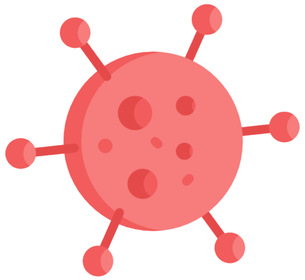


Breast Cancer Detection

Using Machine Learning and Deep Learning

Subject: Data Mining

Lecturer: PHAUK Sokkhey & CHAN Sophal
Presenter: CHOENG Veyseng

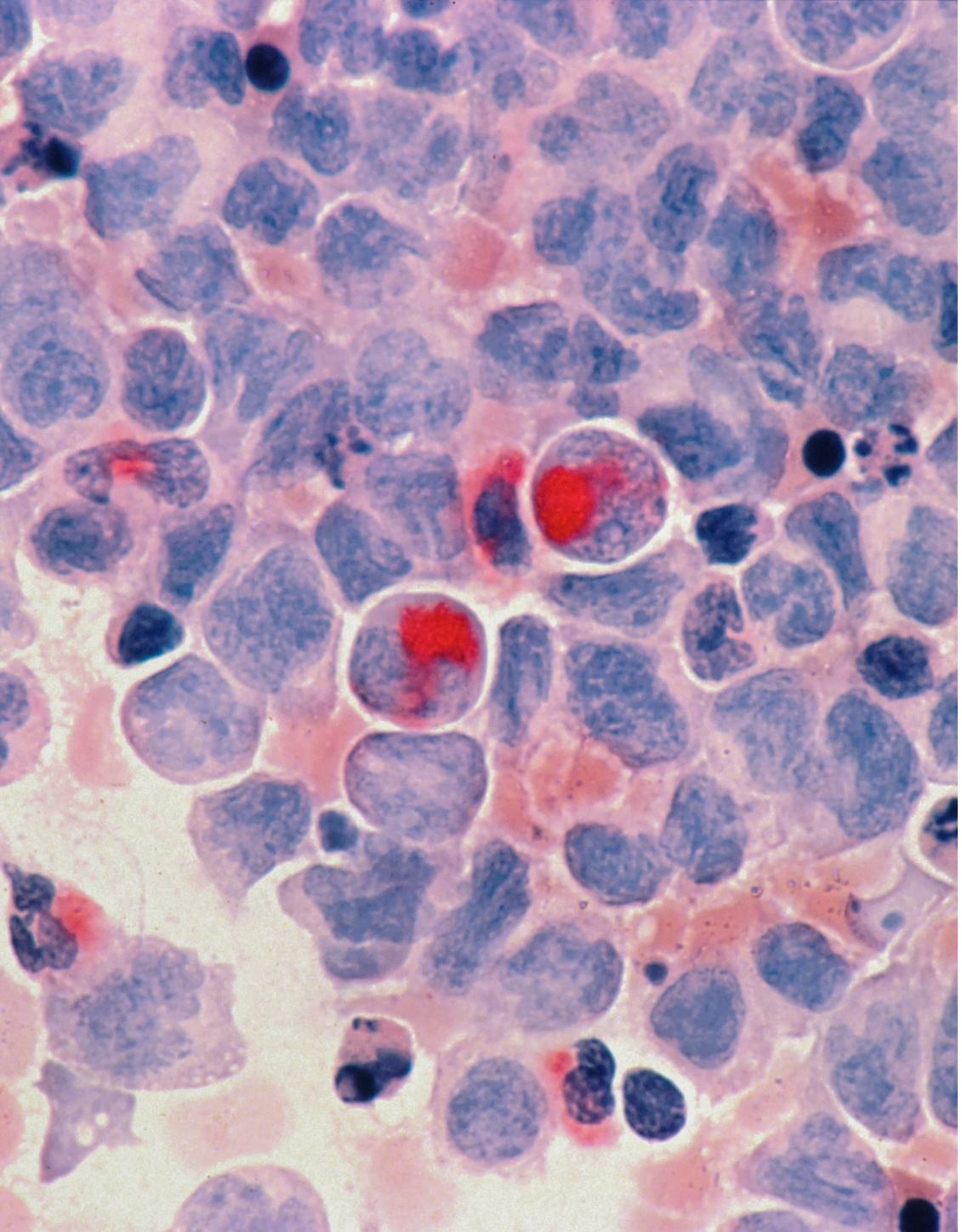
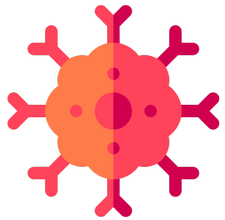


