Machine Learning Approach for Early Breast Cancer Detection: Insights from Wisconsin Diagnostic Dataset, Coimbra Dataset and Mammographic Mass Dataset

Srinivas Bobba
sribob-2@student.ltu.se

Mini Project-D7041E, Applied Artificial Intelligence
Group Number: *MINI-PROJECT 19*

Agenda

- Introduction
- Problem Definition
- Data Exploration
- Data Pre-processing
- Methodology
- Implementation
- Experiments and Results
- Comparative Analysis
- Conclusion
- Recomendations

Introduction

- In 2020, there were around 2.26 million new cases of breast cancer among females. The tables below list the ten nations with the greatest rates of female breast cancer deaths and the largest number of female breast cancer fatalities in the same year.
- Putting even the best healthcare systems across the world under tremendous pressure.
- Regular mammography and other screening methods are crucial for early detection, leading to better treatment outcomes.
- In need of and well suited for early detection.
- The focus of this study is on utilizing machine learning approach to extract uncover features that contribute to predicting whether a tumor is malignant or benign.

Breast cancer rates

Rank	Country	Number	ASR/100,000
	World	*2,261,419*	*47.8*
1	Belgium	11,734	113.2
2	The Netherlands	15,725	100.9
3	Luxembourg	497	99.8
4	France	58,083	99.1
5	France, New Caledonia	185	99.0
6	Denmark	5,083	98.4
7	Australia	19,617	96.0
8	New Zealand	3,660	93.0
9	Finland	5,228	92.4
10	US	253,465	90.3

Breast cancer deaths

Rank	Country	Number	ASR/ 100,000
	World	*684,996*	*13.6*
1	Barbados	111	42.2
2	Fiji	184	41.0
3	Jamaica	637	34.1
4	Bahamas	80	31.0
5	Papua New Guinea	847	27.7
6	Somalia	1,189	27.2
7	Mali	1,425	26.6
8	Dominican Republic	1,577	26.4
9	Syria	1,946	26.2
10	Samoa	21	25.6

https://www.wcrf.org/cancer-trends/breast-cancer-statistics/

Introduction: Why and for whom

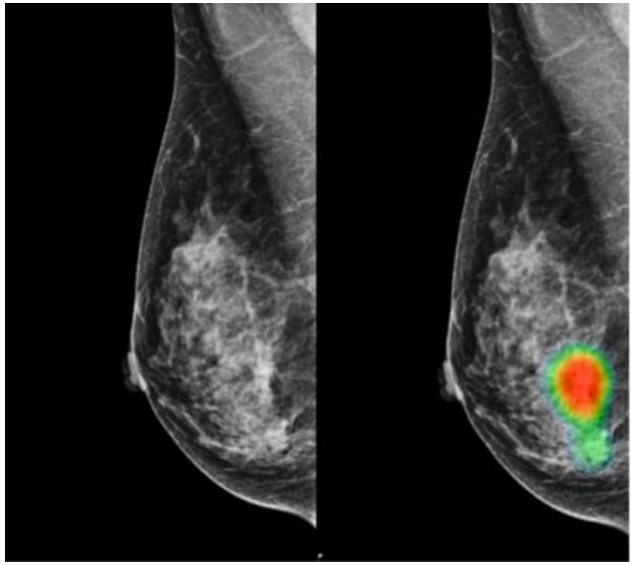
- Health care providers

  Early and correct diagnosis will:

  - Stop the spread of cancer to other parts of the body.
  - Reduce the need for health care
  - Significantly lower the health care costs

- Patients

  - Early and correct diagnosis can increase survival rate
  - Eliminating the risk of surgery



Problem Definition

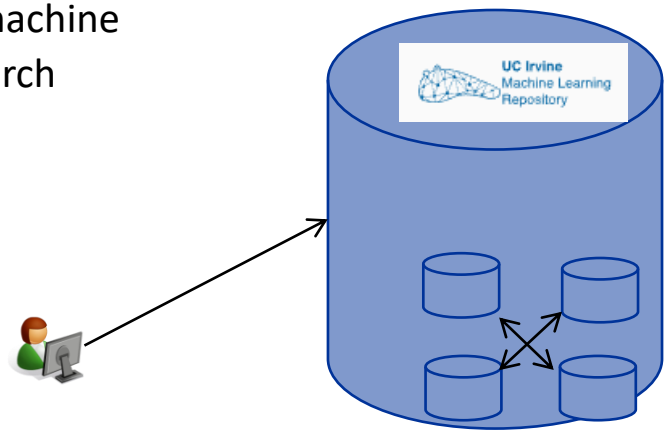
- The objective is to predict malignancy presence based on the features, with a focus on evaluating the performance of machine learning models in breast cancer detection utilizing Wisconsin (Diagnostic), Coimbra, and mammographic mass data.

- Machine learning approach is important here because it can analyze complex patterns in the datasets and helps create models that make breast cancer detection more accurate and efficient using smart algorithms.

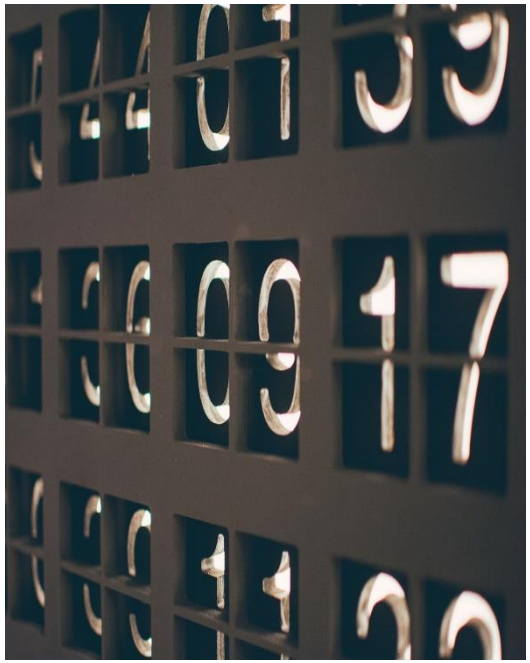
Introduction : Data Collection

The Datasets(Vectorized) below are collected from the UCI machine learning repository which is accessible to the public for research purposes.

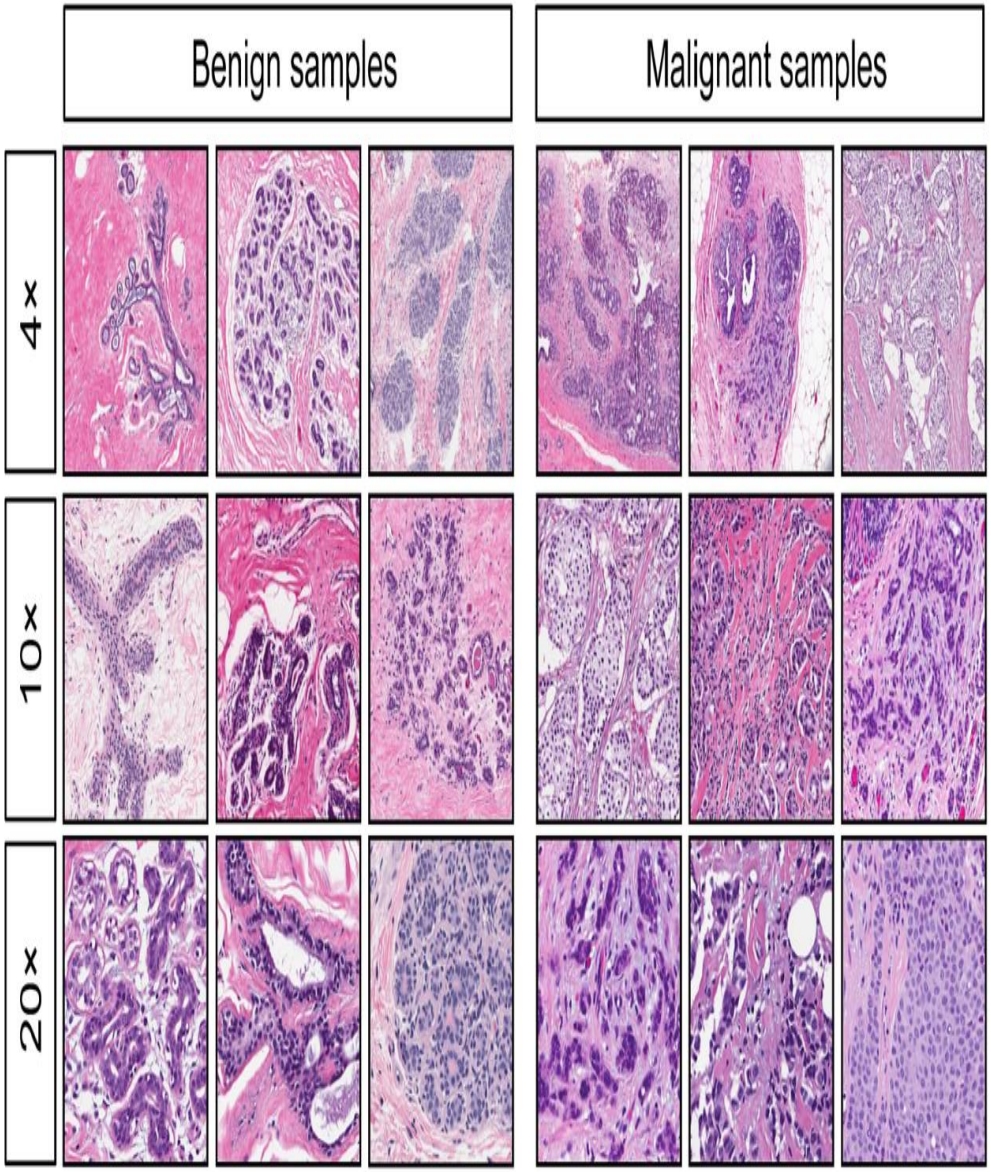
- Breast Cancer Wisconsin(Diagnostic) Dataset available at : https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- Breast Cancer Coimbra Dataset avalibale at : https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra
- Breast Cancer Mammographic Mass Dataset available at : https://archive.ics.uci.edu/dataset/161/mammographic+mass



Description of the Datasets
Breast cancer Wisconsin(Diagnostic) Dataset- It comprises features computed from digitized images of fine needle aspirates (FNA) of breast masses. Each instance in the dataset represents a diagnosed breast cancer case and includes various attributes such as mean, standard error, and worst values for key features like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The binary dependent variable indicates the presence or absence of breast cancer.
Breast Cancer Coimbra Dataset- It comprises clinical features collected from 64 patients with breast cancer and 52 healthy controls. It falls under the subject area of Health and Medicine and is tailored for classification tasks. The dataset includes 10 quantitative predictors, representing anthropometric data and parameters obtainable through routine blood analysis. The binary dependent variable indicates the presence or absence of breast cancer.
Breast Cancer Mammographic Mass Dataset – It pertains to mammographic mass lesions in breast cancer diagnosis. It includes features derived from digitized mammograms, such as shape and margin attributes, age of the patient, and BI-RADS assessment. The binary dependent variable indicates the presence or absence of breast cancer.



