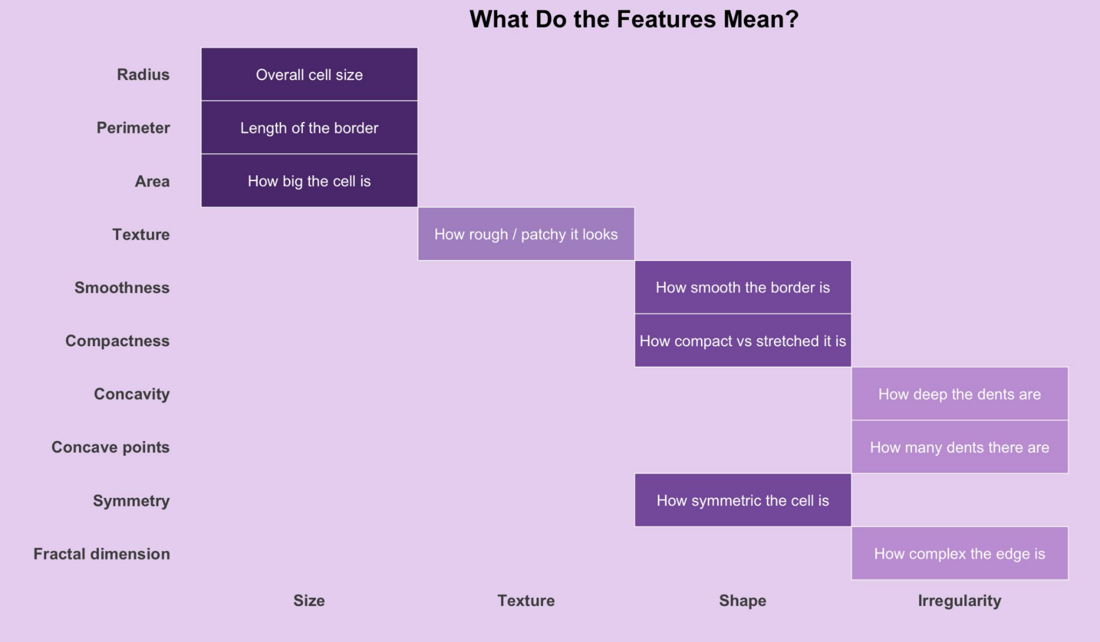
- Breast Cancer Wisconsin (Diagnostic) Data Set
- 569 samples
- Target: malignant vs benign (diagnosis)



DATA SET USED



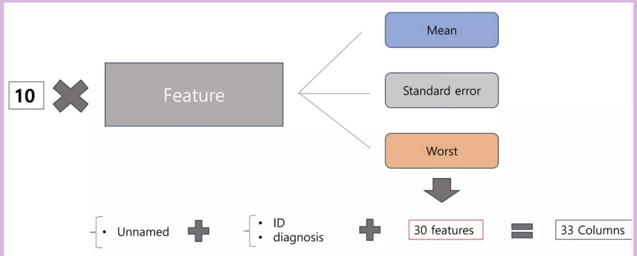
DATA SET USED



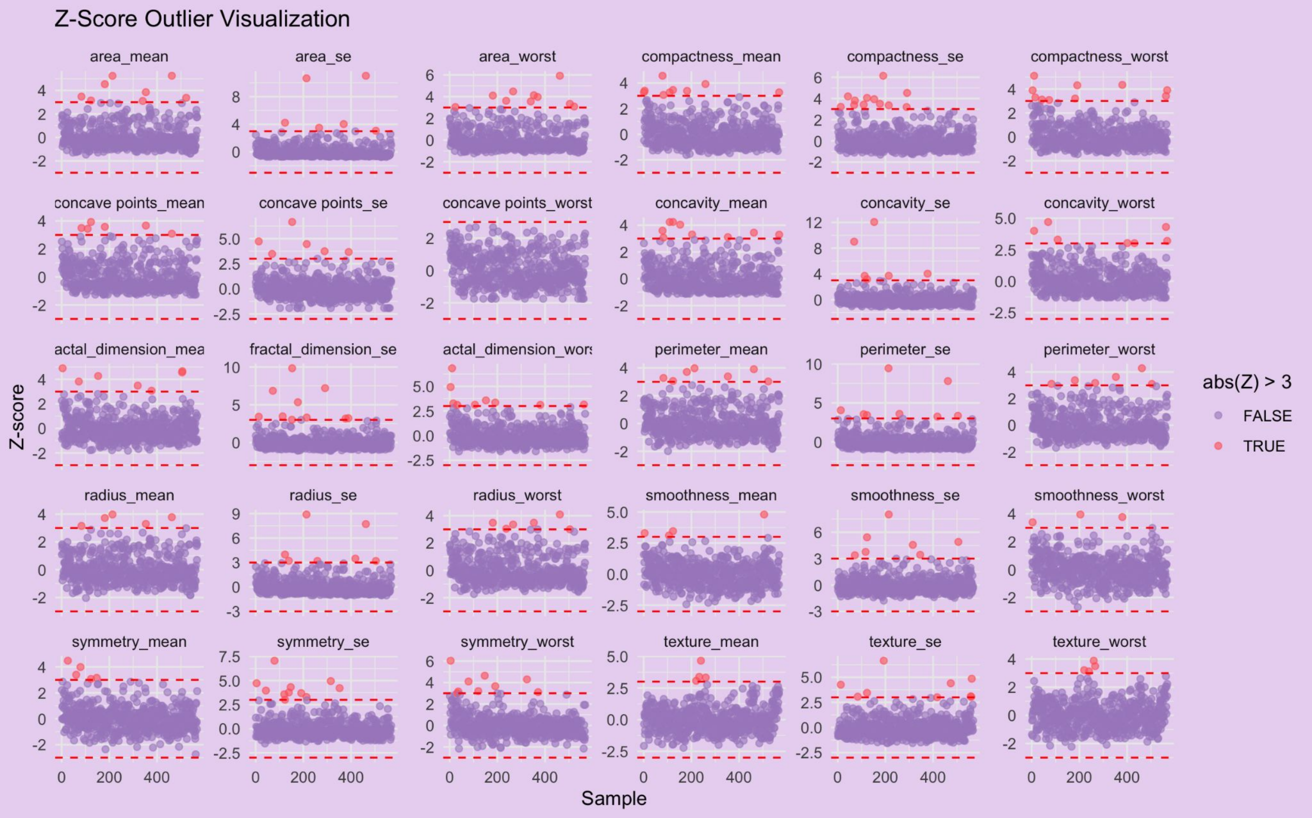
What Do the Features Mean?

