MICHIGAN INSTITUTE FOR DATA & AI IN SOCIETY
UNIVERSITY OF MICHIGAN

Schmidt Sciences

BINARY CLASSIFICATION TO PREDICT TUMOR TYPE (BENIGN OR MALIGNANT)

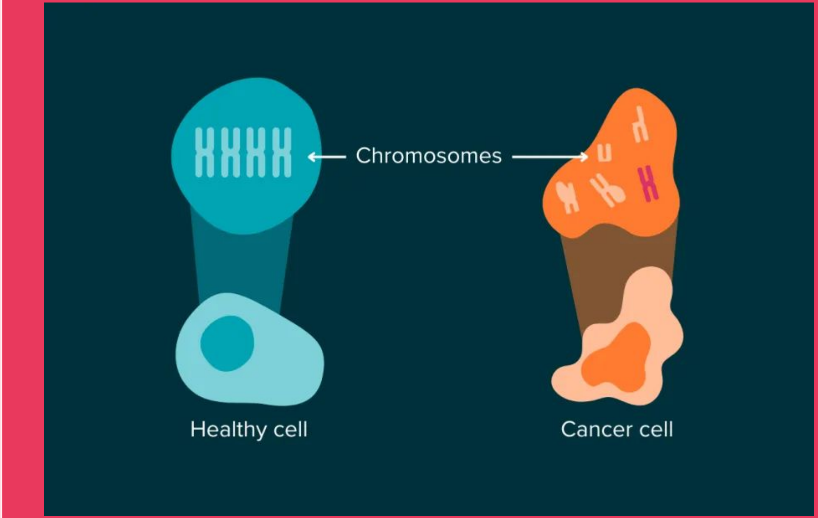
BENIGN BY DESIGN 💅
Paige Bowling
Long-Jing Hsu
Yan Ying Tan
Samuel Akingbade

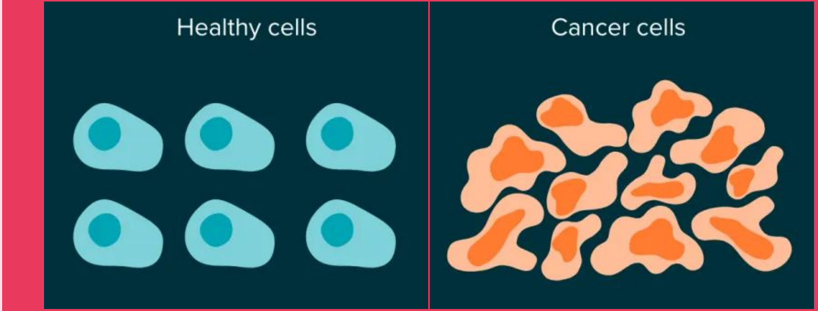
U-M Knowledge-Guided Machine Learning (KGML) Workshop | August 8, 2025

INTRODUCTION

What could go wrong?

- The contents of the nucleus are very different for cancer cells
- Looking at cells, it can be determined if they are
  - M (Malignant): cancerous
  - B (Benign): non-cancerous





WHAT ARE WE MEASURING?

- The dataset features describing cell nuclei characteristics.
  - Each instance represents a single breast mass sample.
- Categories:
  - Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, and Fractal dimension
- Three measurements/features each:
  - Mean, Standard Error, and Worst

(Kaggle.com)