Table of Content

Problem



Breast cancer is one of the most harmful diseases that negatively effects for women public health

Detecting the tumor on breast is not an easy task which is required medical technical and support from technology

Project Ideation

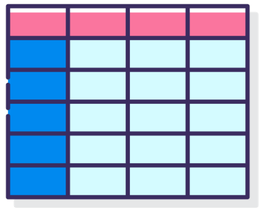
Breast Cancer Images

Breast Cancer Data

Machine Learning Models

Data Introduction

Breast Cancer Data From Kaggle



Breast Cancer Wisconsin (Diagnostic) Dataset [1] is one of the most popular dataset for predicting the input data is benign or malignant as the tumors of breast cancer.

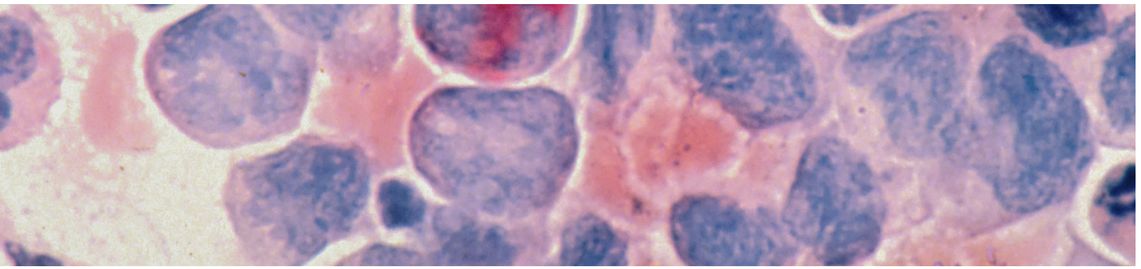
Data Introduction

Breast Cancer Data From Kaggle

569
Rows in total

357 benign

212 malignant



30 Features

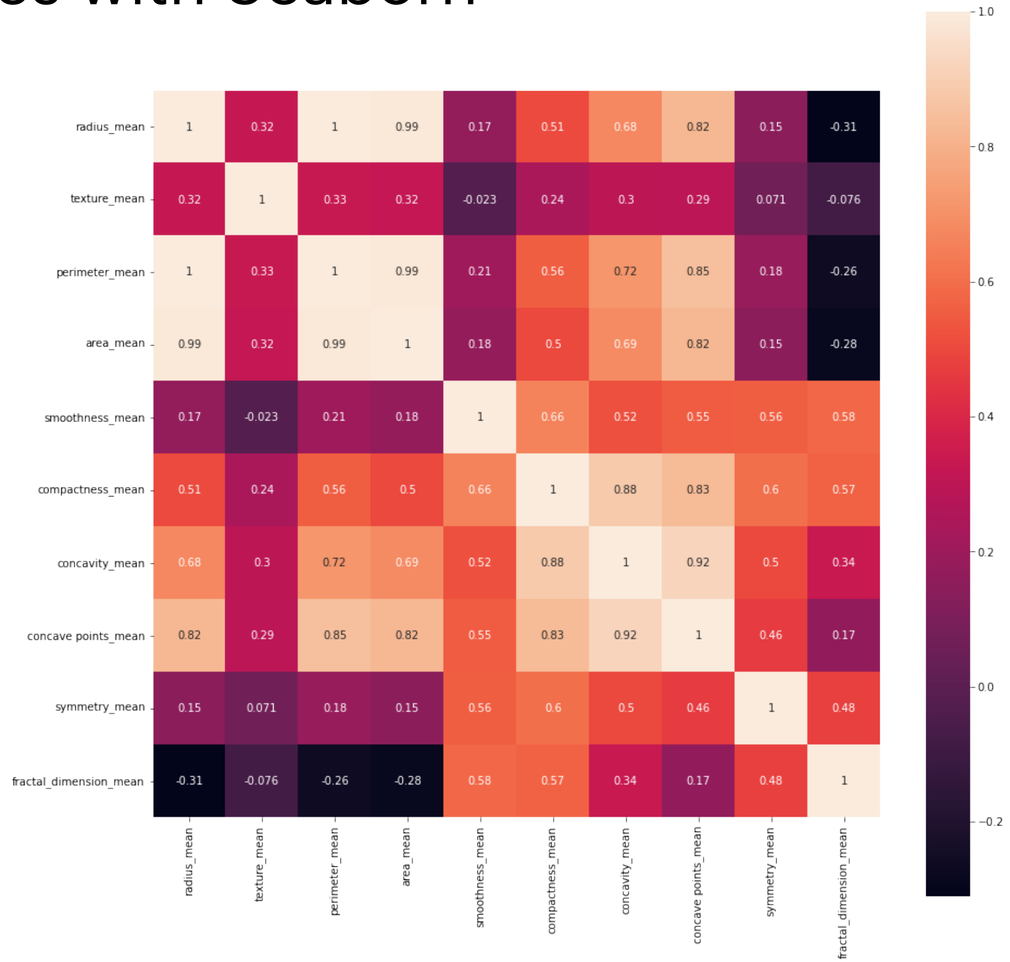
Extract and compute from digital images of diagnosis