	Size	Texture	Shape	Irregularity
Radius	Overall cell size			
Perimeter	Length of the border			
Area	How big the cell is			
Texture		How rough / patchy it looks		
Smoothness			How smooth the border is	
Compactness			How compact vs stretched it is	
Concavity				How deep the dents are
Concave points				How many dents there are
Symmetry			How symmetric the cell is	
Fractal dimension				How complex the edge is

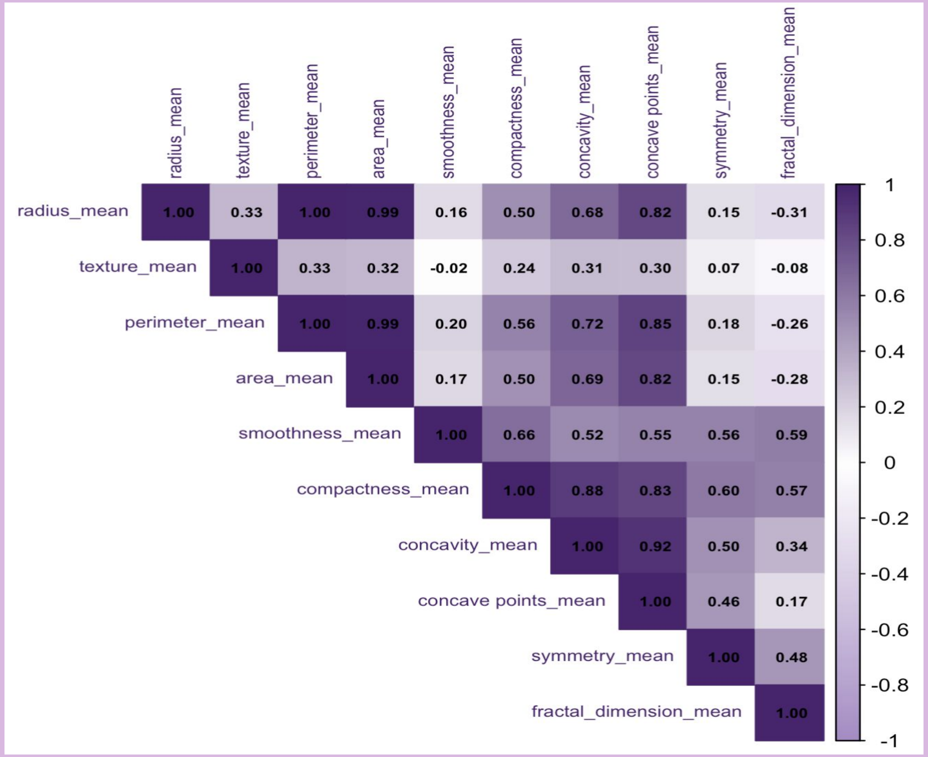
- 30 numeric features
- Mean, SE, “worst” values