SIMPLE MODELS ARE A SCIENTISTS BEST FRIEND

Reliability

Accountability

Interpretability

Sustainability

Executability

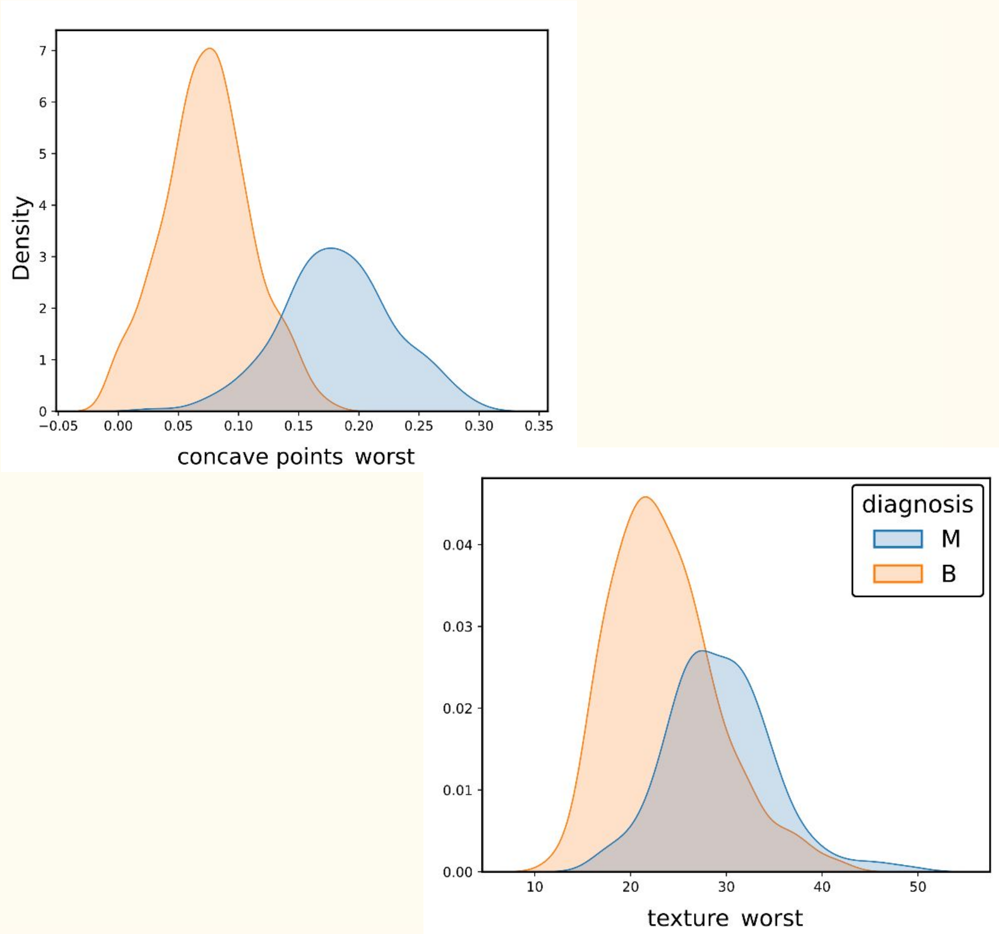
(Forbes 2021)

RESEARCH QUESTION

- **How many features are necessary to accurately predict a patient's breast cancer diagnosis using simple ML models?**
- **Measurements of success:**
  - High accuracy & Low false positive rate