2 Labels

Benign as 0 and Malignant as 1

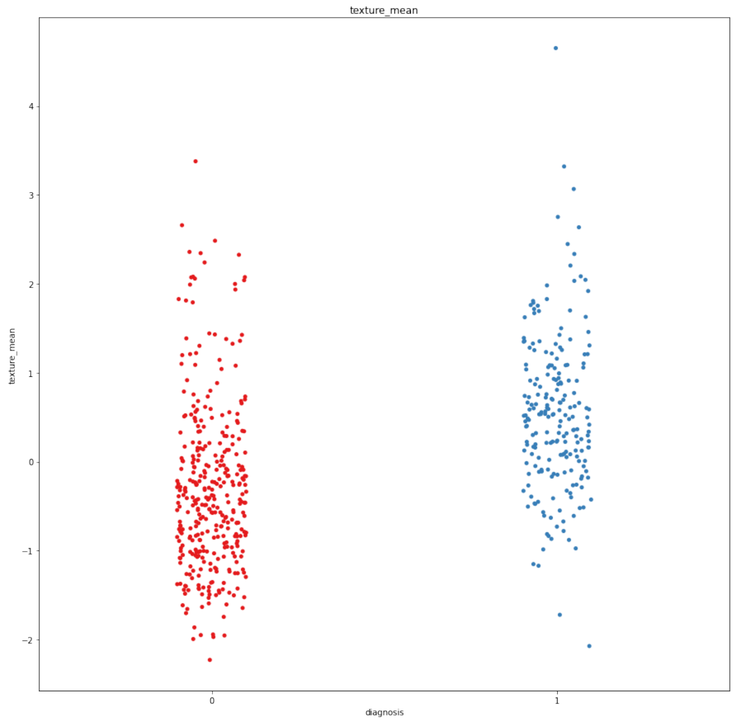
Data Visualization

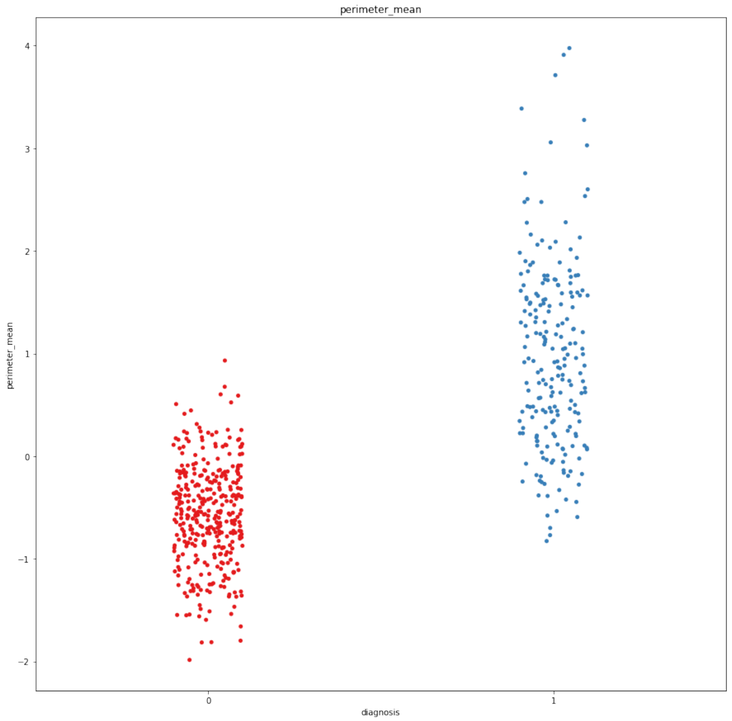
Heatmap of 10 features with Seaborn

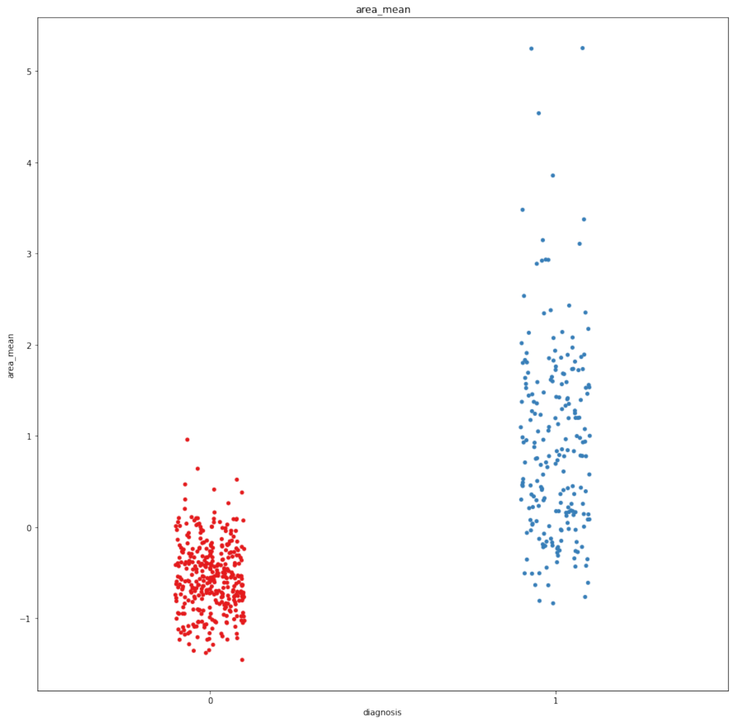


Data Visualization

Strip Plot Using Seaborn of 3 features with 2 classes

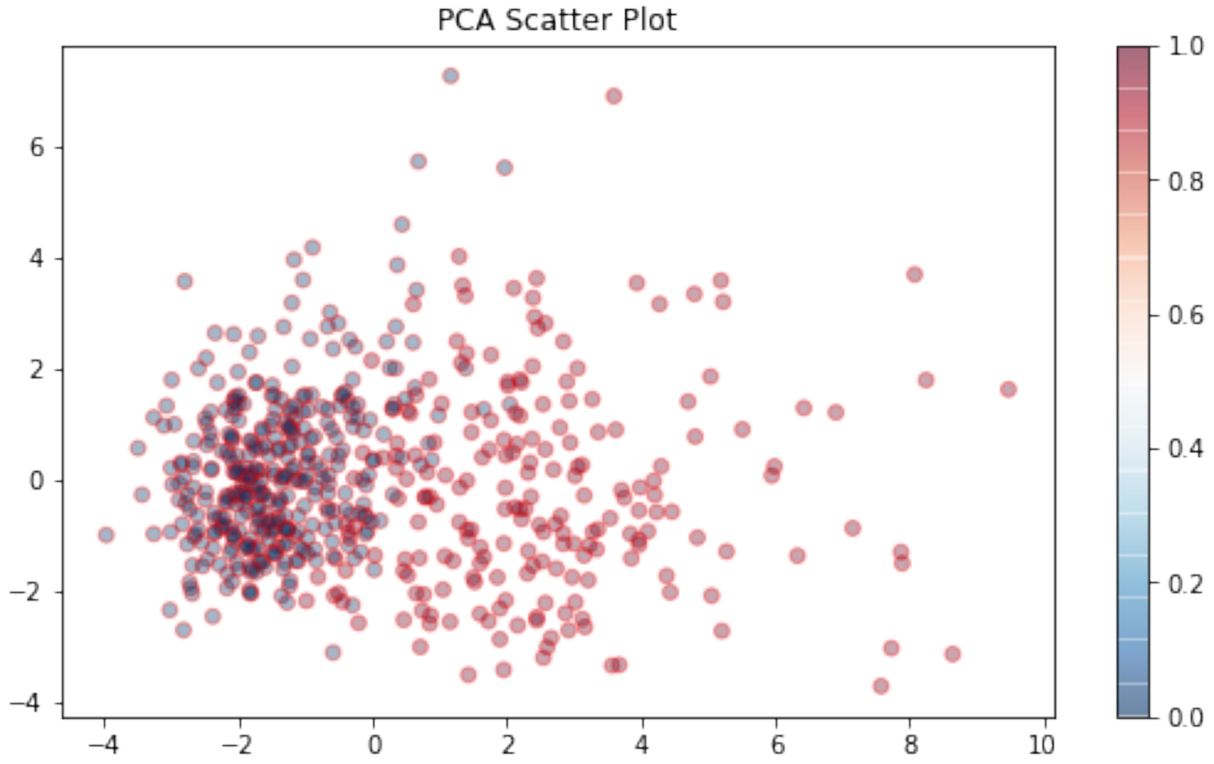