DATA SET USED

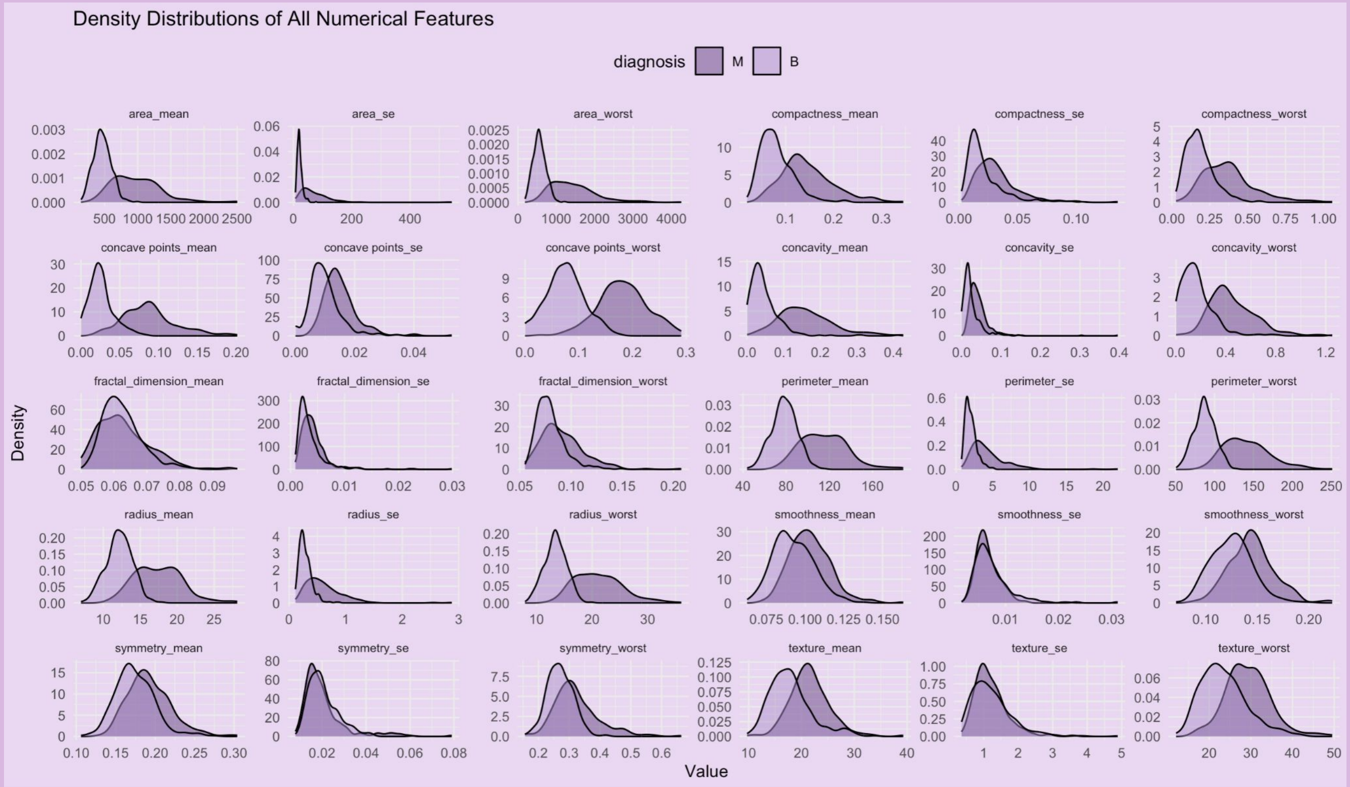


DATA SET USED



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	1.00	0.33	1.00	0.99	0.16	0.50	0.68	0.82	0.15	-0.31
texture_mean		1.00	0.33	0.32	-0.02	0.24	0.31	0.30	0.07	-0.08
perimeter_mean			1.00	0.99	0.20	0.56	0.72	0.85	0.18	-0.26
area_mean				1.00	0.17	0.50	0.69	0.82	0.15	-0.28
smoothness_mean					1.00	0.66	0.52	0.55	0.56	0.59
compactness_mean						1.00	0.88	0.83	0.60	0.57
concavity_mean							1.00	0.92	0.50	0.34
concave points_mean								1.00	0.46	0.17
symmetry_mean									1.00	0.48
fractal_dimension_mean										1.00