Data Analysis

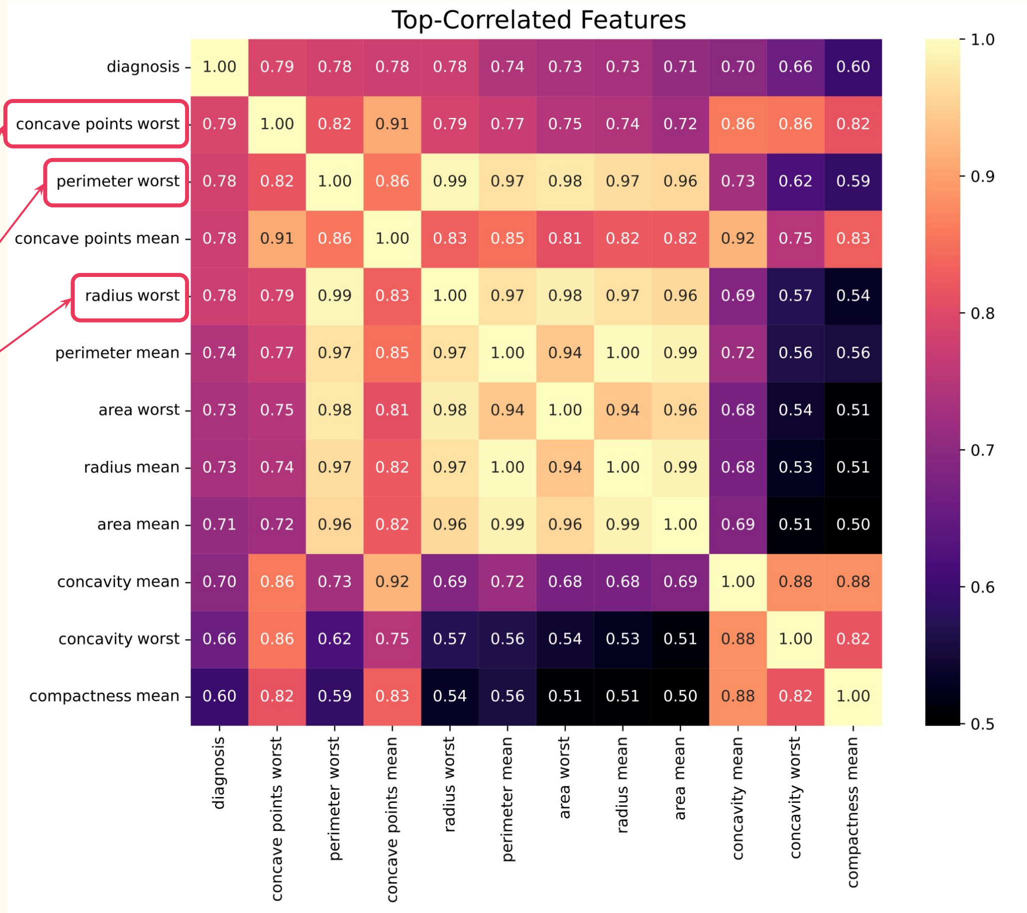
Bimodal Distribution



Data Analysis

Top-Correlated Features

Top 3 Features



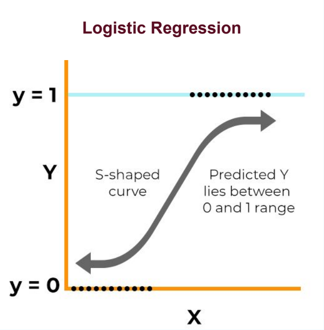
Top-Correlated Features

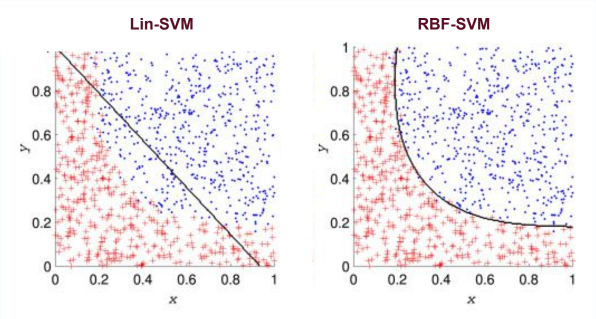
	diagnosis	concave points worst	perimeter worst	concave points mean	radius worst	perimeter mean	area worst	radius mean	area mean	concavity mean	concavity worst	compactness mean
diagnosis	1.00	0.79	0.78	0.78	0.78	0.74	0.73	0.73	0.71	0.70	0.66	0.60
concave points worst	0.79	1.00	0.82	0.91	0.79	0.77	0.75	0.74	0.72	0.86	0.86	0.82
perimeter worst	0.78	0.82	1.00	0.86	0.99	0.97	0.98	0.97	0.96	0.73	0.62	0.59
concave points mean	0.78	0.91	0.86	1.00	0.83	0.85	0.81	0.82	0.82	0.92	0.75	0.83
radius worst	0.78	0.79	0.99	0.83	1.00	0.97	0.98	0.97	0.96	0.69	0.57	0.54
perimeter mean	0.74	0.77	0.97	0.85	0.97	1.00	0.94	1.00	0.99	0.72	0.56	0.56
area worst	0.73	0.75	0.98	0.81	0.98	0.94	1.00	0.94	0.96	0.68	0.54	0.51
radius mean	0.73	0.74	0.97	0.82	0.97	1.00	0.94	1.00	0.99	0.68	0.53	0.51
area mean	0.71	0.72	0.96	0.82	0.96	0.99	0.96	0.99	1.00	0.69	0.51	0.50
concavity mean	0.70	0.86	0.73	0.92	0.69	0.72	0.68	0.68	0.69	1.00	0.88	0.88
concavity worst	0.66	0.86	0.62	0.75	0.57	0.56	0.54	0.53	0.51	0.88	1.00	0.82
compactness mean	0.60	0.82	0.59	0.83	0.54	0.56	0.51	0.51	0.50	0.88	0.82	1.00