Data Visualization

Scatter Plot Using PCA by Selecting 2 Features to Represent on Graph



Machine Learning Concepts

Logistic Regression

Logistic Regression (LR) is a supervised learning model that forecasts **binary outcomes** using **logistic function** as the core element that can be as true or false, 1 or 0, Yes or No...

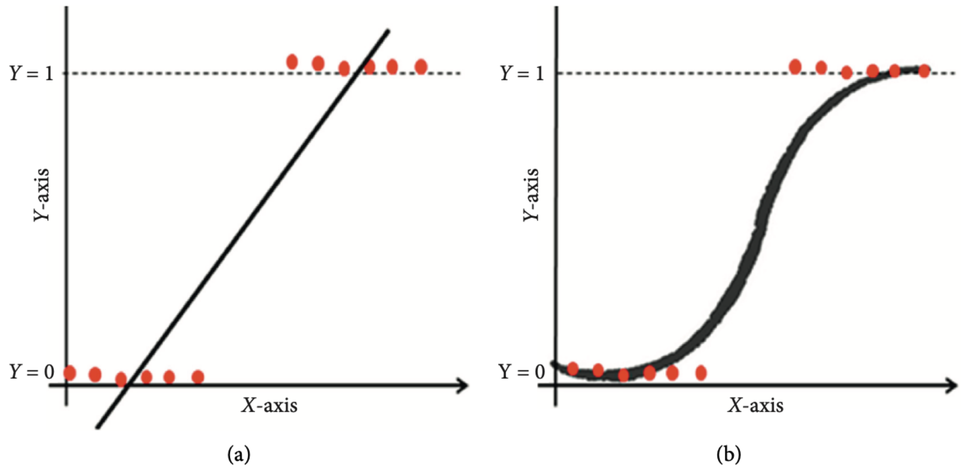


Figure: (a) Linear Regression and (b) Logistic Regression

Machine Learning Concepts

Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm that focuses on **hyperplane formation** that set a fair boundary for distinguishing the two different datasets.

