# Breast Cancer Prediction

## Machine Learning Project

*Bikash Thapa Magar, Jiamin Yao, Olivia Clark*

# Motivation and Research Aim

**Motivation**

- 2.3 million women diagnosed with breast cancer; 685,000 deaths globally (2020)
- Use data to infer correlations between cancer diagnoses and chosen features

---

**The Plan**

| | |
|---|---|
| *Input* | {Size, Texture, Compactness, etc} |
| *Methods* | {Classification, Clustering} |
| *Output* | {Malignant, Benign} |

# Related Work

- Logistic regression, KNN, SVM, naïve Bayes, decision tree, and random forest with a breast cancer dataset [1]
  - Removed non-numerical values and normalized numeric values [1]
  - Heat map and other plots created using R studio, Minitab, and Python [1]

- ML better for diagnoses compared to traditional methods (human interpretation) [2, 3, 4]
  - AI system had higher area under AUC-ROC curve than human detector by 11.5% [2]
  - Image analysis was used with ML to increase accuracy of breast cancer testing [3, 4]

- LDA and SVM resulted in increased accuracy [5]

# Dataset & Features

Breast Cancer Dataset by M Yasser H from Kaggle

**32 features, 569 examples (75%/25% train/test)**

## Features:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | ID | 9 | symmetry mean | 17 | smoothness se | 25 | perimeter worst |
| 2 | diagnosis | 10 | concavity mean | 18 | compactness se | 26 | area worst |
| 3 | radius mean | 11 | concave points mean | 19 | concavity se | 27 | smoothness worst |
| 4 | texture mean | 12 | fractal dimension mean | 20 | concave points se | 28 | compactness worst |
| 5 | perimeter mean | 13 | radius se | 21 | symmetry se | 29 | concavity worst |
| 6 | area mean | 14 | texture se | 22 | fractal dimension se | 30 | concave points worst |
| 7 | smoothness mean | 15 | perimeter se | 23 | radius worst | 31 | symmetry worst |
| 8 | compactness mean | 16 | area se | 24 | texture worst | 32 | fractal dimension worst |

Citation: Yasser H., M. (2022). Breast Cancer Dataset. Retrieved January 2024 from https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset.

# Methods - Algorithms

## Classification

- Logistic Regression
- K-Nearest Neighbors
- Naive Bayes
- Decision Tree
- Random Forest
- Boosting
- Support Vector Machines

---

## Clustering

- K-Means
- Hierarchical
- Mean-Shift

# Methods - Metrics

## Classification

- Accuracy
- Recall
- Precision
- F1-Score

---

## Clustering

**Internal Metrics**

- Silhouette Score
- Davies-Bouldin index
- Calinski-Harabasz Index

**External Metrics**

- Adjusted Rand Index
- Normalized Mutual Information

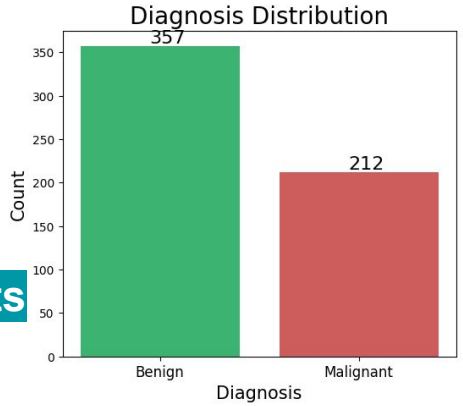
# Results - Data Process & Feature Selection