Experiment-1: Breast Cancer Prediction Analysis on Wisconsin Dataset



- Data Exploration and Pre-processing.
- Data Transformation and Feature Selection.
- Build the models using Supervised Machine Learning Algorithms.
- Evaluate model performance.
- Present the results.

Data Exploration(Wisconsin Dataset)

Breast Cancer Wisconsin (Diagnostic)

Donated on 10/31/1995

Diagnostic Wisconsin Breast Cancer Database.

Dataset Characteristics	Subject Area	Associated Tasks	
Multivariate	Health and Medicine	Classification	
Feature Type	**# Instances**	**# Features**	**Has Missing Values?**
Real	569	30	No

Data Retrieval From UCI Repository Using Url

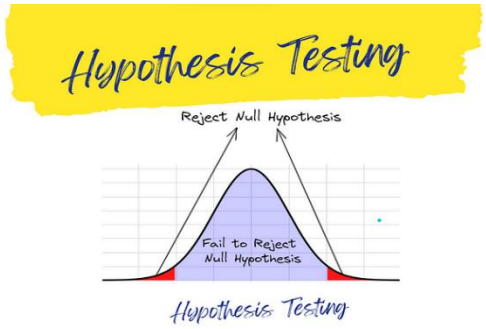
```
class BreastCancerData:
    def __init__(self, url):
        self.url = url
        self.column_names = ["id", "diagnosis", "mean_radius", "mean_texture", "mean_perimeter", "mean_area",
                "mean_smoothness", "mean_compactness", "mean_concavity", "mean_concave_points",
                "mean_symmetry", "mean_fractal_dimension", "se_radius", "se_texture", "se_perimeter",
                "se_area", "se_smoothness", "se_compactness", "se_concavity", "se_concave_points",
                "se_symmetry", "se_fractal_dimension", "worst_radius", "worst_texture", "worst_perimeter",
                "worst_area", "worst_smoothness", "worst_compactness", "worst_concavity", "worst_concave_points",
                "worst_symmetry", "worst_fractal_dimension"]

    def read_data(self):
        self.df = pd.read_csv(self.url, names=self.column_names)
        print(self.df)  # prints the dataframe

    def get_data_frame(self):
        return self.df

    def separate_data(self):
        self.benign_radius = self.df.loc[self.df["diagnosis"] == "B", "mean_radius"]
        self.malignant_radius = self.df.loc[self.df["diagnosis"] == "M", "mean_radius"]


data = BreastCancerData(url='https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data')
data.read_data()
BC_data = data.get_data_frame()
```

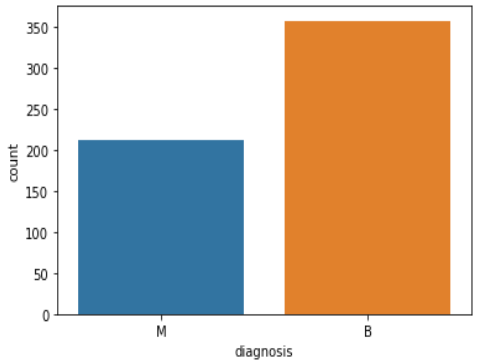



```
[569 rows x 32 columns]
Reject the null hypothesis. The mean radius of benign tumors is less than the mean radius of malignant tumors.
Reject the null hypothesis. The mean radius of benign tumors is less than the mean radius of malignant tumors.
Test results:
------------------------------
t-statistic: -22.208797758464524
p-value: 1.6844591259582747e-64
critical value: 1.6475454734678237
degree of freedom: 567
confidence level: 0.95
alpha: 0.05
```

```
# Display information about the da
BC_data.isnull().sum()
```

```
id                         0
diagnosis                  0
mean_radius                0
mean_texture               0
mean_perimeter             0
mean_area                  0
mean_smoothness            0
mean_compactness           0
mean_concavity             0
mean_concave_points        0
mean_symmetry              0
mean_fractal_dimension     0
se_radius                  0
se_texture                 0
se_perimeter               0
se_area                    0
se_smoothness              0
se_compactness             0
se_concavity               0
se_concave_points          0
se_symmetry                0
se_fractal_dimension       0
worst_radius               0
worst_texture              0
worst_perimeter            0
worst_area                 0
worst_smoothness           0
worst_compactness          0
worst_concavity            0
worst_concave_points       0
worst_symmetry             0
worst_fractal_dimension    0
dtype: int64
```

Class Count




```
B    357
M    212
Name: diagnosis, dtype: int64
```

Data Pre-processing(Wisconsin Dataset)

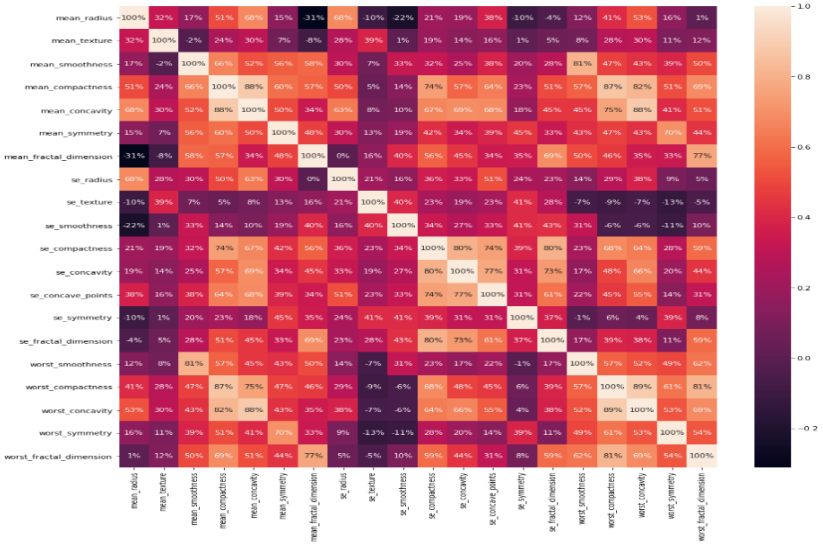
1. **Feature Grouping:**
   - Grouped features into 'worst,' 'mean,' and 'se' categories.

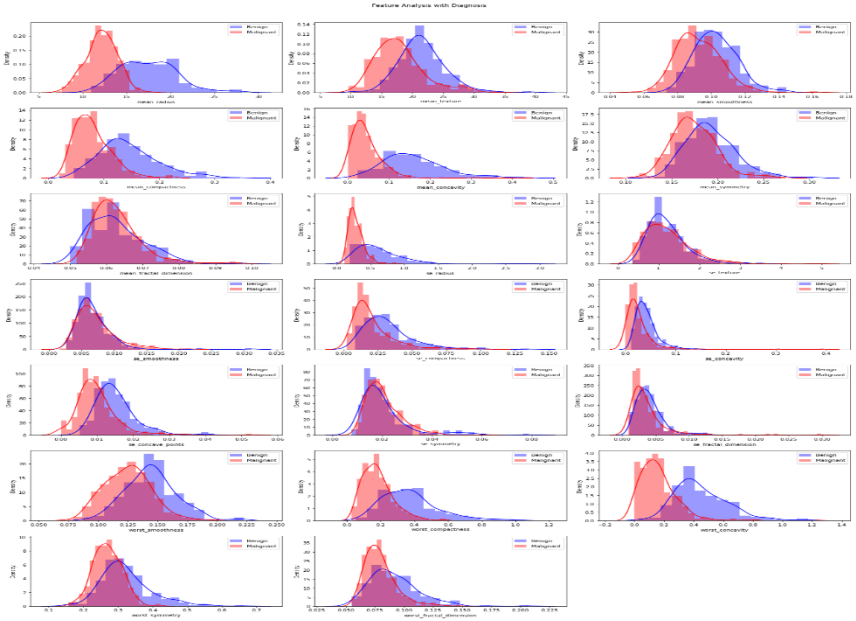
2. **Diagnostic Result Visualization**:
   - Plotted histograms for each feature group by diagnostic result.

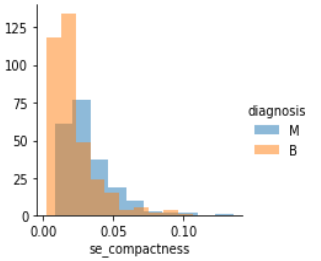
3. **Correlation Analysis:**
   - Calculated correlation matrix for all features (excluding 'id' and 'diagnosis').

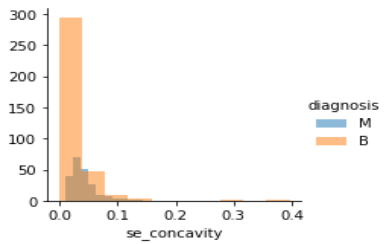
4. **Feature Filtering:**
   - Removed highly correlated features (correlation >= 0.9).

5. **Final Feature Selection:**
   - Selected a subset of 20 features for further analysis.









Data Transformation (Wisconsin Dataset)

1. Standardization

- Scaled numerical features using `StandardScaler`.

2. PCA Dimensionality Reduction:

- Applied PCA to reduce dimensions to 10 components.
- Selected features explaining over 95% of variance.

3. Visualization:

- Visualized data distribution and confirmed linearity.

4. Feature Selection:

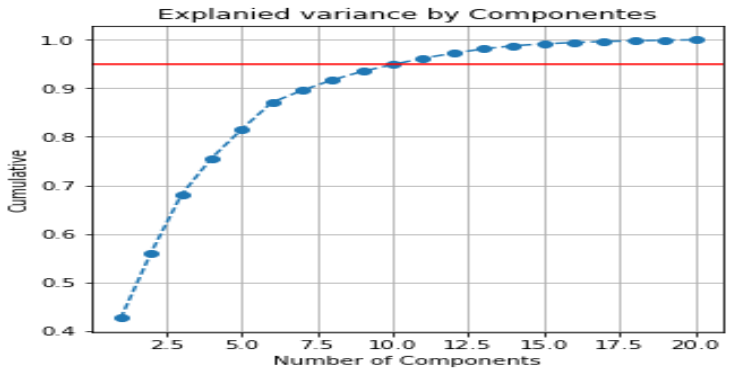
- Identified most important features in each principal component.