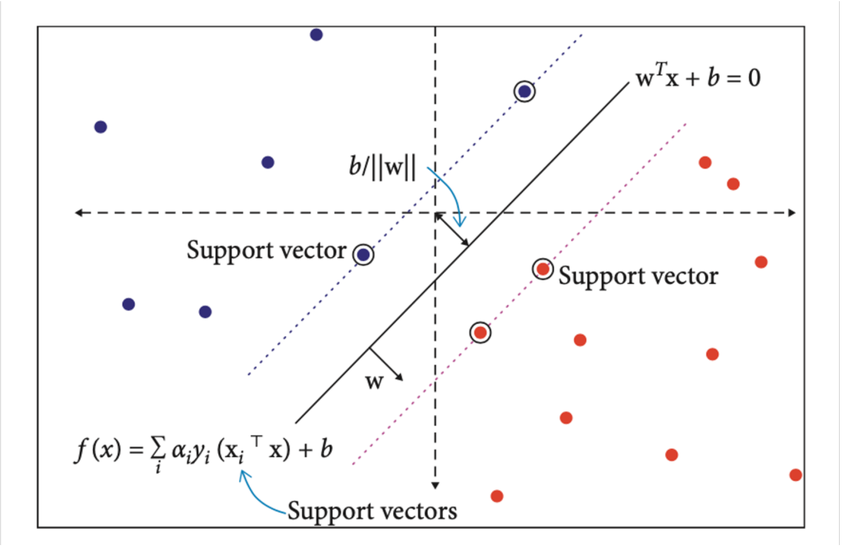


Figure: SVM Formula

Machine Learning Concepts

K-Nearest neighbor Classifier

K-nearest neighbour is a machine learning algorithm that **identifies the pattern** and explores the closest or nearest relations to **K**.

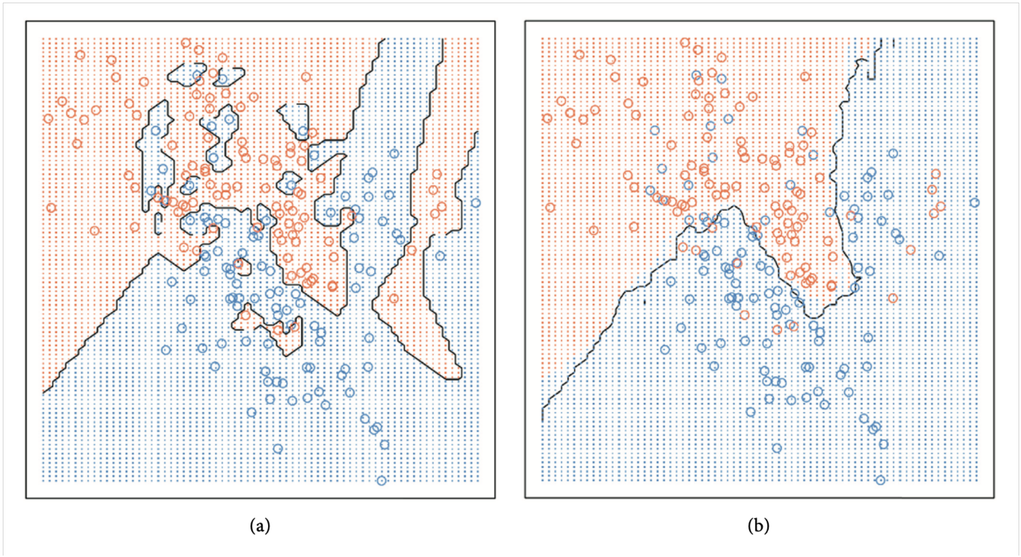


Figure: (a) K=1 and (b) K = 20

Machine Learning Concepts

Random Forest Classifier

Random Forest Classifier is a supervised learning technique follow the method to classify the large number of data with **decision tree algorithm** through the concept of **ensemble learning**.