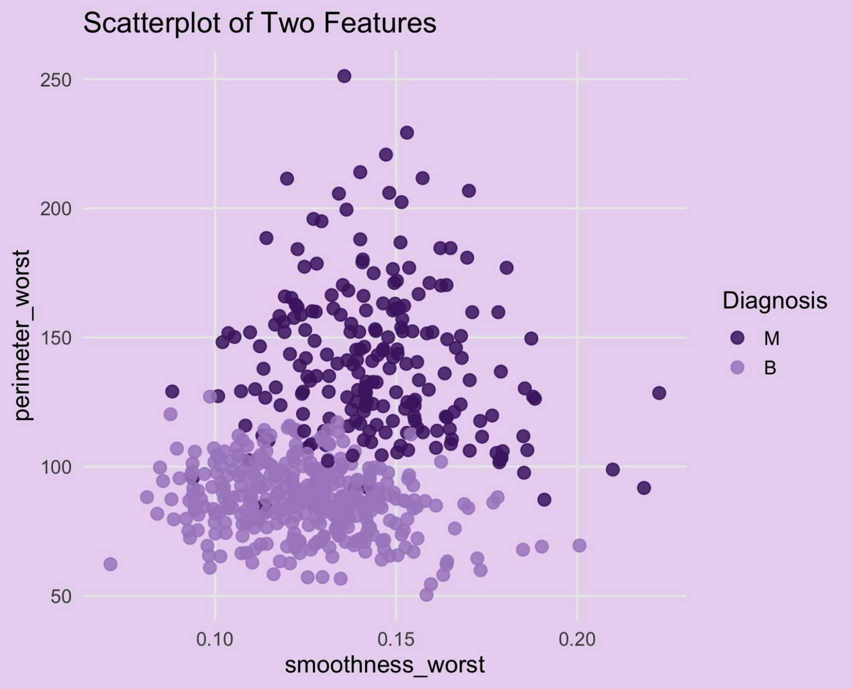
DATA SET USED

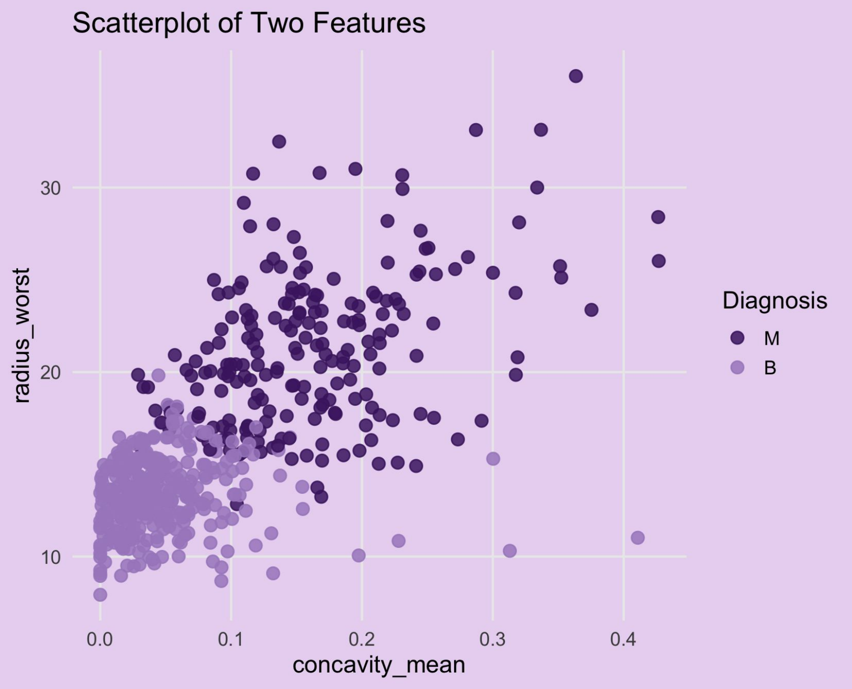


-> Clear class differences

-> Early signs of separability

DATA SET USED



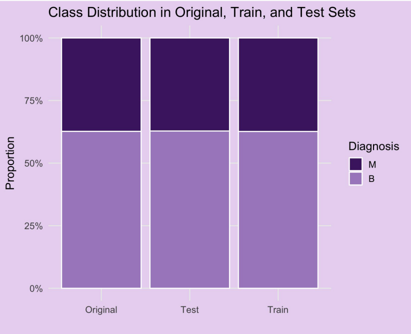


DATA PREPARATION

1

Removed "unnamed" and "id" column

2



Training, test split (70-30)

3

Scaling of variables

LINEAR SVM

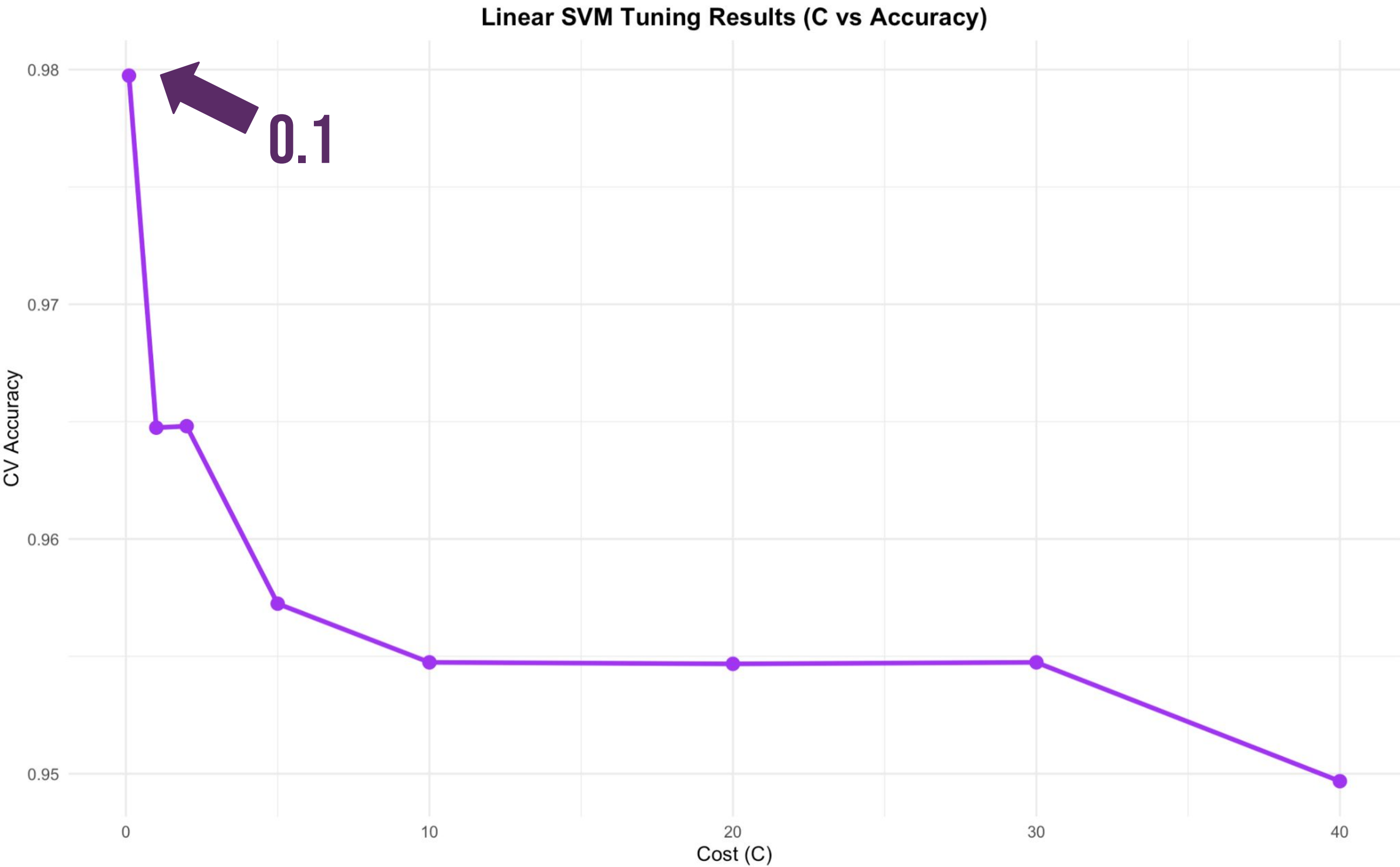
HYPERPARAMETER TUNING

The Cost (C)