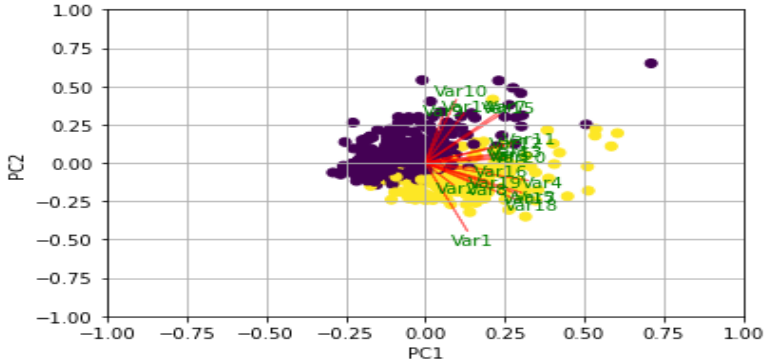
5. New Dataset:

- Created a 569x11 dataset for final analysis.

6. Resulting Data:

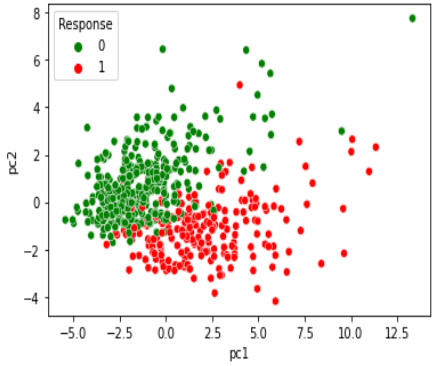
- 10 principal components and a 'Response' column.
- 'Response' indicates cancer (1) or non-cancer (0).





Breast Cancer Wisconsin (Diagnostic)

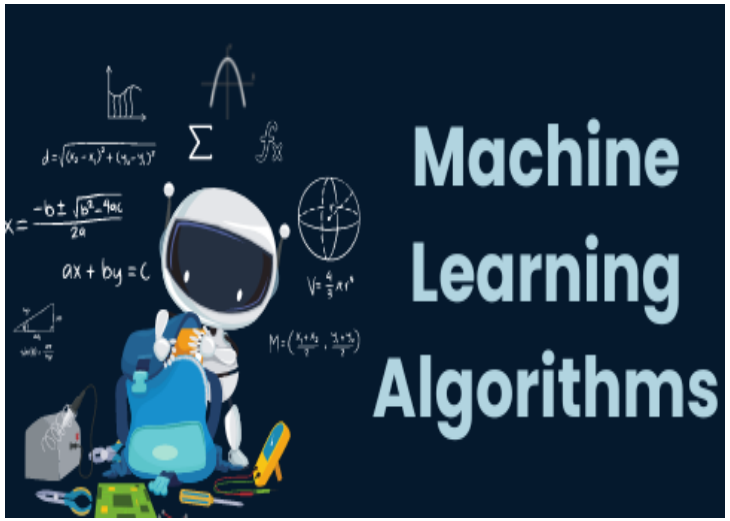
	0	1
0	PC0	compactness_mean
1	PC1	radius_mean
2	PC2	radius_se
3	PC3	symmetry_mean
4	PC4	symmetry_worst
5	PC5	texture_mean
6	PC6	smoothness_se
7	PC7	texture_se
8	PC8	symmetry_mean
9	PC9	concavity_se



Methodology

Building Models using supervised Machine learning Algorithms

- Logistic Regression: Linear model for binary classification.
- Decision Trees: Non-linear model for classification and regression.
- Random Forest: Ensemble learning method using multiple decision trees.
- k-Nearest Neighbors (KNN): Instance-based learning algorithm.
- Support Vector Machines (SVM): Linear and non-linear classification algorithm.
- Multi-Layer Perceptron (MLP):Feedforward neural network with multiple layers.
- Artificial Neural Networks (ANN):General term for interconnected nodes arranged in layers.
- Convolutional Neural Networks (CNN):Specialized neural network architecture for processing grid-like data.

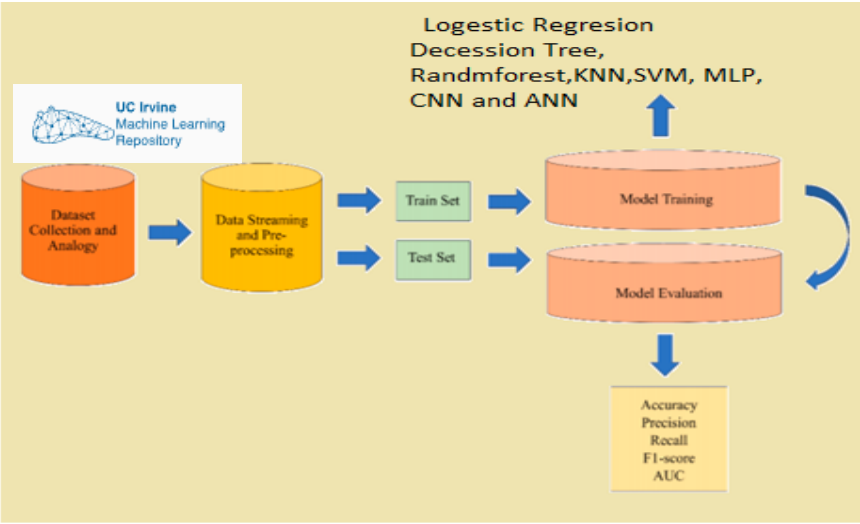


Model Evaluation

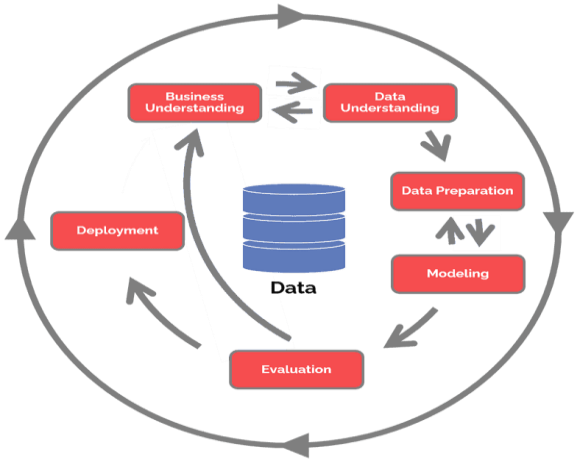
The Performance metrics are chosen for our analysis because due to imbalanced datasets

1. **Sensitivity (Recall):** True positive rate.
2. **Specificity:** True negative rate.
3. **F1 Score:** Harmonic mean of precision and recall.
4. **Accuracy:** Overall correctness.
5. **AUC (Area Under the Curve):** Discrimination ability.

Implementation



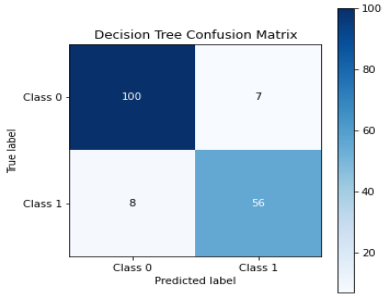
Breast Cancer Prediction Process Model

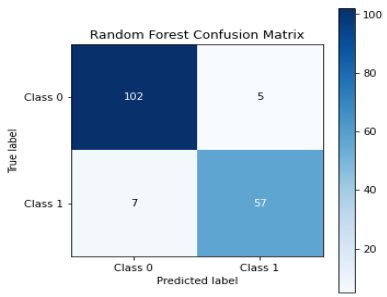


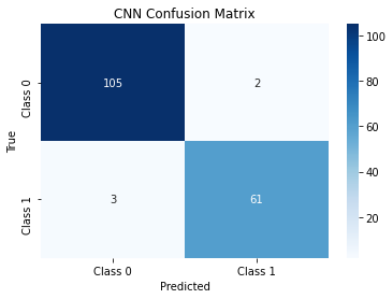
The CRoss Industry Standard Process for Data Mining (**CRISP-DM**)

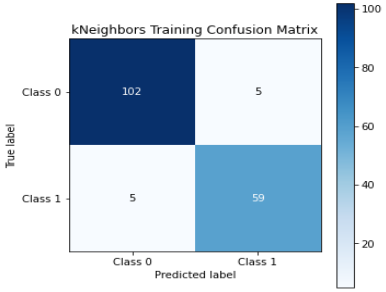
Motivation: https://www.datascience-pm.com/crisp-dm-2/

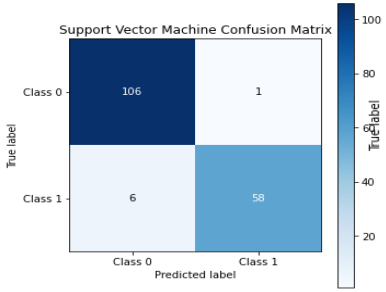
Results(Wisconsin Diagnostic Dataset)

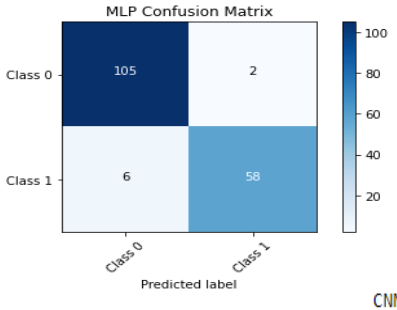


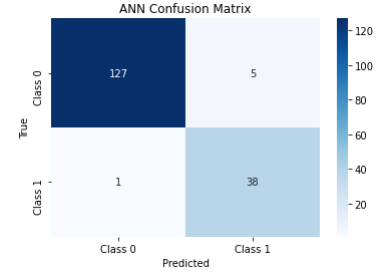













```
Logistic Regression metrics:
AUC: 0.959
Recall: 0.938
Specificity: 0.981

Decision Tree metrics:
AUC: 0.888
Recall: 0.859
Specificity: 0.916

Random Forest metrics:
AUC: 0.922
Recall: 0.891
Specificity: 0.953

kNeighbors Training metrics:
AUC: 0.939
Recall: 0.906
Specificity: 0.972

Support Vector Machine metrics:
AUC: 0.975
Recall: 0.969
Specificity: 0.981
```

MLP Accuracy: 0.9707602339181286
Classification Report:

	precision	recall	f1-score	support
0	0.97	0.98	0.98	107
1	0.97	0.95	0.96	64
accuracy			0.97	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.97	0.97	0.97	171

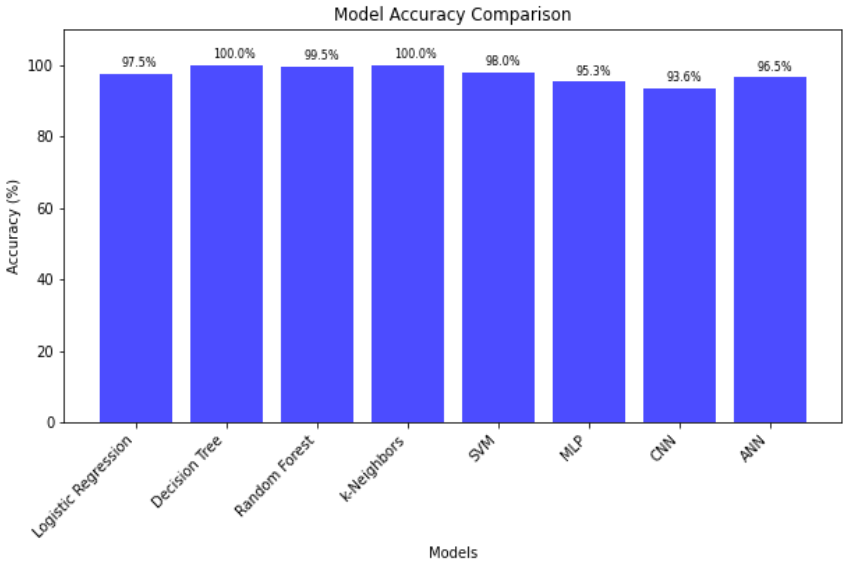
CNN Accuracy: 0.9707602339181286
Classification Report:

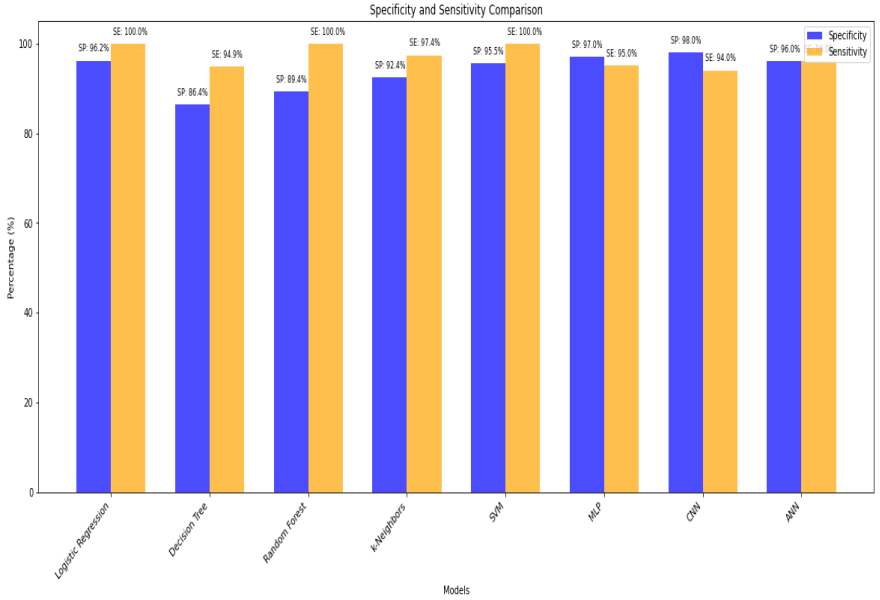
	precision	recall	f1-score	support
0	0.97	0.98	0.98	107
1	0.97	0.95	0.96	64
accuracy			0.97	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.97	0.97	0.97	171

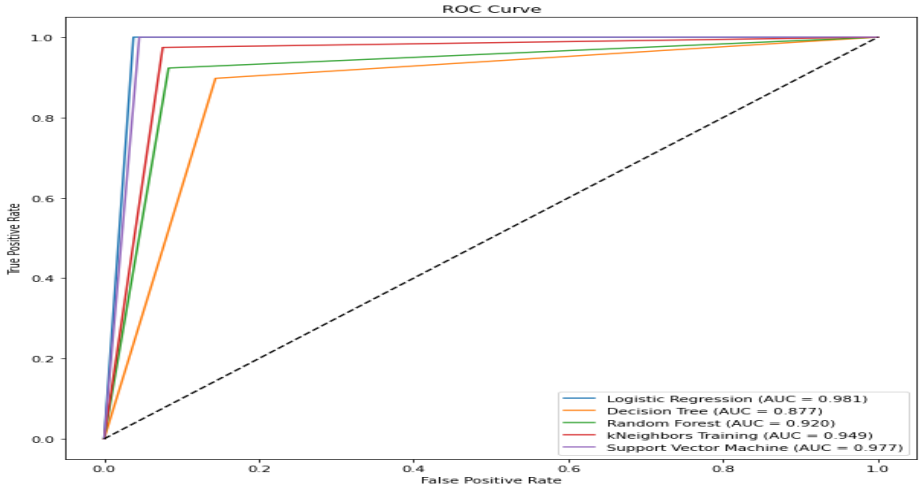
ANN Accuracy: 96.49%
ANN Classification Report:

	precision	recall	f1-score	support
0	0.99	0.96	0.98	132
1	0.88	0.97	0.93	39
accuracy			0.96	171
macro avg	0.94	0.97	0.95	171
weighted avg	0.97	0.96	0.97	171

Interpretation of the Results(Wisconsin Diagnostic Dataset)







- If our primary concern is accuracy, Decision Tree, Random Forest, k-Neighbors, and Support Vector Machine all achieved perfect training accuracy. However, training accuracy alone might not represent the model's generalization to new data.
- Additionally, it's crucial to consider metrics like recall, precision, and specificity, especially in medical applications where false negatives (missing cancer cases) can be critical.
- **Based on this analysis, the Logistic regression, Support Vector Machine and k-Neighbors models seem to be strong contenders for cancer prediction.**

Experiment-2: Breast Cancer Prediction Analysis on Coimbra Dataset

Breast Cancer Coimbra Dataset

University Hospital Centre of Coimbra

2018

- Data Exploration and Pre-processing.
- Data Transformation and Feature Selection.
- Build the models using Supervised Machine Learning Algorithms.
- Evaluate model performance.
- Present the results.

Data Exploration(Coimbra Dataset)

Breast Cancer Coimbra

Donated on 3/5/2018

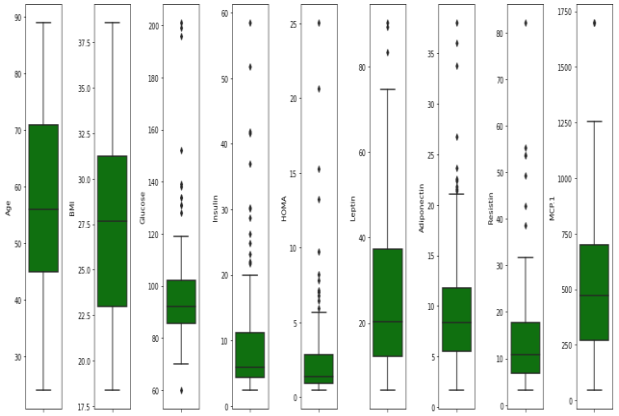
Clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Health and Medicine	Classification
Feature Type	**# Instances**	**# Features**
Integer	116	10

```
: coimbra_data.isnull().sum()

: Age             0
  BMI             0
  Glucose         0
  Insulin         0
  HOMA            0
  Leptin          0
  Adiponectin     0
  Resistin        0
  MCP.1           0
  Classification  0
  dtype: int64
```

Outliers



Data Retrieval From UCI Repository and Data Exploration

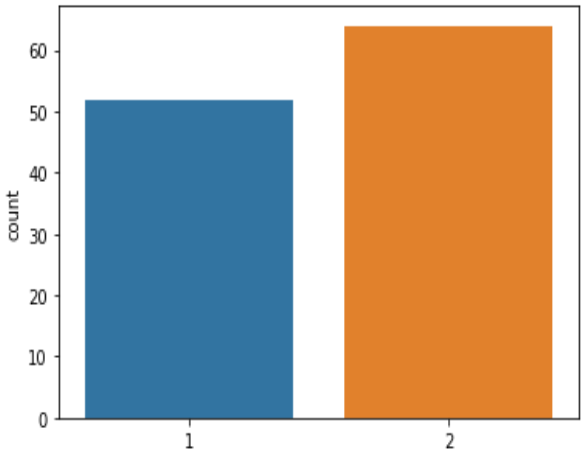
```
coimbra_data = pd.read_csv("Coimbra_dataR2.csv")
X = coimbra_data.drop(['Classification'],axis=1)
Y = coimbra_data['Classification']

coimbra_data.head()
```

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1

116 rows × 10 columns

Class Count




```
2    64
1    52
Name: Classification, dtype: int64
```

Data Pre-processing(Coimbra Dataset)

1. Feature Identification:

- Originally, the classes were represented as 1 and 2. we converted with a more conventional representation, where 0 often signifies one class (e.g., benign), and 1 signifies another class (e.g., malignant).

2. Diagnostic Result Visualization:

- Plotted histograms for each feature group by diagnostic result.

3. Correlation Analysis:

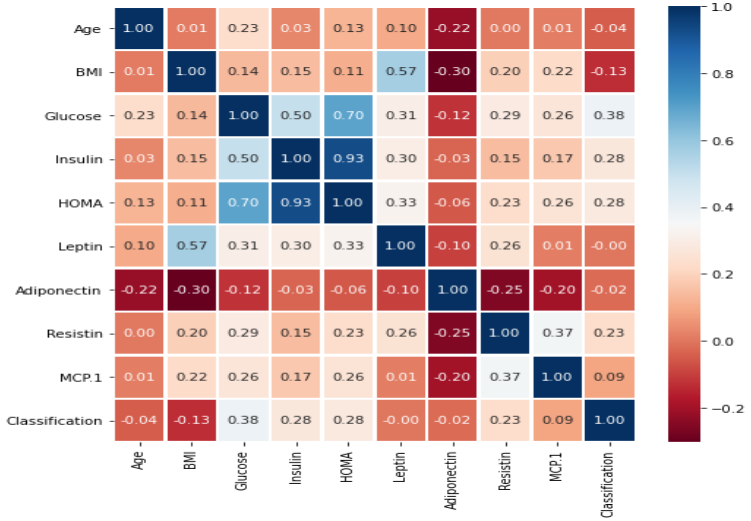
- Calculated correlation matrix for all features (excluding 'id' and 'diagnosis').

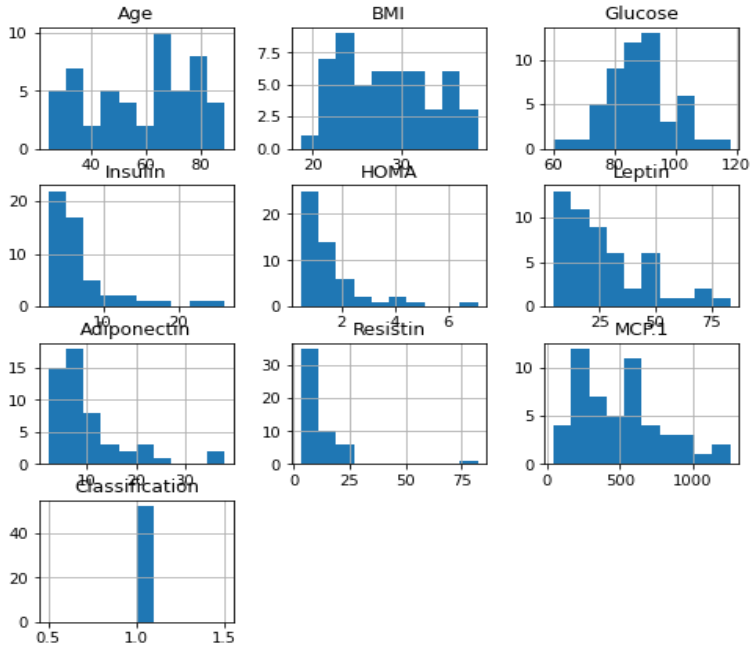
4. Feature Filtering:

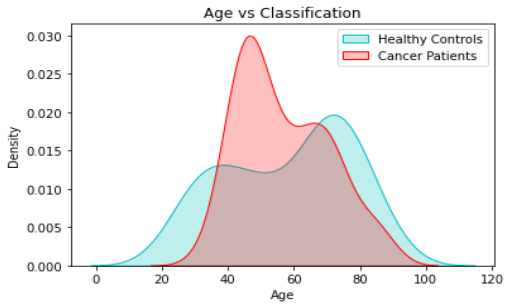
- Removed highly correlated features (correlation >= 0.9).