PREDICTING THE DIAGNOSIS

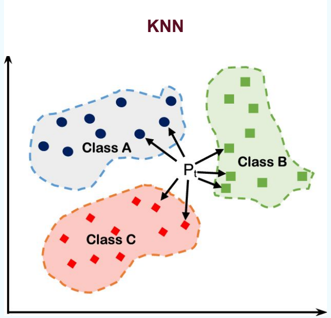
Setup

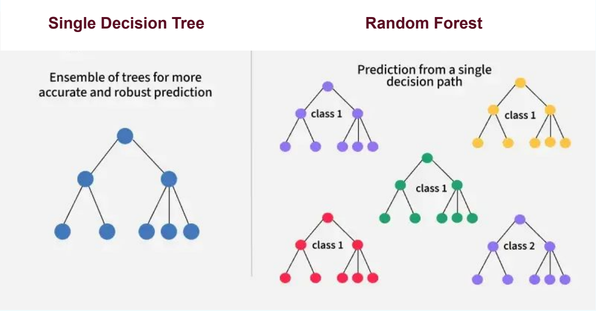
- Split data set 80/20
- Tested 8 different ML models
    - Logistic Regression, SVM (Lin & RBF), KNN, Decision Tree, Random Forest, Gradient Boost, LightGBM
- Tested use of different features (of the 30 possible)


Logistic Regression

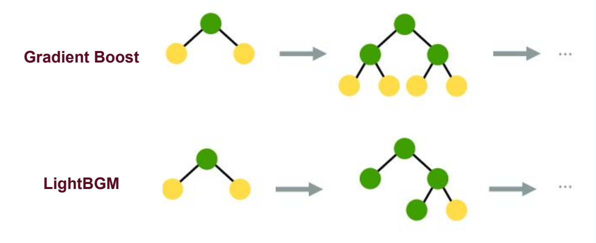

Lin-SVM

RBF-SVM

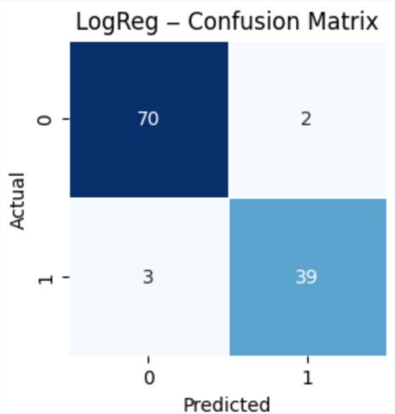

KNN


Single Decision Tree

Random Forest
Ensemble of trees for more accurate and robust prediction
Prediction from a single decision path
class 1
class 1
class 1
class 1
class 2

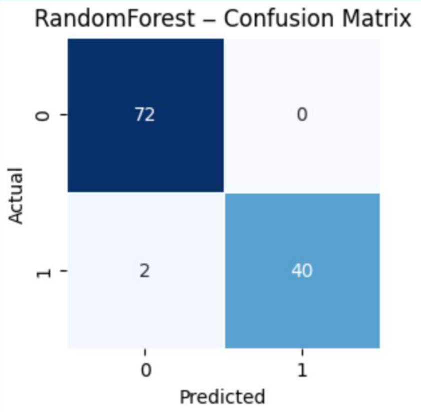


Preliminary Results



LogReg – Confusion Matrix

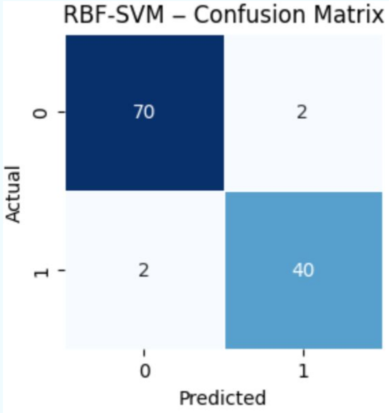
Actual \ Predicted	0	1
0	70	2
1	3	39



RandomForest – Confusion Matrix

Actual \ Predicted	0	1
0	72	0
1	2	40

Models	Accuracy	ROC-AUC
LogReg	0.956	0.994
Random Forest	**0.982**	**0.996**
RBF-SVM	0.965	0.995