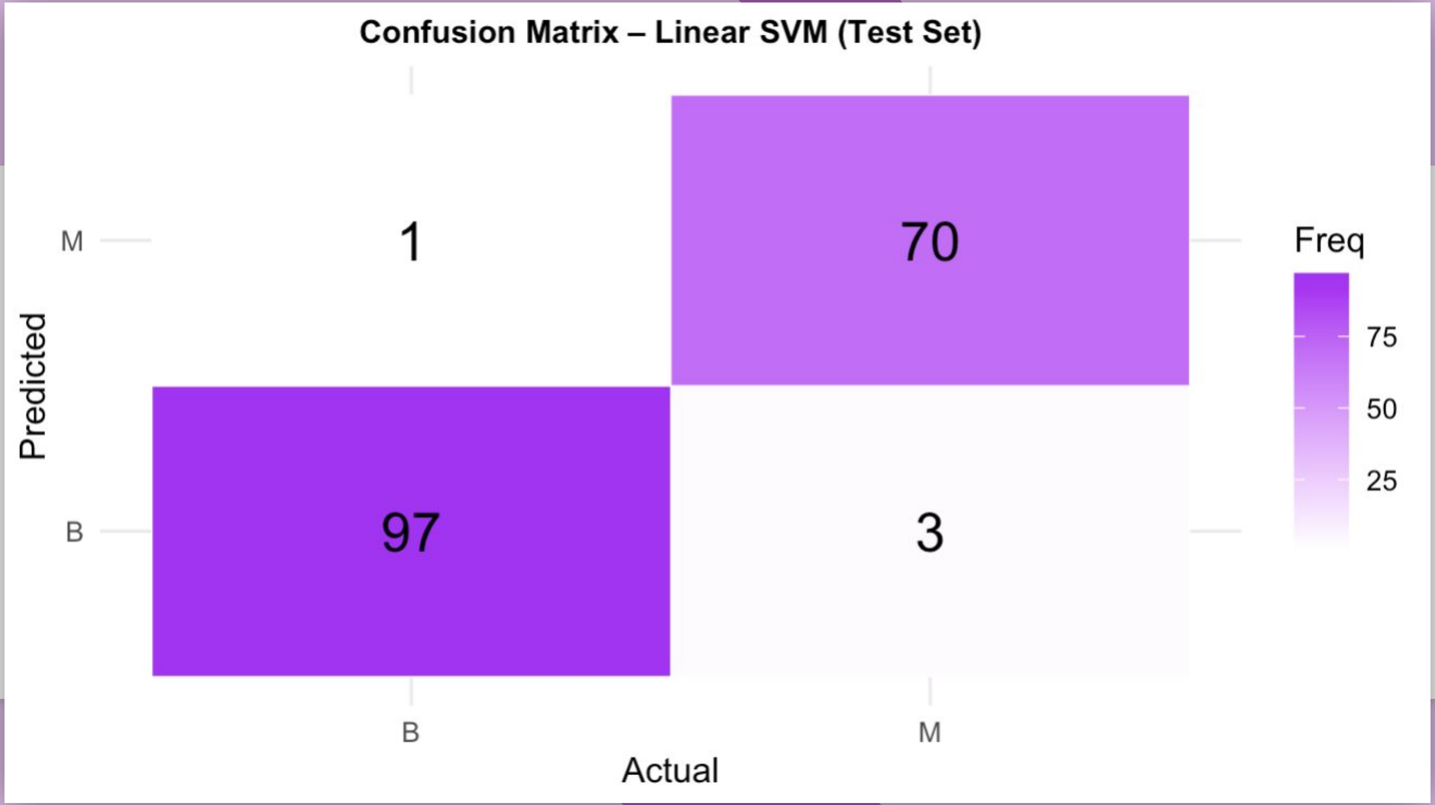
- How strictly it separates the classes
  - Low C: softer margin
  - High C: harder margin (risk overfitting)
- Only this parameter (simple & fast)
- Several C values (0.1 - 40)

10-FOLD CROSS VALIDATION

- Best balance of bias and variance
  - 5-fold → higher variance (less stable)
  - 20-fold → lower variance but too slow
- Standard in ML research
- Works well with data of our size


Linear SVM Tuning Results (C vs Accuracy)
CV Accuracy
0.98
0.97
0.96
0.95
0
10
20
30
40
Cost (C)

0.1



Confusion Matrix – Linear SVM (Test Set)

Predicted \ Actual	B	M
M	1	70
B	97	3

RESULTS - PERFORMANCE METRICS ✓

TRAIN ACCURACY	TEST ACCURACY	KAPPA	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY
0.987	0.977	0.952	0.959	0.99	0.974

- Strong generalization
- Model performs far better than chance
- Detects most malignant tumours

- Detects almost all benign tumours
- Model treats both classes fairly despite the imbalance

Erafus

RADIAL SVM

HYPERPARAMETER TUNING

The Cost (C)

- How strictly it separates the classes
- Several C values (0.1 - 40)

The Sigma (γ, gamma)

- Controls how “wiggly” the decision boundary is
- Values from 0.001 to 0.2

10-Fold Cross Validation

- Good balance between reliability and computational cost
- Repeated **3 times** to make the results more stable (reduces randomness)

IN SUMMARY...

