Q3:

Model Performance Comparison on Breast Cancer Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| SVM | 0.97 | 0.98 | 0.96 | 0.97 |
| k-NN (k=3) | 0.956 | 1.0 | 0.891 | 0.943 |
| Perceptron | 0.947 | 0.955 | 0.913 | 0.933 |
| Logistic Regression | 0.956 | 0.977 | 0.913 | 0.944 |

Based on the results from the **Breast Cancer dataset**, **SVM** stands out as the best performing model with an accuracy of 97%, precision of 98%, recall of 96%, and an F1-score of 97%. This high performance is due to SVM's ability to maximize the margin between classes, making it highly effective in distinguishing between malignant and benign tumors. In contrast, **k-NN** (with k=3) achieved an accuracy of 95.6% and perfect precision (100%), but its recall was lower at 89.1%, meaning it missed a few malignant cases. The F1-score of 94.3% reflects this trade-off between precision and recall. **Perceptron** performed well with a test accuracy of 94.7%, precision of 95.5%, and recall of 91.3%, but it showed signs of overfitting, as indicated by a perfect training accuracy of 100%. **Logistic Regression**, with an accuracy of 95.6%, precision of 97.7%, recall of 91.3%, and an F1-score of 94.4%, also provided a strong balance between precision and recall. While both Logistic Regression and k-NN performed similarly in terms of accuracy, SVM consistently outperformed them in terms of overall metrics. Therefore, **SVM** is the best model for this dataset, offering the most reliable performance in classifying breast cancer cases.

Q4.

The **K-Means clustering (from scratch)** achieved an accuracy of **91%**, with a **precision** of **93%**, **recall** of **82%**, and an **F1-score** of **87%** on the Breast Cancer dataset. While the overall performance of the clustering algorithm is relatively high, especially in terms of accuracy and precision, it still falls short when compared to supervised learning models like **SVM** and **Logistic Regression**. The lower **recall** of 82% indicates that the K-Means algorithm missed some malignant cases, which is critical in medical applications where identifying malignant tumors is of utmost importance.

The high precision, however, shows that when K-Means does predict a malignant tumor, it is highly likely to be correct. The balance between precision and recall, as captured by the **F1-score** of **87%**, demonstrates that K-Means clustering can reasonably group the data into two clusters but may struggle to capture all malignant cases effectively.

Compared to supervised models like **SVM** (which had higher accuracy and recall), K-Means is less effective because it does not have access to class labels during training. Supervised models benefit from this labeled information, which allows them to learn a more accurate decision boundary, whereas K-Means relies solely on the inherent structure of the data, which can lead to less optimal cluster separation, especially when dealing with complex datasets like this one.

In summary, while **K-Means clustering** can provide reasonable classification in an unsupervised setting, it cannot match the performance of supervised models in terms of recall and overall reliability, particularly for critical tasks like breast cancer detection.