Report ~ Prediction of Breast Cancer

1)Investigating The Data:
The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is a well-known dataset in the field of machine learning and medical research. Here's an analysis of its reliability and potential use:

Source and Creators:

The dataset was created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian, all associated with the University of Wisconsin in the early 1990s.(1992) Dr. Wolberg is affiliated with the General Surgery Department, while Street and Mangasarian are affiliated with the Computer Sciences Department. These individuals have academic backgrounds and expertise in their respective fields, lending credibility to the dataset's creation.

Data Collection:

The dataset consists of features computed from digitized images of fine needle aspirates (FNAs) of breast masses. These features describe characteristics of cell nuclei present in the images. Images were digitized and analyzed, and relevant features were extracted for analysis.

Past Usage:

The dataset has been used in various academic publications, including those by the creators themselves and other researchers in the field. It has been cited in multiple publications, indicating its widespread use and recognition within the scientific community. The dataset has been used for breast cancer diagnosis, prognosis, and predictive modeling in machine learning research.

Reliability:

The dataset has been extensively used in academic research and cited in reputable journals, indicating a level of reliability and acceptance within the scientific community. It has been used for various purposes such as diagnostic prediction, prognosis, and machine learning model development. The fact that the dataset has been used in different contexts and by different researchers speaks to its versatility and reliability.

Data Quality:

The dataset's features were selected based on an exhaustive search and expert knowledge in the field of breast cancer diagnosis. The datasets are labeled with the diagnosis (benign or malignant), indicating clear distinctions between the classes.

Applicability:

Given its reliability, past usage, and well-defined features, the WDBC dataset can be considered suitable for various applications related to breast cancer diagnosis and prognosis. However, since the dataset is from the early 1990s, researchers should consider whether advancements in technology and medical understanding have made any features or conclusions outdated.

Additionally, it's important to verify that any predictive models trained on this dataset generalize well to current medical practices and patient populations.

In conclusion, the Wisconsin Diagnostic Breast Cancer dataset appears to be a reliable and valuable resource for research in breast cancer diagnosis and prognosis. Its extensive usage in academic publications and its creators' expertise lend credibility to its quality and relevance. However, researchers should exercise caution and ensure that any conclusions drawn from the dataset are validated against current medical knowledge and practices.

2) Exploratory Data Analysis: First of all, we need to read the CSV file and make some arrangements such as aligning columns, applying numeric adjustments, and adjusting classes.

a) Checking Histograms: Looking at the general distribution of the data will give us some insight into how the data must be preprocessed, including whether it requires Z-Score normalization and how outliers should be handled. We can clearly see that all of the features scale from 1 to 10, and most of their distributions resemble each other. Therefore, we don’t have to perform Z-Score normalization. However, we do need to address handling outliers, as there are some unbalanced histogram charts present.

b) Checking Null and Unique Values: Since we already know that the features scale from 1 to 10, we are aware that there isn’t a redundant column containing numerous unique values. However, we will still check the unique values of each column. Additionally, for null values, we can handle them differently for each column depending on which column has null values. Depending on the balance of the column, we can adopt different approaches. If a column is balanced, we can simply fill the missing values with the mean. However, if the column isn't balanced, we'll need to implement a different process.

c) Handling Null Values: The column ‘Bare_nuclei’ isn’t balanced at all so filling it with the mean value wouldn't be appropriate. However, we observe that the frequency of 1 is greater than the rest. Therefore, we can handle it using the mode. But if features such as ‘Clump_thickness’ or ‘Bland_chromatin’ would have null values we could easily handle with the mean.

d) Checking Outliers: We observed in the histograms that the distribution of the data is somewhat messy; therefore, we need to examine boxplot graphs to determine whether any outliers exist. Upon checking the boxplot, we can clearly see that outliers are present.

e) Handling Outliers: We will address the extreme outliers using the Interquartile Range (IQR) method. Additionally, we will identify which column has the most outliers and determine the best approach to handle them. We removed the extreme outliers, all of which

belonged to one specific column. However, most of these outliers were instances of Class 1. This situation may pose a problem because the removed instances could be crucial data for Class 1, and nearly 40% of Class 1's instances are now missing.

f) Trying the both datasets: When we compare the performance of both datasets—one with outliers and one without outliers—we can clearly observe that the dataset with outliers performs better. The primary reason behind this is that when we handle outliers, instances belonging to Class 1 lose significant insights for interpreting the data, leading to a decrease in performance.