1

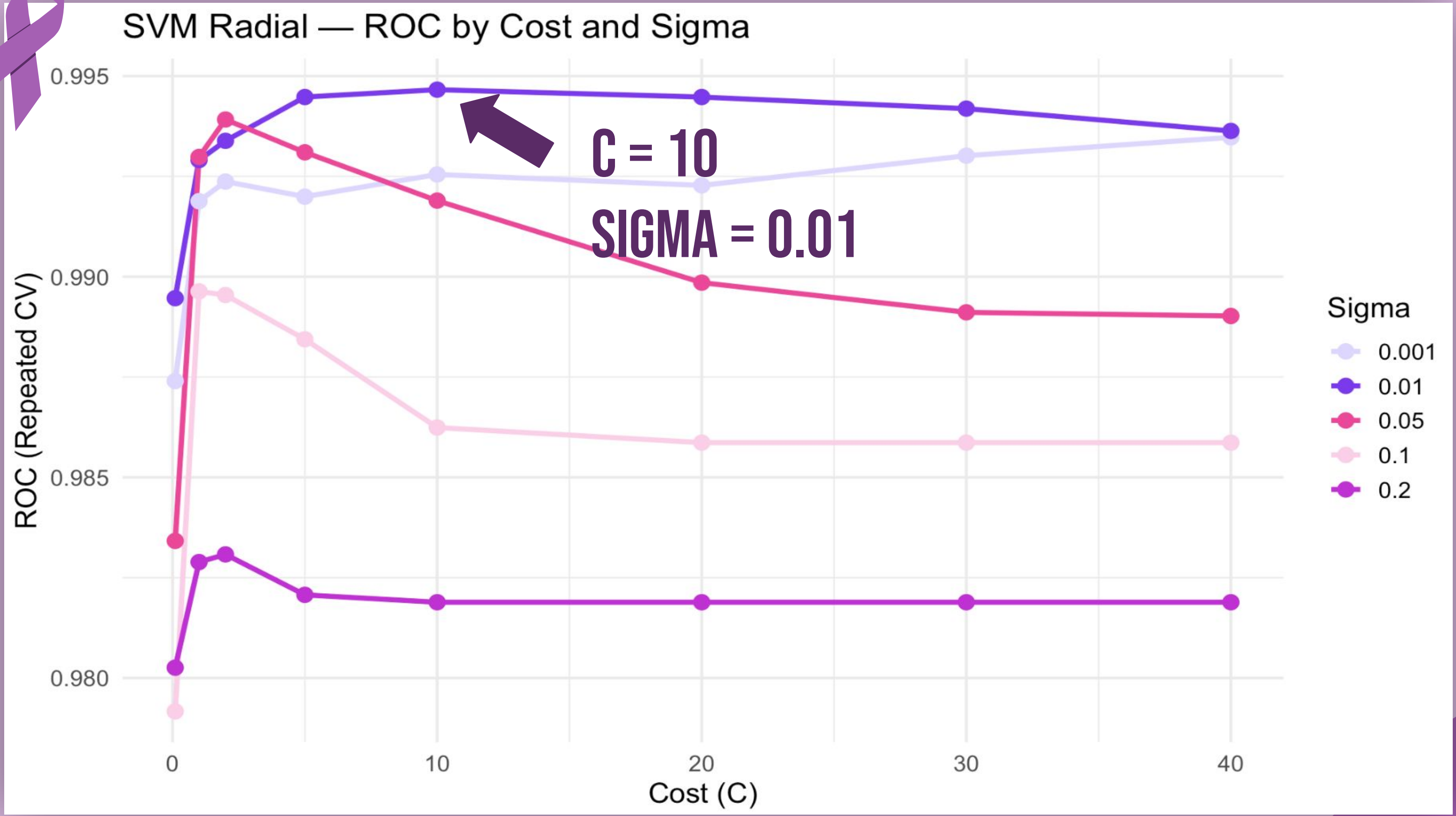
Created a grid with several values of C and Sigma