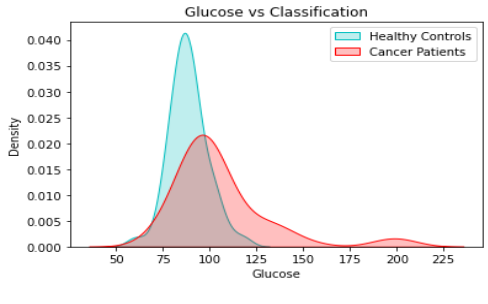
5. Final Feature Selection:

- Selected all 10 features for further analysis.









Data Transformation (Coimbra Dataset)

1. Standardization

- Scaled numerical features using `StandardScaler`.

2. PCA Dimensionality Reduction:

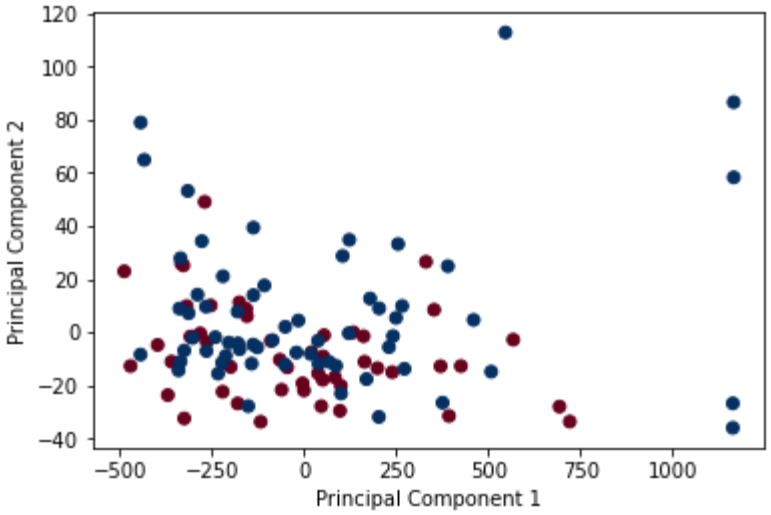
- Applied PCA to reduce dimensions. No apparent Clustering.

3. Visualization:

- Visualized data distribution and confirmed linearity.

4. Feature Selection:

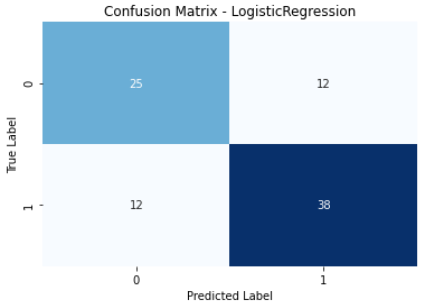
- Identified most important features(Variable Importance) using Extra tree classifier.

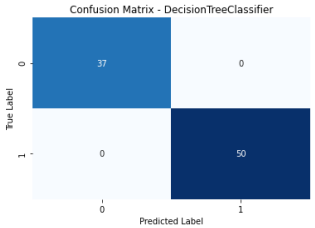


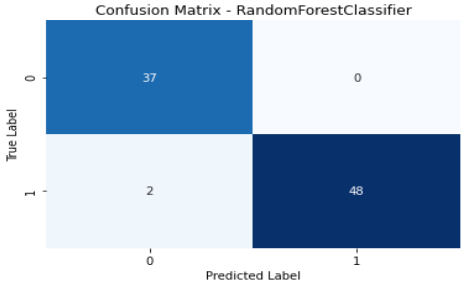
Breast Cancer Coimbra

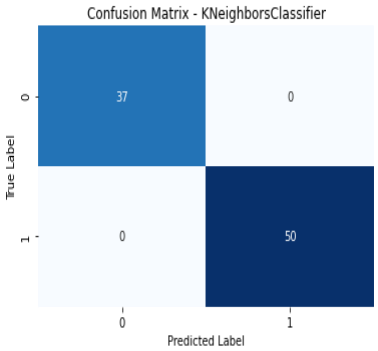
Glucose	Importance: 0.41
Age	Importance: 0.2
Resistin	Importance: 0.16
BMI	Importance: 0.07
Insulin	Importance: 0.05
HOMA	Importance: 0.05
Leptin	Importance: 0.05
MCP.1	Importance: 0.02
Adiponectin	Importance: 0.0

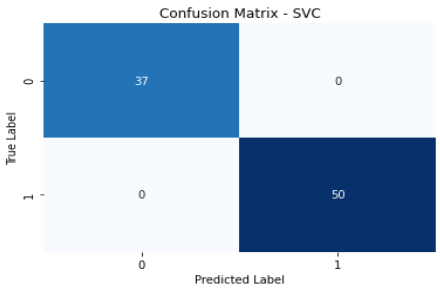
Results(Coimbra Dataset)

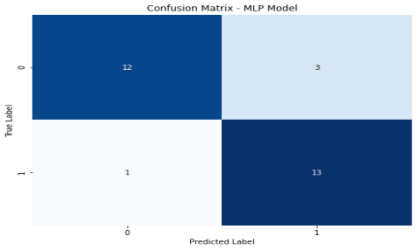


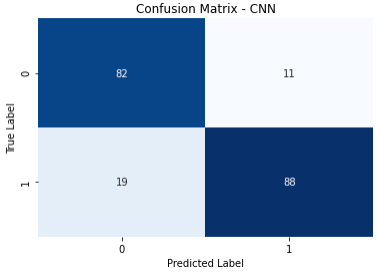


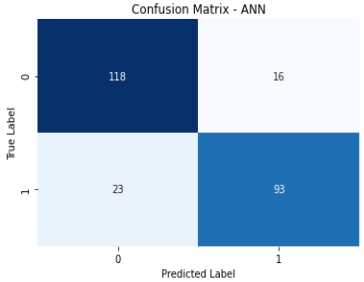












MLP Accuracy: 0.8620689655172413
Classification Report:

	precision	recall	f1-score	support
0	0.92	0.80	0.86	15
1	0.81	0.93	0.87	14
accuracy			0.86	29
macro avg	0.87	0.86	0.86	29
weighted avg	0.87	0.86	0.86	29

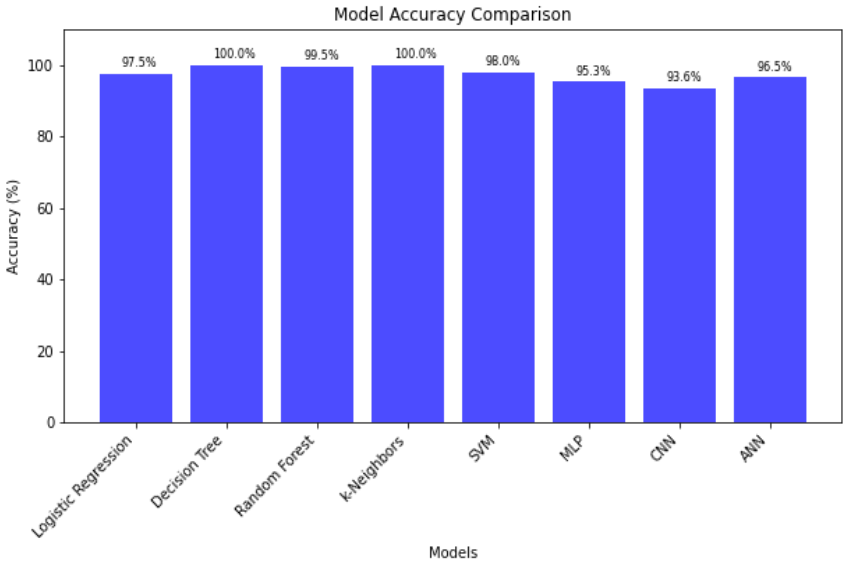
CNN Accuracy: 0.85
Classification Report:

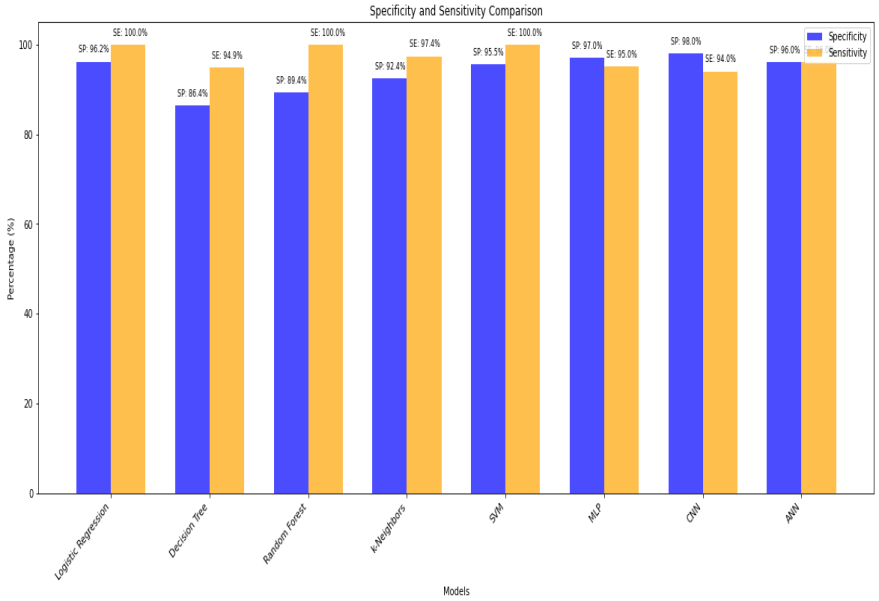
	precision	recall	f1-score	support
0	0.81	0.88	0.85	93
1	0.89	0.82	0.85	107
accuracy			0.85	200
macro avg	0.85	0.85	0.85	200
weighted avg	0.85	0.85	0.85	200