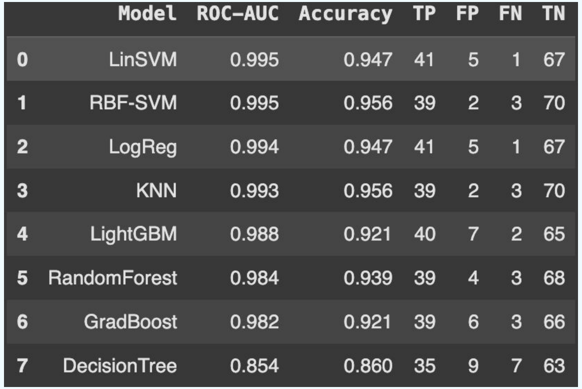
RBF-SVM – Confusion Matrix

Actual \ Predicted	0	1
0	70	2
1	2	40

Representative Model

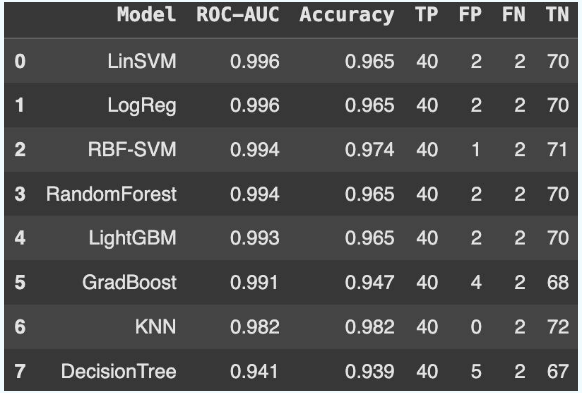
	ROC-AUC			Accuracy		
Models	**3**	**10**	**All**	**3**	**10**	**All**
LogReg	0.994	**0.996**	0.994	0.947	0.965	0.956
Random Forest	0.984	0.994	**0.996**	0.939	0.965	**0.982**
RBF-SVM	**0.995**	0.994	0.995	**0.956**	**0.974**	0.965

3 Parameters



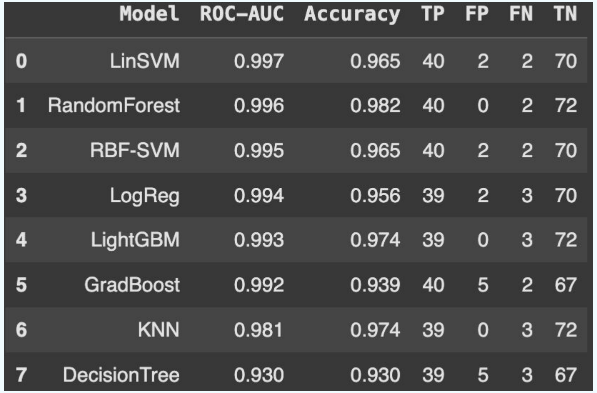
	Model	ROC-AUC	Accuracy	TP	FP	FN	TN
0	LinSVM	0.995	0.947	41	5	1	67
1	RBF-SVM	0.995	0.956	39	2	3	70
2	LogReg	0.994	0.947	41	5	1	67
3	KNN	0.993	0.956	39	2	3	70
4	LightGBM	0.988	0.921	40	7	2	65
5	RandomForest	0.984	0.939	39	4	3	68
6	GradBoost	0.982	0.921	39	6	3	66
7	DecisionTree	0.854	0.860	35	9	7	63

10 Parameters



	Model	ROC-AUC	Accuracy	TP	FP	FN	TN
0	LinSVM	0.996	0.965	40	2	2	70
1	LogReg	0.996	0.965	40	2	2	70
2	RBF-SVM	0.994	0.974	40	1	2	71
3	RandomForest	0.994	0.965	40	2	2	70
4	LightGBM	0.993	0.965	40	2	2	70
5	GradBoost	0.991	0.947	40	4	2	68
6	KNN	0.982	0.982	40	0	2	72
7	DecisionTree	0.941	0.939	40	5	2	67

30 Parameters



	Model	ROC-AUC	Accuracy	TP	FP	FN	TN
0	LinSVM	0.997	0.965	40	2	2	70
1	RandomForest	0.996	0.982	40	0	2	72
2	RBF-SVM	0.995	0.965	40	2	2	70
3	LogReg	0.994	0.956	39	2	3	70
4	LightGBM	0.993	0.974	39	0	3	72
5	GradBoost	0.992	0.939	40	5	2	67
6	KNN	0.981	0.974	39	0	3	72
7	DecisionTree	0.930	0.930	39	5	3	67

CONCLUSION

What can we conclude? Future directions?

1

Simple models

- Easy-to-understand algorithms
- Choose the ones most linked to the result
- Avoid using similar features

2

You don’t need 30 features to predict a diagnosis

- With only 3 of top features achieves >95% fidelity
- Using one feature from each category is approx. same as using all

3

Future improvements

1. Can combine (stack or blend) models
2. Find larger dataset with more diverse data (not interdependent measurements)

“KEEP IT SIMPLE”