2

Compared all combinations using repeated cross-validation

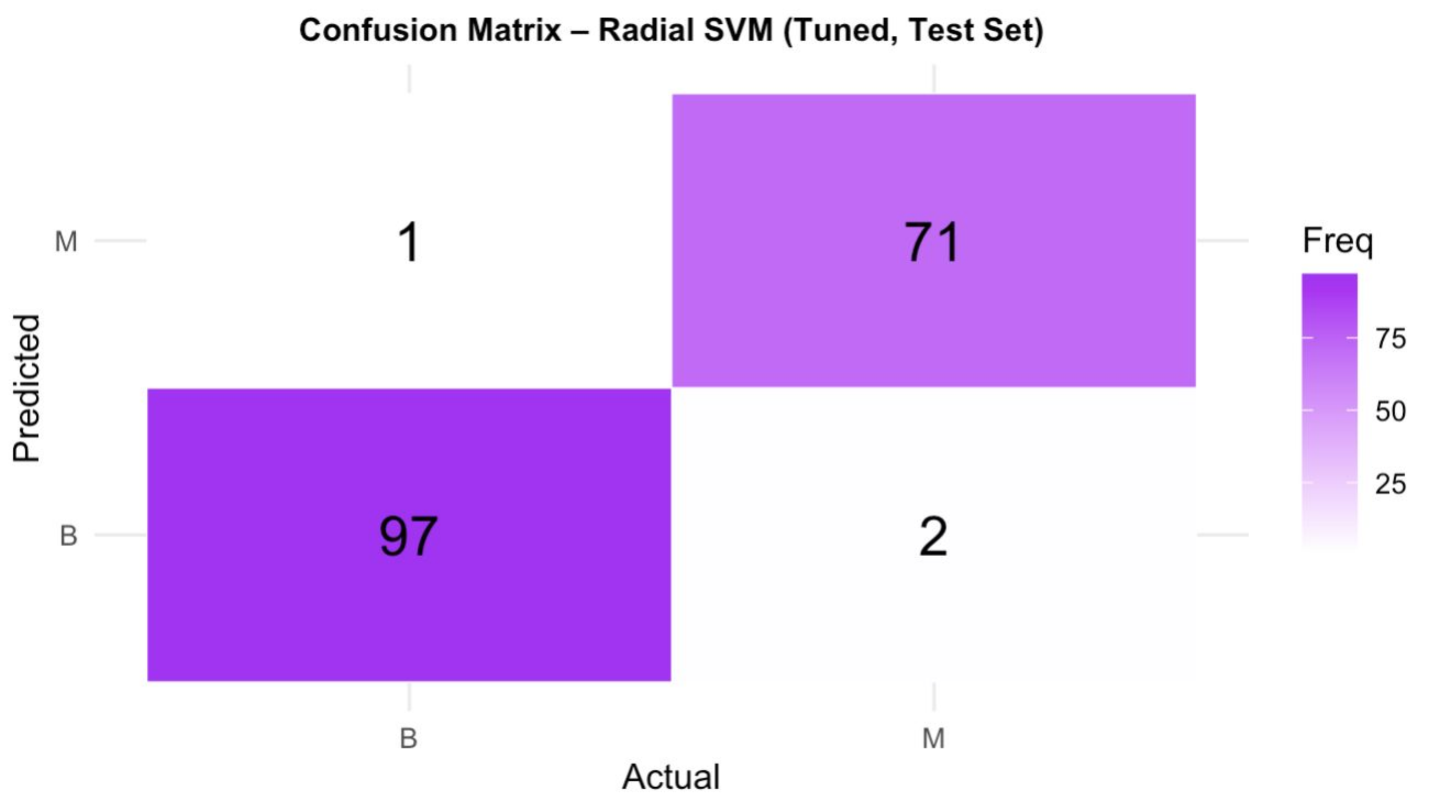
3

Selected the one with the highest ROC AUC



C = 10

SIGMA = 0.01



Confusion Matrix – Radial SVM (Tuned, Test Set)

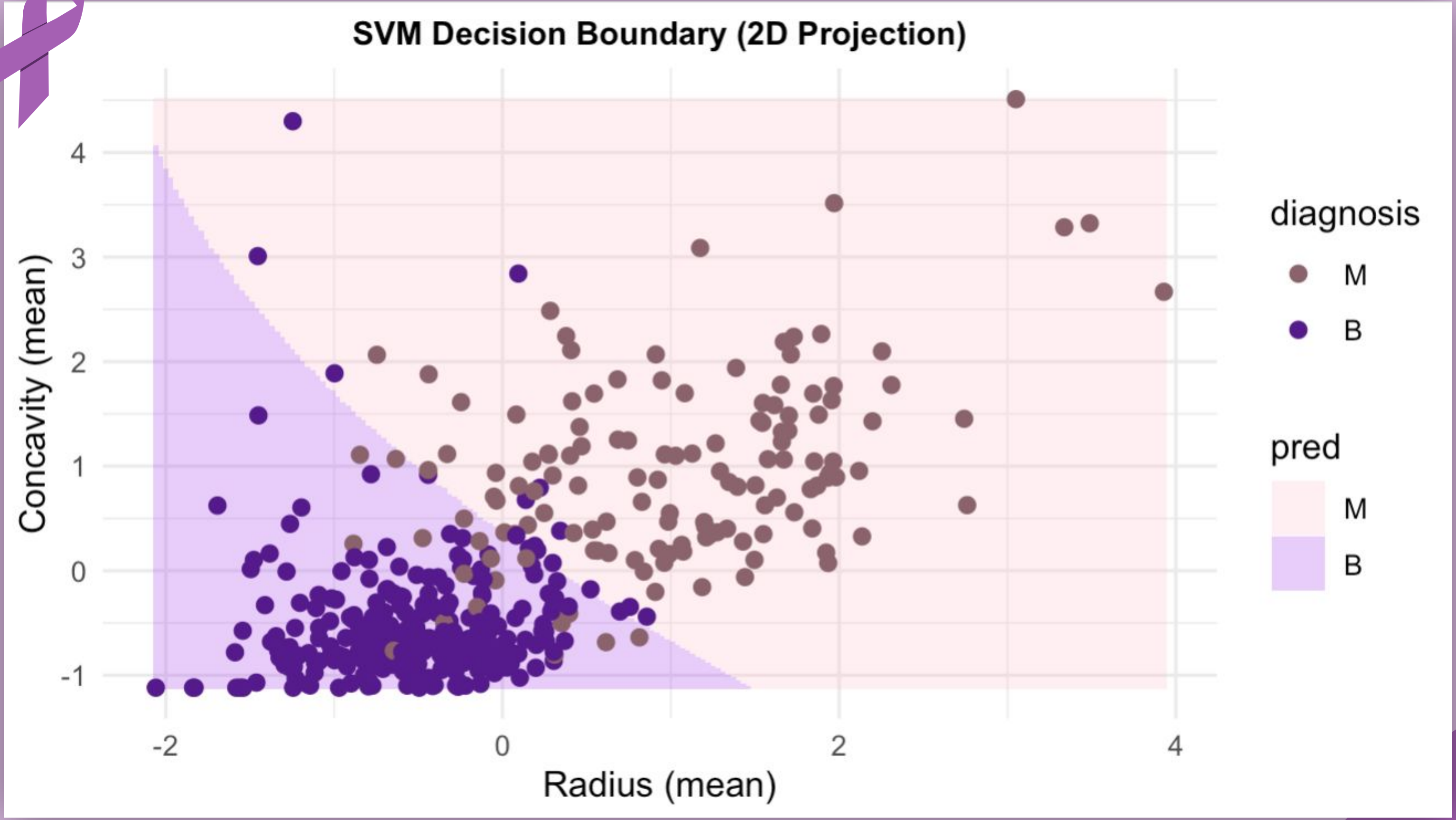
Predicted \ Actual	B	M
M	1	71
B	97	2

RESULTS - PERFORMANCE METRICS

TRAIN ACCURACY	TEST ACCURACY	KAPPA	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY
0.987	0.982	0.964	0.973	0.99	0.981