3) Comparing Models: If we want to compare models, we need to understand what accuracy, precision, recall, and ROC AUC score are.

Accuracy Analogy: Imagine you're a basketball player aiming for the hoop. Accuracy is like the percentage of shots you make out of all the shots you attempt.

Explanation: A high accuracy means you consistently make shots, hitting the hoop more often than not. It measures how well you're hitting the target compared to the total number of attempts.

Precision Analogy: Imagine 100 people take a COVID test, and all of them test positive, but not all of them are actually sick. The ratio of true positives to all positive test results, including both true positives and false positives, is known as precision.

Explanation: Precision measures how consistent and uniform your actions are. A high precision in classification means your model's predictions are consistent and reliable, with minimal variation.

Recall Analogy: Recall is like a detective trying to catch all the suspects in a criminal investigation.

Explanation: Recall measures how thorough your search is in finding all the relevant instances. Similar to a detective's goal of capturing all suspects, a high recall in classification means your model identifies most of the positive instances, minimizing the chance of missing any relevant cases.

ROC AUC Analogy: ROC AUC score is like a medical test's ability to distinguish between healthy and sick patients.

Explanation: ROC AUC score evaluates the model's ability to differentiate between positive (sick) and negative (healthy) instances. Just as you'd want a medical test to accurately identify sick patients while minimizing false positives (healthy patients wrongly classified as

sick), a high ROC AUC score indicates the model's effectiveness in correctly classifying positive cases while keeping false positives low.

Actually, for this dataset aimed at detecting breast cancer, it's crucial to detect cancer accurately. We want to avoid administering chemotherapy to individuals who do not have cancer. Therefore, both recall and precision are the most important metrics for us to evaluate models.

Here is the metrics table:

Model	Accuracy	Precision	Recall	ROC AUC Score
XGBoost Classification	0.97	0.97	0.98	0.94
Logistic Regression	0.97	0.97	0.98	0.94
Random Forest Classification	0.97	0.97	1.0	0.94
Support Vector Classification	0.97	0.97	1.0	0.94
Neural Network Classification	0.97	0.97	0.98	0.94
KNN	0.98	0.98	1.0	0.98

KNN Classification:

KNN achieves the highest accuracy, precision, and ROC AUC score among all the models listed. It also has perfect recall, indicating strong performance in identifying all positive instances.

Random Forest Classification:

Random Forest performs consistently well across all metrics, with high accuracy, precision, and perfect recall. However, its ROC AUC score is slightly lower compared to KNN.

Support Vector Classification:

Support Vector Classification achieves high accuracy, precision, and perfect recall, similar to Random Forest. However, its ROC AUC score is slightly lower.

XGBoost Classification:

XGBoost performs well across all metrics, with high accuracy, precision, and recall. Its ROC AUC score is slightly lower than Random Forest and Support Vector Classification.

Logistic Regression:

Logistic Regression also performs consistently well across all metrics. It's a simple and interpretable model, but its performance is slightly lower compared to Random Forest, Support Vector Classification, and XGBoost.

Neural Network Classification:

The neural network model performs well, similar to logistic regression and XGBoost. However, it might require more computational resources and tuning compared to simpler models.

Based on this comparison, KNN is ranked as the most usable model due to its highest overall performance across all metrics. Random Forest and Support Vector Classification follow closely behind, while logistic regression, XGBoost, and neural network classification exhibit slightly lower performance but are still viable options depending on other considerations such as model complexity and interpretability.