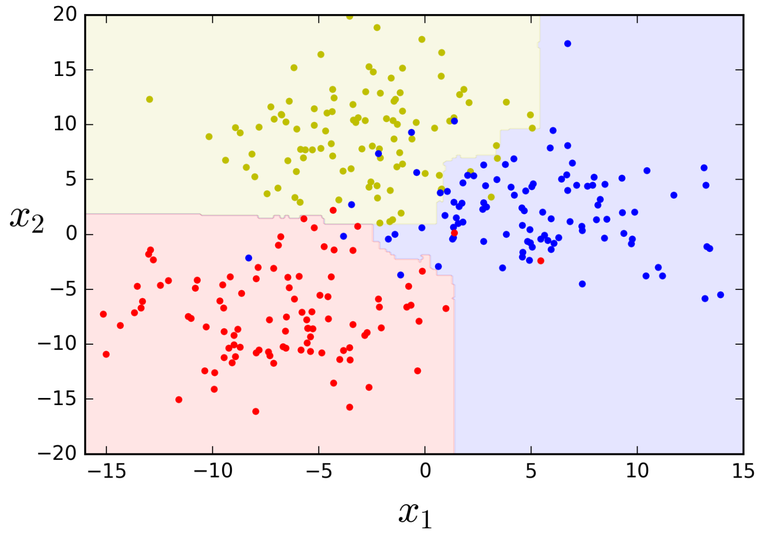


Figure: An Example of Random Forest Classifier use case

Feature Selection

Principal Component Analysis (PCA)

Principal Component Analysis is a method to reduce the dimensions follow the linear algebra for the computation.

It is useful for graph representation by reducing the parameters to only 2 parameters and feature selection use case.

Evaluation Matrices

Classification Report

Accuracy: The sum of all correct predictions divides by total predictions
Recall: Focus on Ratio of True Positive correctly classified to the total number of positive (TP + FN)
Precision: Focus on Ratio of True Positive correctly classified to the total number of classified positive samples (TP + FP)
F1-Score: Combine with Recall and Precision

Evaluation Matrices

Example of Classification Report

	precision	recall	f1-score	support
0	0.77	0.86	0.81	37584
1	0.84	0.75	0.79	37577
accuracy			0.80	75161
macro avg	0.81	0.80	0.80	75161
weighted avg	0.81	0.80	0.80	75161

Methodology

Preprocessing

Split data into 2 main sets: **Training Set for 80%** and **Testing Set for 20%**

Using **Standard Scaler** for scaling the features according to this formula:

$$z = \frac{x - \mu}{\sigma}$$

Methodology

Model 1: Logistic Regression
Model 2: Logistic Regression with PCA (Features = 2)
Model 3: K-Nearest Neighbor Classifier (K = 2)
Model 4: K-Nearest Neighbor Classifier (K = 3)
Model 5: Support Vector Machine (Kernel = rbf)
Model 6: Support Vector Machine (Kernel = linear)
Model 7: Random Forest Classifier
Model 8: Random Forest Classifier with PCA (Features = 2)

Result

Output with Testing Set

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
97.37%	92.98%	93.86%	94.74%	**98.25%**	96.46%	93.86%	92.98%

Support Vector Machine (Kernel = rbf)

Conclusion

- Cleaning Data is one of the most important steps to reach the maximum results from visualization to data prediction
- Understand on how to do preprocessing and visualization
- Successfully implemented the models and evaluation them
- Apply dimensionality reduction on the models and would be benificial in the long run dealing with large amounts of features
- Clearly identify the algorithms and how to use them
- Boost the creativity and customization to adjust the right use case

References

1. *Breast Cancer Wisconsin (Diagnostic) Data Set*. (n.d.). Retrieved July 6, 2022, from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
2. Alanazi, S. A., Kamruzzaman, M. M., Islam Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2021). Boosting Breast Cancer Detection Using Convolutional Neural Network. Journal of Healthcare Engineering, 2021, e5528622. https://doi.org/10.1155/2021/5528622
3. Brownlee, J. (2016, March 31). Logistic Regression for Machine Learning. *Machine Learning Mastery*. https://machinelearningmastery.com/logistic-regression-for-machine-learning/
4. Christopher, A. (2021, February 3). K-Nearest Neighbor. *The Startup*. https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4
5. Yiu, T. (2021, September 29). *Understanding Random Forest*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
6. Gandhi, R. (2018, July 5). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
7. *Data simple - Random Forests*. (n.d.). Retrieved July 12, 2022, from http://luisvalesilva.com/datasimple/random_forests.html

THANK YOU!