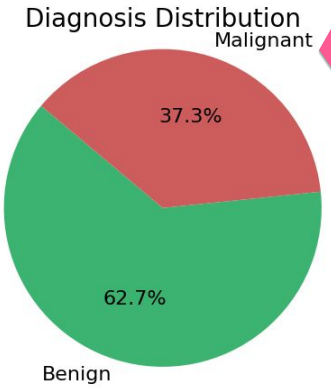
**Remove empty value**

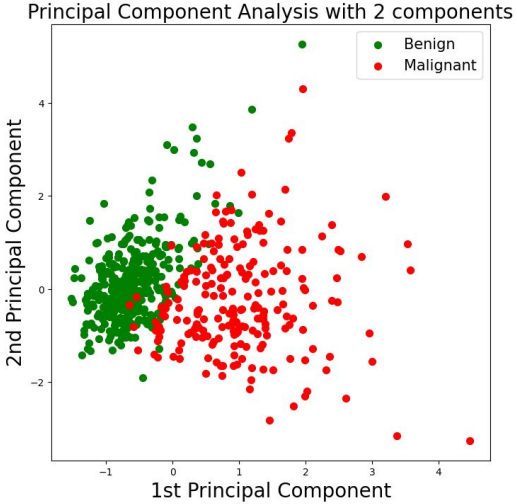
**Encode diagnosis (B->0, M->1)**

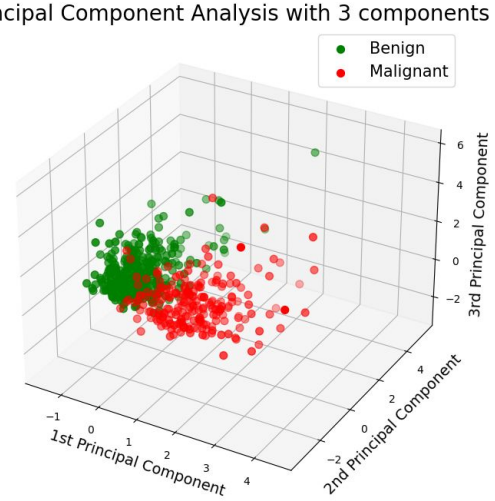
**Standard Scaler**

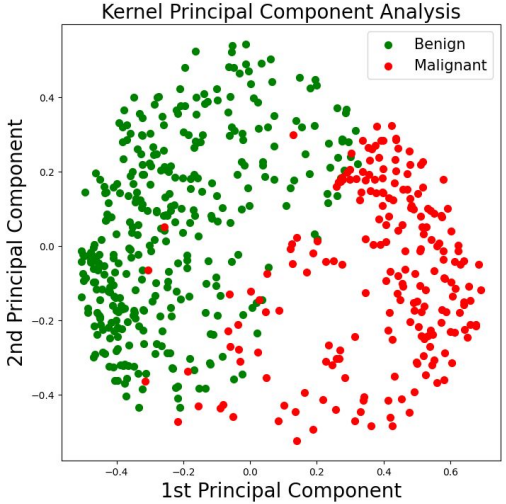
**Split data into train(75%) and test(25%) sets**



# Results - Model Training

## Find hyperparameters before training model

- Grid Search to find hyperparameters for a model
- Find the best score
- Classification use Recall; Clustering use ARI

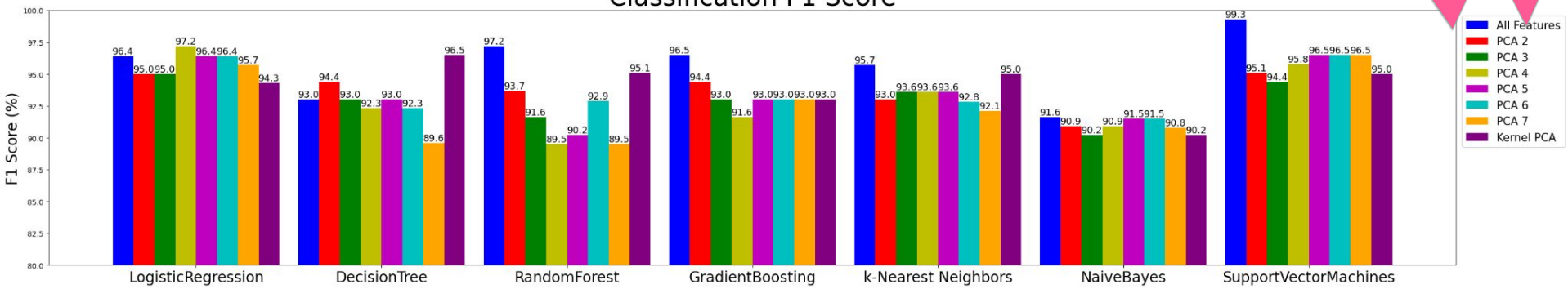
## Get Evaluation Metrics with hyperparameters