- Both models perform very well
- Radial SVM performs slightly better

- Suggests relationship between features and diagnosis is not perfectly linear


SVM Decision Boundary (2D Projection)
Concavity (mean)
4
3
2
1
0
-1
-2
0
2
4
Radius (mean)
diagnosis
● M
● B
pred
M
B

MODEL PERFORMANCE COMPARISON

RESULTS - COMPARISON

MODEL	TEST ACCURACY	KAPPA	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY
Logistic Regression	0.965	0.929	0.973	0.959	0.966
Linear SVM	0.977	0.952	0.959	0.99	0.974
Radial SVM (Default)	0.971	0.940	0.932	1.00	0.966
Radial SVM	0.982	0.964	0.973	0.99	0.981

THE RESEARCH QUESTION

“Is Support Vector Machines an accurate method for predicting whether a cell is malignant or benign?”

YES

Method is a good predictor

HIGH ACCURACY

0.982

BEST MODEL

Better performance than baseline models

LIMITATIONS AND ADVANTAGES

LIMITATIONS

- Slow Training
- Parameter Tuning Difficulty
- Classes Overlapping
- Sensitive To Scaling

ADVANTAGES

- Good With High Dimensionality
- Nonlinear Capability
- Can Handle Outliers
- Memory Efficient

REAL WORLD IMPLEMENTATION

ILLUSTRATE PATIENT CONTEXT

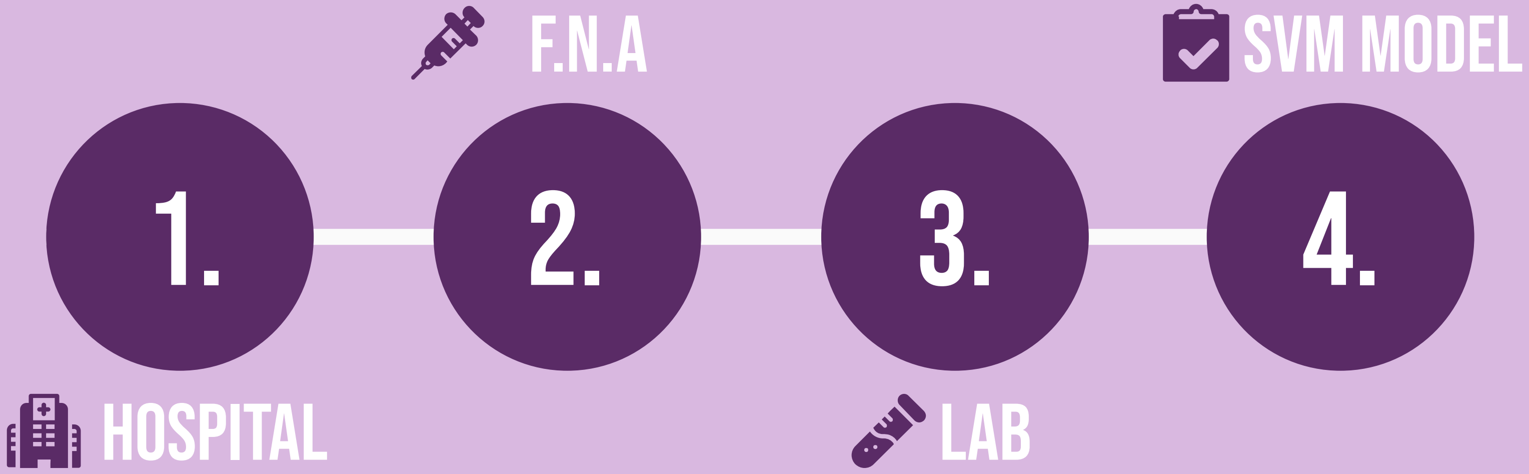
SARA

42

GRAPHIC DESIGNER

TIRED & SLIGHT LUMP

HOSPITAL



1. HOSPITAL

2. F.N.A

3. LAB

4. SVM MODEL

INVESTIGATING TEST RESULTS

VARIABLE	TEST VALUE
radius_mean	12.47
texture_mean	18.60
area_mean	481.9
smoothness_mean	0.09965
compactness_mean	0.10580
concavity_mean	0.08005

CONCLUSION

CONCLUSION

Investigate whether Support Vector Machines can predict if cell is benign or malignant

Important to properly scale the data and tune the hyperparameters

Compared the model with a logistic regression and similar SVM models

Support Vector Machine handles high dimensionality, outliers and has good nonlinear capabilities

A tuned radial SVM performed the best with an Accuracy of 0.982

SVM effective in classification tasks in medical diagnostic context

REFERENCES

- https://www.datacamp.com/tutorial/support-vector-machines–r
- Coussement and van den Poel (2008).Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34, 313–327
- Kim, H.-S. and S.-Y. Sohn (2010): Support vector machines for default prediction of SME's based on technology credit. European Journal of Operational Research, 201, 838-846
- Jae H. Min, Young-Chan Lee (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28, 603–614
- Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018 Jan-Feb;15(1):41-51. doi: 10.21873/cgp.20063. PMID: 29275361; PMCID: PMC5822181.
- Xulei Yang, Qing Song and A. Cao, "Weighted support vector machine for data classification," *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Montreal, QC, Canada, 2005, pp. 859-864 vol. 2, doi: 10.1109/IJCNN.2005.1555965.
- https://medium.com/@RobuRishabh/support-vector-machines-svm-27cd45b74fbb