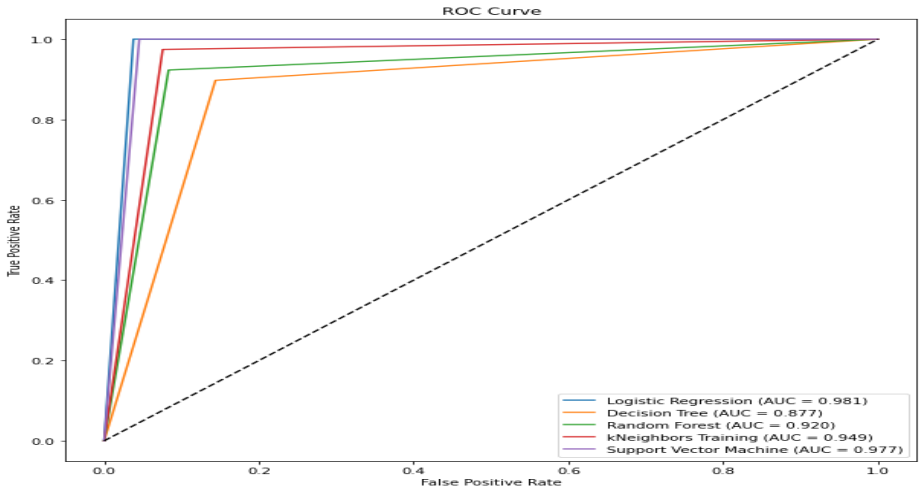
ANN Classification Report:

	precision	recall	f1-score	support
0	0.85	0.90	0.88	134
1	0.88	0.82	0.85	116
accuracy			0.86	250
macro avg	0.87	0.86	0.86	250
weighted avg	0.86	0.86	0.86	250

Interpretation of the Results(Coimbra Dataset)



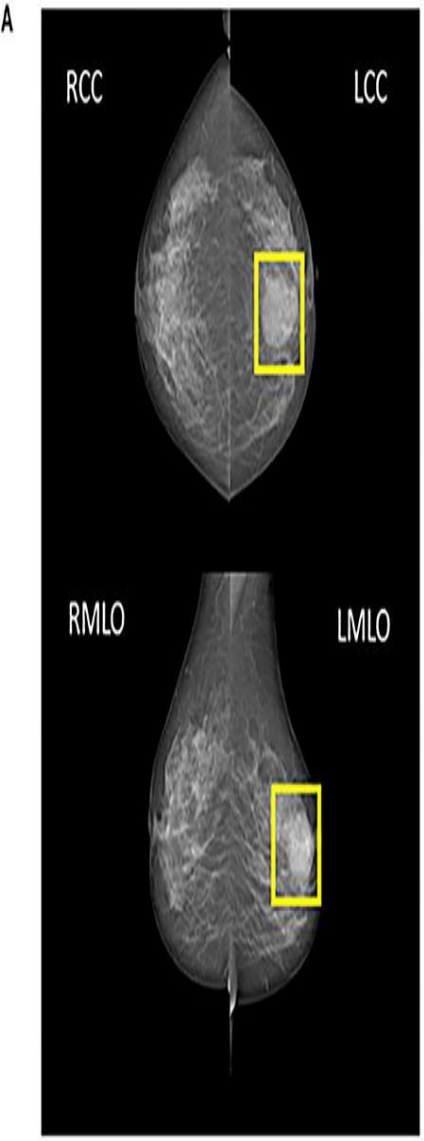


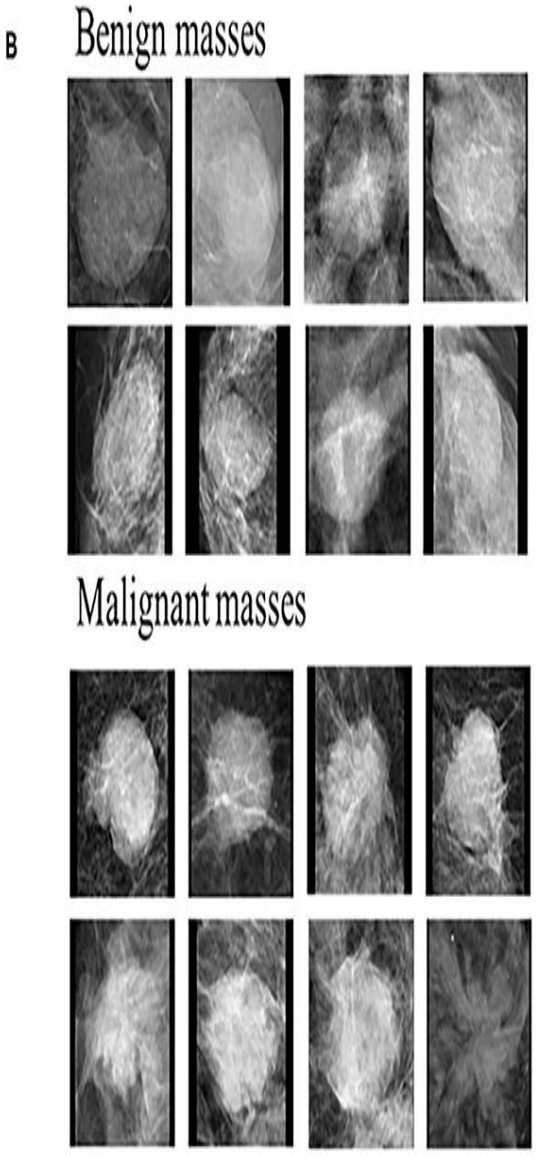


Based on the above results, We observed that

- Decision Tree and k-Nearest Neighbors models exhibit potential overfitting due to perfect training accuracy.
- Random Forest and Support Vector Machine models show good generalization.
- Neural network models, especially the CNN, demonstrate competitive performance, outperforming the ANN.

Experiment-3: Breast Cancer Prediction Analysis on Mammographic Mass





- Data Exploration and Pre-processing.
- Data Transformation and Feature Selection.
- Build the models using Supervised Machine Learning Algorithms.
- Evaluate model performance.
- Present the results.

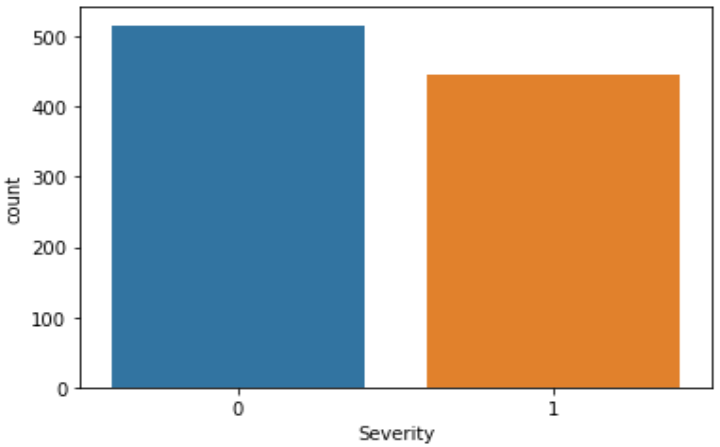
Data Exploration(Mammographic Mass Dataset)

Mammographic Mass

Donated on 10/28/2007

Discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Health and Medicine	Classification
Feature Type	**# Instances**	**# Features**
Integer	961	6




```
0    516
1    445
Name: Severity, dtype: int64
```

Data Retrieval From UCI Repository and Data Exploration

```
# Read in dataset
url = "http://archive.ics.uci.edu/ml/machine-learning-databases/mammographic-masses/mammographic_masses.data"
Mamm_data = pd.read_csv(url, header=None)
```

```
# Name the columns of the dataset
Mamm_data .columns = ["BI-RADS", "Age", "Shape", "Margin", "Density", "Severity"]

# Get the number of observations and attributes of the Mamm dataset
print('The shape of the dataset is: ', Mamm_data.shape)

# Get the datatypes of each feature in the Mamm dataset
Mamm_data.dtypes
```

	BI-RADS	Age	Shape	Margin	Density	Severity
0	5	67	3	5	3	1
1	4	43	1	1	?	1
2	5	58	4	5	3	1
3	4	28	1	1	3	0
4	5	74	1	5	?	1
...	...	...	...	...	...	...
956	4	47	2	1	3	0
957	4	56	4	5	3	1
958	4	64	4	5	3	0
959	5	66	4	5	3	1
960	4	62	3	3	3	0

961 rows × 6 columns

Data Pre-processing(Mammographic Mass Dataset)

1. Impute Function:

- Converts a column to numeric, replacing non-numeric values with the median.

2. Replace Function:

- Replaces outliers in a numeric column with an upper limit.

3. ZNorm Function:

- Z-normalizes numeric values in a column.

4. Decode Function:

- Decodes numeric values into categorical names.

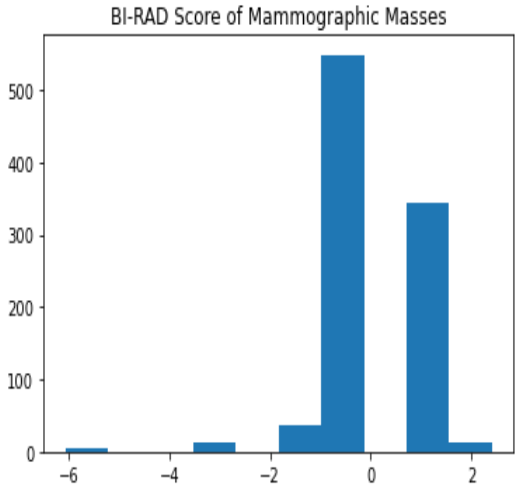
5. Consolidate Function:

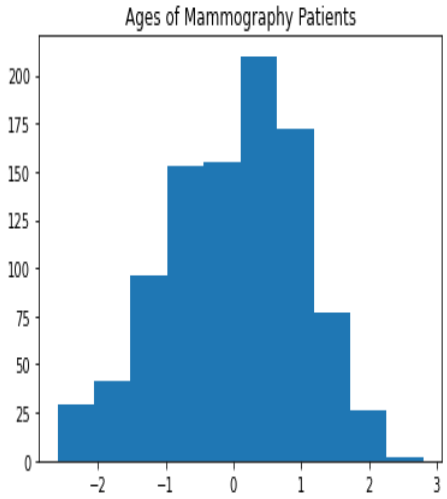
- Consolidates categorical variables in a given column.