- Fit the model for original dataset and 7 kinds of transformed dataset
- Classfication get Train/Test Accuracy, Recall, Precision, F1 Score
- Clustering get Silhouette Score, DBI, CHI, ARI, NMI

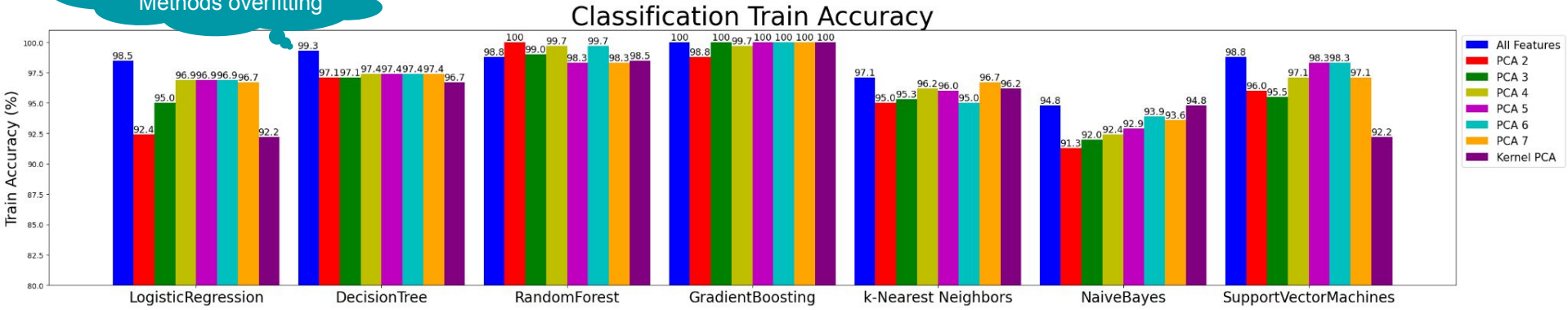
**Decision Tree Hyperparameters**

| Parameter / Data | max_depth | min_samples_split | min_samples_leaf |
|---|---|---|---|
| All Features | None | 2 | 10 |
| PCA 2 | None | 1 | 10 |
| PCA 3 | None | 1 | 10 |
| PCA 4 | 20 | 4 | 5 |
| PCA 5 | 10 | 4 | 2 |
| PCA 6 | None | 4 | 10 |
| PCA 7 | 10 | 2 | 2 |
| Kernel PCA | 10 | 2 | 10 |

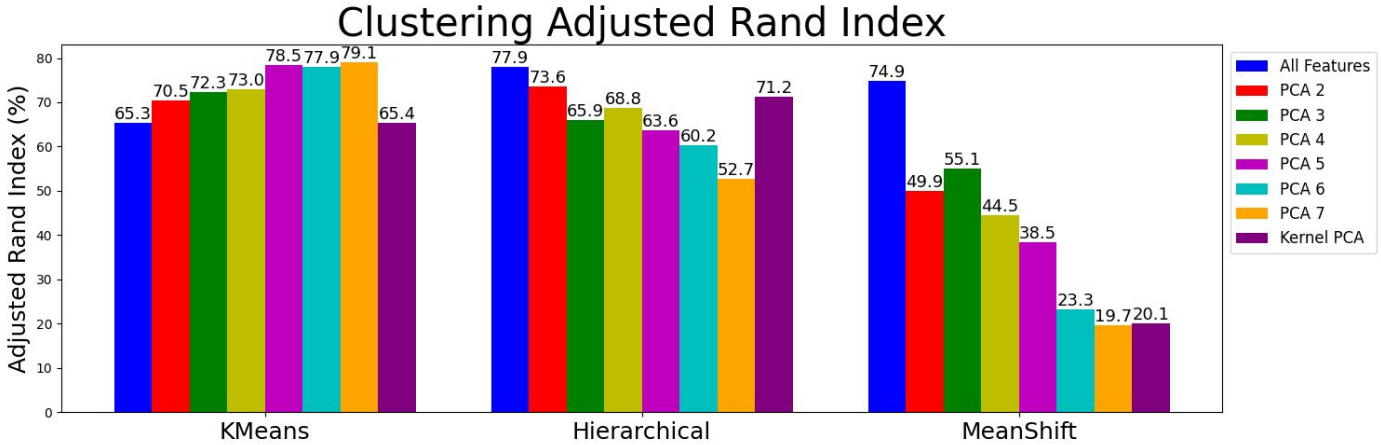
# Results - Classification

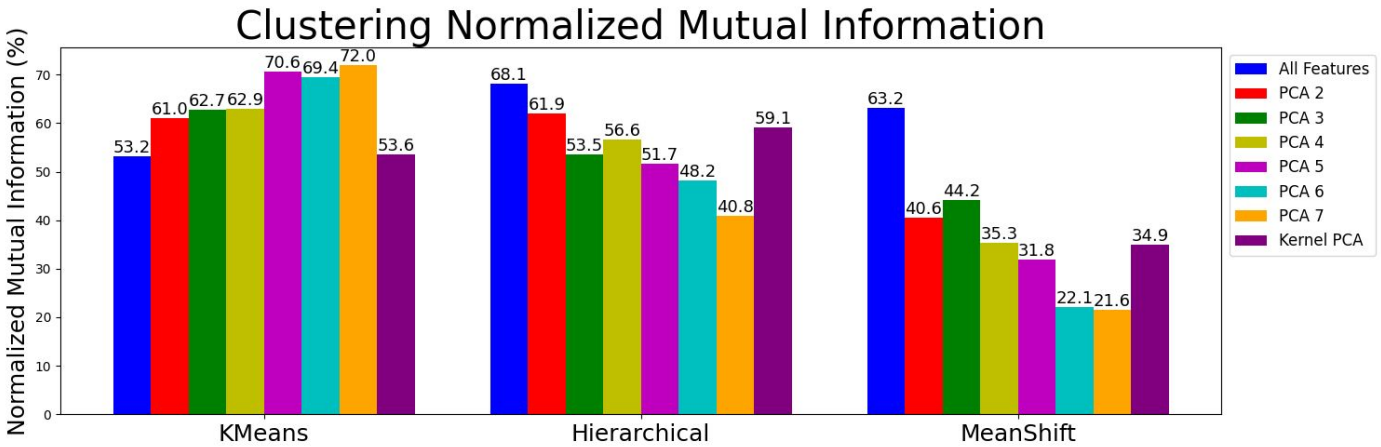



Tree-Based Methods overfitting



# Results - Clustering




62.7%

# Conclusion

**Overall, Classification performance is better than Clustering**

**Classification:**

- Original dataset performance is better than other transformed data for most of classification methods.
- SVM in original dataset is the best(99.3%);

**Clustering:**

- Hierarchical and MeanShift in original dataset performance is better than transformed data.
- K-Means PCA performance is better than original dataset and kernel PCA, and the performance is getting better when PCA components increase from 2 to 7.
- K-Means in PCA with 7 components is the best(79.1%) in clustering.

# Discussion

**Classification:**

- Any feature selection can help to improve classification performance?
- How to prevent Tree-Based methods' overfitting?

**Clustering:**

- Any feature selection can help to improve performance for Hierarchical and MeanShift?
- K-Means can meet best performance with what PCA components?

# References


[1] Ak, M.F. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare 2020, 8, 111. https://doi.org/10.3390/healthcare8020111

[2] McKinney, S.M., Sieniek, M., Godbole, V. et al. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94 (2020). https://doi.org/10.1038/s41586-019-1799-6

[3] W. Nick Street, W. H. Wolberg, and O. L. Mangasarian "Nuclear feature extraction for breast tumor diagnosis", Proc. SPIE 1905, Biomedical Image Processing and Biomedical Visualization, (29 July 1993); https://doi.org/10.1117/12.148698

[4] Wolberg WH, Street WN, Mangasarian OL. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Anal Quant Cytol Histol. 1995 Apr;17(2):77-87. PMID: 7612134.

[5] David A. Omondiagbe et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 495 012033

[6] Yasser H., M. (2022). Breast Cancer Dataset. Retrieved January 2024 from https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset.

# Thank You