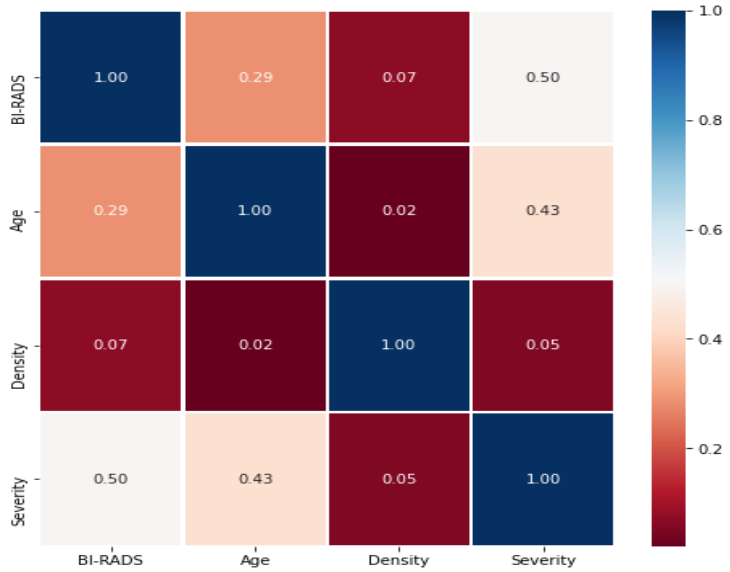
6. OneHotEncode Function:

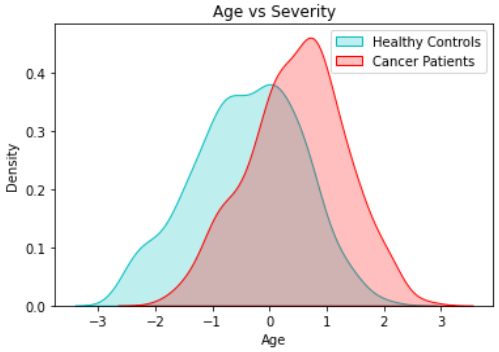
- One-hot encodes categorical data, creating binary values in a new column based on a specific categorical variable.

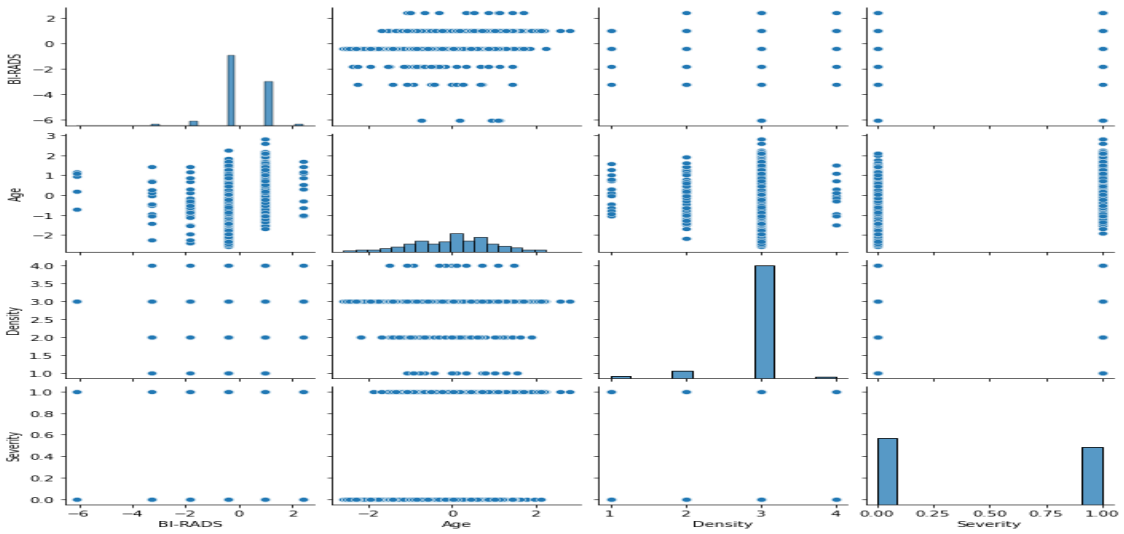
Data Visualization(Mammographic Mass Dataset)

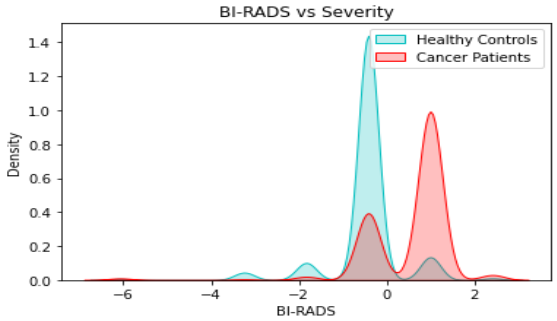












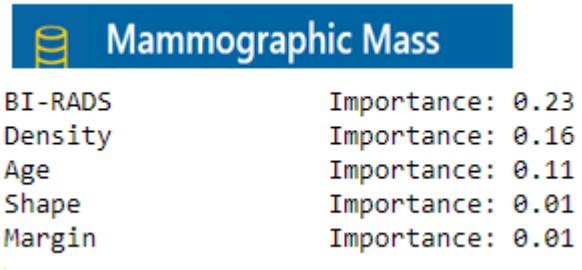
Data Transformation (Mammographic Mass Dataset)

1.Dataset Splitting:

- Utilized `train_test_split` from `sklearn.model_selection` to split data into training and test sets (75% training, 25% test).

2.Feature Preprocessing:

- Used `SimpleImputer` for missing values (mean for numeric, most frequent for categorical).
- Applied `StandardScaler` for numeric features to standardize them.



3. Column Transformation:

- Used `ColumnTransformer` to independently preprocess numeric and categorical features.

4.Pipeline for Transformation:

- Constructed pipelines for numeric and categorical transformations.

5. One-Hot Encoding:

- Applied `OneHotEncoder` to handle categorical variables, ignoring unknown values.

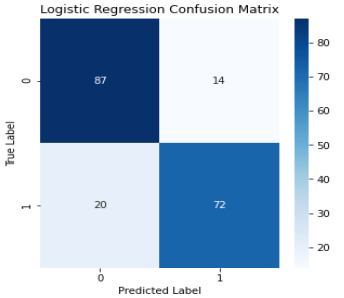
6. Feature Scaling:

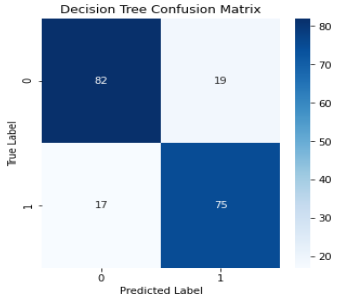
- Standardized features using `StandardScaler` independently for training and test sets.

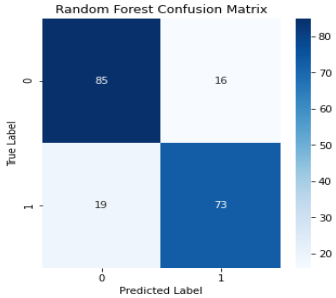
7. Feature Selection:

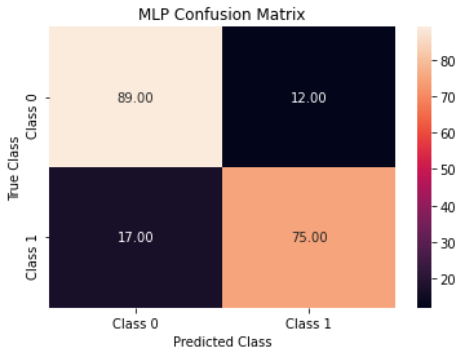
- Identified most important features(Variable Importance) using Extra tree classifier.

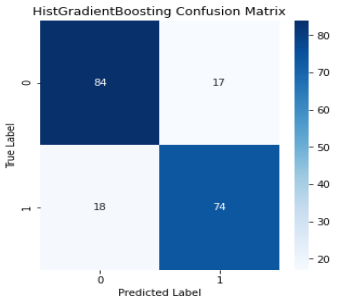
Results(Mammographic Mass Dataset)

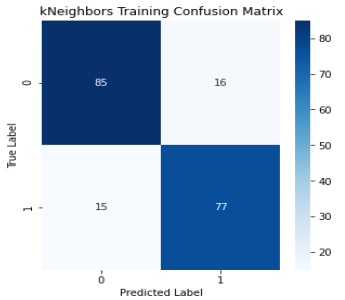


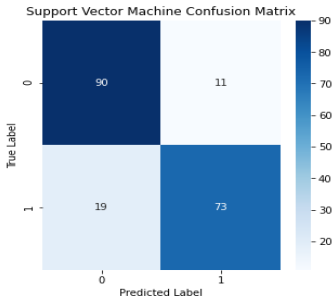


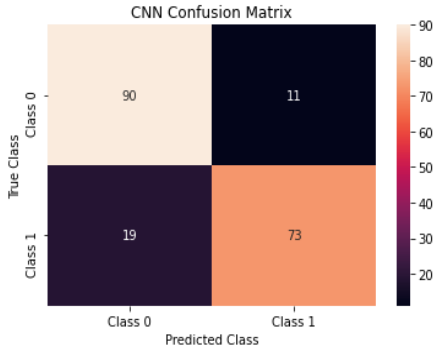


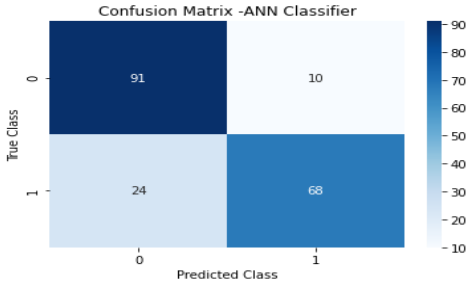












MLP Accuracy: 0.8031088082901554
Classification Report:

	precision	recall	f1-score	support
0	0.80	0.83	0.82	101
1	0.81	0.77	0.79	92
accuracy			0.80	193
macro avg	0.80	0.80	0.80	193
weighted avg	0.80	0.80	0.80	193

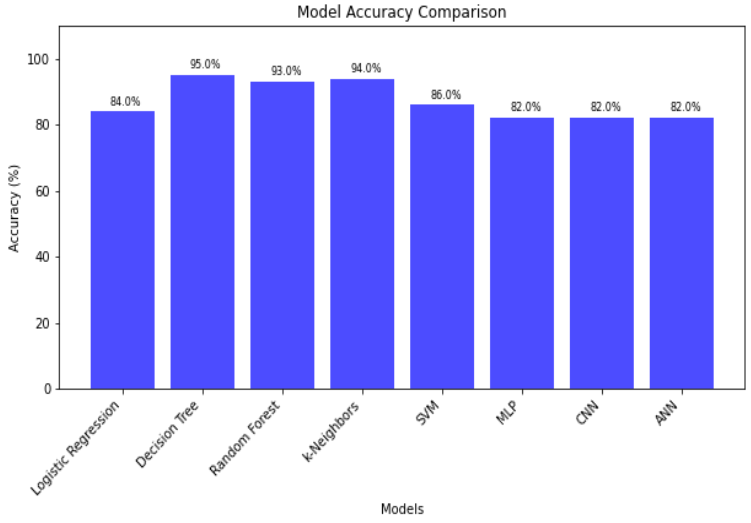
CNN Accuracy: 0.5233160621761658
Classification Report:

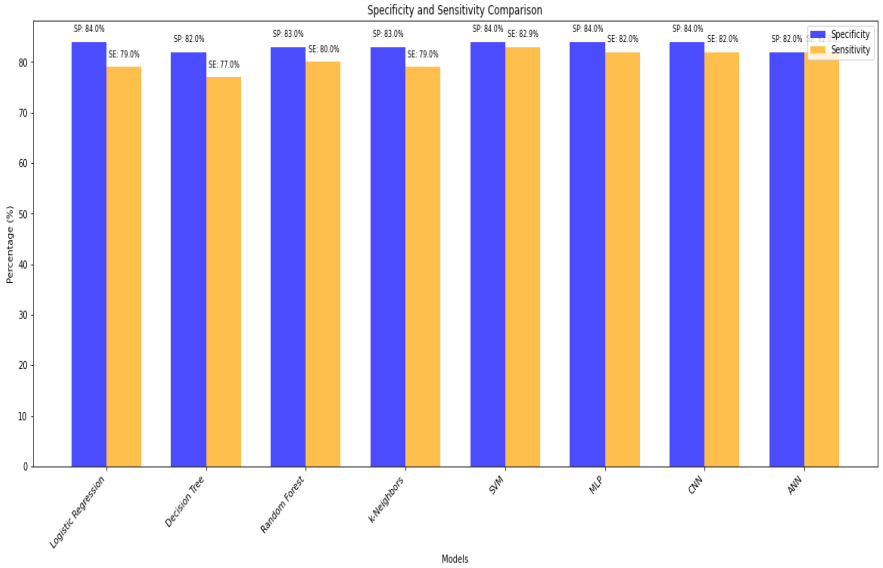
	precision	recall	f1-score	support
0	0.52	1.00	0.69	101
1	0.00	0.00	0.00	92
accuracy			0.52	193
macro avg	0.26	0.50	0.34	193
weighted avg	0.27	0.52	0.36	193

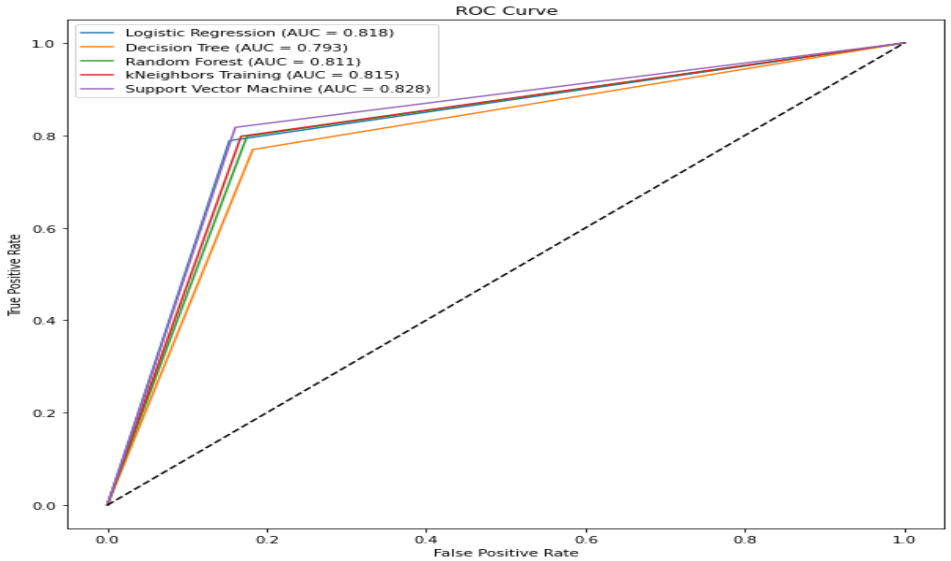
ANN Classification Report:

	precision	recall	f1-score	support
0.0	0.83	0.78	0.80	95
1.0	0.80	0.85	0.82	97
accuracy			0.81	192
macro avg	0.81	0.81	0.81	192
weighted avg	0.81	0.81	0.81	192

Interpretation of the Results(Mammographic Mass Dataset)





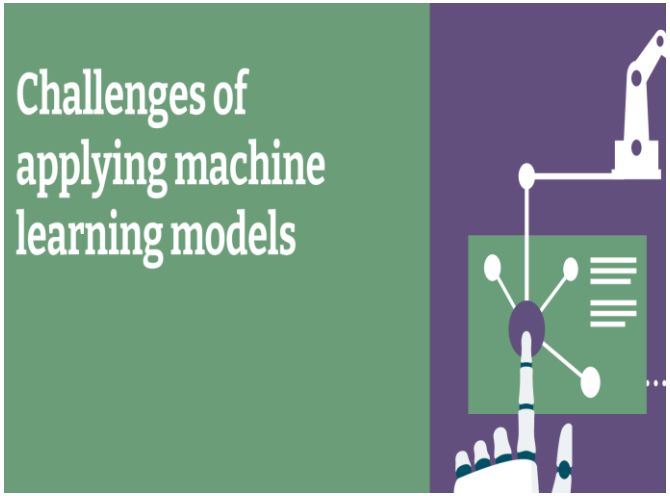


1. The ensemble models (Random Forest, k-Nearest Neighbors) outperform individual models, showcasing the power of combining multiple learners.
2. Support Vector Machine and Logistic Regression provide a good balance between sensitivity and specificity.
3. Neural networks (ANN and CNN) offer competitive performance, but their interpretability may be limited compared to traditional machine learning models.

Challenges

The challenges we faced during detecting breast cancer with machine learning approach :

- Data Variability: Diverse datasets (Wisconsin, Coimbra, mammographic) differ in collection methods, formats, and scales, posing integration challenges.
- Imbalanced Data: More benign cases can imbalance datasets, affecting the model's accuracy in predicting malignancy.
- Model Interpretability: Complex models like neural networks can be hard to interpret, hindering trust in clinical settings.
- Ethical Concerns: Using machine learning in healthcare raises ethical issues related to privacy, consent, and algorithm biases.
- Generalization Challenges: Models may struggle to generalize to diverse populations, requiring robustness across different demographics.



Refection on Experiment-1 (Wisconsin Dataset)

- Hypothesis testing suggested that the mean radius of benign tumors is less than malignant tumors.
- Features like ['radius_mean', 'texture_mean', 'smoothness_mean', 'compactness_mean','concavity_mean', 'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'smoothness_se', 'compactness_se', concavity_se', 'concave points_se', 'symmetry_se','fractal_dimension_se', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'symmetry_worst', 'fractal_dimension_worst'],,' were identified as indicative of tumor malignancy.
- Machine learning algorithms (Logistic Regression, Decision Tree, Random Forest, k-Neighbors, SVM, MLP, CNN, ANN) were employed.
- High sensitivity (minimizing false negatives) was crucial.
- Recommendations:

✓ For high sensitivity: Logistic regression, Support Vector Machine and k-Neighbors models seem to be strong contenders for cancer prediction.

Refelection on Experiment-2 (Coimbra Dataset)

- Features like Glucose, Age, Resistin, BMI, and Insulin were indicative of tumor nature.
- Machine learning algorithms (Decision Tree, Random Forest, k-Neighbors, SVM, MLP, CNN, ANN) were applied.
- Observations:

Based on the results, We observed that

✓ Decision Tree and k-Nearest Neighbors models exhibit potential overfitting due to perfect training accuracy.
✓ Random Forest and Support Vector Machine models show good generalization.
✓ Neural network models, especially the CNN, demonstrate competitive performance, outperforming the ANN.

Reflection on Experiment-3 (Mammography Mass Dataset)

- Features such as BI-RADS, Density and Age were important.
- Machine learning algorithms (Logistic Regression, Decision Tree, Random Forest, k-Neighbors, SVM, MLP, CNN, ANN) were employed.
- Observations.

✓ The ensemble models (Random Forest, k-Nearest Neighbors) outperform individual models, showcasing the power of combining multiple learners.

✓ Support Vector Machine and Logistic Regression provide a good balance between sensitivity and specificity.

✓ Neural networks (ANN and CNN) offer competitive performance, but their interpretability may be limited compared to traditional machine learning models

Conclusion

- In conclusion, our analysis across Wisconsin, Coimbra, and Mammographic Mass datasets highlights the significance of machine learning in breast cancer diagnosis.
- Features like mean radius, Age , Insulin, Density, Glucose, Resistin and specific biomarkers prove crucial.
- For high sensitivity, models like Logistic Regression, Decision Tree, Random Forest, SVM, MLP, CNN, and ANN shine, but careful consideration is needed to balance interpretability.
- Ensemble models, especially Random Forest, and SVM demonstrate robust generalization.
- Moving forward, validating models on diverse datasets, collaborating with experts, and adapting to evolving research are key for successful implementation in clinical settings.
- Detecting breast cancer early is crucial, and machine learning provides valuable tools for this essential part of breast cancer treatment.

General Recommendations-Technical Perspective

- Consider the task-specific requirements for model selection.
- Prioritize high sensitivity for early detection in breast cancer diagnosis.
- Evaluate the trade-off between model complexity and interpretability.
- Ensemble models, such as Random Forest, can offer improved performance.
- Regularly validate models on independent datasets to ensure generalizability.
- Collaborate with medical experts for a more comprehensive understanding of feature importance and model outputs.

General Recommendations- Medical Perspective

- Regular Monitoring:

- Monitor individuals with malignancy-indicative features regularly.

- Focused Assessments:

- Assess 'Glucose', 'Age', 'Resistin', 'BMI', and 'Insulin' for tumor nature.

- Tailored Screening:

- Customize breast cancer screenings based on 'BI-RADS', 'Density', and 'Age'.

- Patient Education:

- Educate patients about feature significance and encourage regular check-ups.

- Consult healthcare professionals for personalized advice.

Meeting the Mark: Grading Criteria-5

The project meets the Grading Criteria-5 based on the following requirements.

Requirements	Grading Criteria-5
Datasets	▪ Used 3 datasets(vectorized data) from UCI Repository ▪ Breast Cancer Wisconsin, Breast cancer Coimbra and Mammography Mass Dataset
	Tasks Performed
Feature Engineering	▪ Data Exploration and Understanding; Handling Missing Data; Variable Encoding; Feature Scaling; Feature Creation; Handling Outliers; Feature Selection; Feature Extraction; Cross-validation.
Machine Learning Algorithms	▪ A total of 8 Machine Learning algorithms applied. Logistic Regression, Decision Tree, Random Forest, k-Neighbors, SVM, MLP, CNN, ANN
Performance Metrics	▪ Sensitivity (Recall): True positive rate | Specificity: True negative rate | F1 Score: Harmonic mean of precision and recall | Accuracy: Overall correctness | AUC (Area Under the Curve): Discrimination ability.
Comparison of models	▪ Compared the performance for two or more different algorithms on three different datasets.
Well documented Results	▪ Documented the results ▪ Reflections and Recommendations are well explained

Thank you