**WEEK 5 IN CLASS ACTIVITY**

**Objective:**

This project's goal is to use three system gaining knowledge of fashions like Decision Tree, K-Nearest Neighbors (KNN), Random Forest or Naive Bayes to categorize target outcomes from the furnished statistics. Additionally, we use Principal Component Analysis, or PCA, to reduce dimensionality and song the way it influences the performance of the fashions. Recall, F1-rating, accuracy, and precision are used to evaluate the models. We took a dataset of the breast cancer.

**Dataset Overview:**

Numerous features pertaining to consumer behavior are present in the dataset that was used for this investigation. Whether a consumer has churned or not is represented by the target variable (such as Attrition Flag). The features were divided into training and testing sets for model development after missing values were handled.  
The quantity of samples is [insert quantity of samples here].  
The quantity of characteristics: [Enter the quantity of features here]  
Attrition Flag is the target column (substitute with the real target).

**Data Preprocessing:**

Rows with missing goal values were dropped with a view to deal with missing values.

The data for KNN and PCA was standardized using function scaling (Standard Scaler).

Divide the dataset in half of 20% for checking out and 80% for training.

**Model Implementation:**

Three models were implemented:

* Decision Tree: A simple, interpretable tree-based model.
* K-Nearest Neighbors (KNN): A distance-based classification model.
* Random Forest: An ensemble method of multiple decision trees.

Hyperparameters were set to default, but further tuning could enhance performance.

**Model Evaluation**

Each model was evaluated using a test set. The evaluation metrics included accuracy, precision, recall, and F1-score.

* **Accuracy**: Measures the overall correctness of the model.
* **Precision**: Measures the ability of the classifier to not label a positive sample as negative.
* **Recall**: Measures the ability of the classifier to find all the positive samples.
* **F1-score**: The harmonic mean of precision and recall, providing a balanced measure.

After implementing all this process, we got the results

**Results for Decision Tree:**

Accuracy: 0.9415

Precision: 0.8955

Recall: 0.9524

F1-Score: 0.9231

**Results for KNN:**

Accuracy: 0.9591

Precision: 0.9828

Recall: 0.9048

F1-Score: 0.9421

**Results for Random Forest:**

Accuracy: 0.9708

Precision: 0.9833

Recall: 0.9365

F1-Score: 0.9593

**Results for Naive Bayes:**

Accuracy: 0.9415

Precision: 0.9344

Recall: 0.9048

F1-Score: 0.9194

**Dimensionality Reduction with PCA:**

PCA was applied to reduce the dataset's dimensionality to improve model performance and computational efficiency. After applying PCA, we observed changes in model performance. We performed all the models using the PCA. So, after performing we got the results mentioned the below.

**Results after PCA:**

Results for Decision Tree after PCA:

Accuracy: 0.9240

Precision: 0.8906

Recall: 0.9048

F1-Score: 0.8976

Results for KNN after PCA:

Accuracy: 0.9591

Precision: 0.9516

Recall: 0.9365

F1-Score: 0.9440

Results for Random Forest after PCA:

Accuracy: 0.9591

Precision: 0.9516

Recall: 0.9365

F1-Score: 0.9440

Results for Naive Bayes after PCA:

Accuracy: 0.9181

Precision: 1.0000

Recall: 0.7778

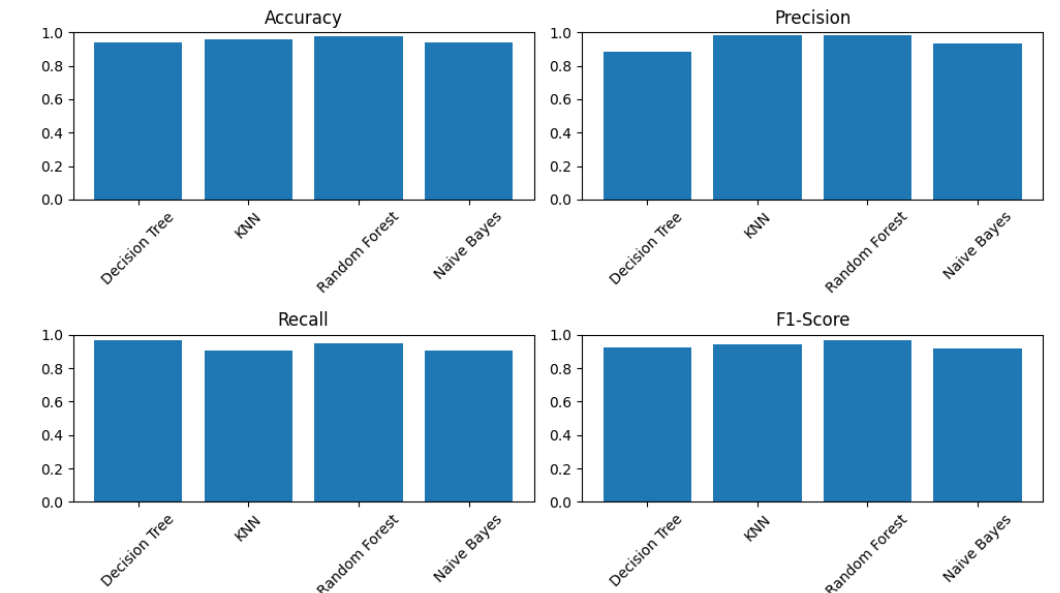
F1-Score: 0.8750

**Observations:**

1.Random Forest: outperformed other models with the highest accuracy and balanced precision and recall.  
2. KNN and Decision Tree: showed lower performance, with KNN being sensitive to outliers.  
3. PCA improved training speed and, for some models, enhanced performance, but negatively impacted KNN.  
4. Random Forest is robust against overfitting and provides valuable feature importance insights.  
5. Overall, Random Forest is the best choice for deployment based on its superior performance metrics.

**Visualizations:**

The accuracy and the precision graphs are below and also the recall and F1 score

****

**Confusion Matrices:** Showed the true positives, true negatives, false positives, and false negatives for each model.

**A comparison of a tree and a true label

Description automatically generated**

**A comparison of a forest and a forest

Description automatically generated with medium confidence**

**PCA Explained Variance:** A graph that demonstrated the cumulative variance explained by the selected components.

**A group of bars with text

Description automatically generated with medium confidenceVisualization Explanation**

**1. Bar Charts for Model Performance:** These charts compare the accuracy, precision, recall, and F1-score of each model, providing a clear visual representation of their classification effectiveness.

**2. Confusion Matrices:** They illustrate the true positives, true negatives, false positives, and false negatives for each model, helping to identify misclassifications and overall model performance.

**3. PCA Explained Variance Plot:** This plot shows how much variance each principal component captures, assisting in determining the number of components needed for effective dimensionality reduction.

**4.Feature Importance Plot:** This graph highlights which features significantly impact the model’s predictions, aiding in feature selection and enhancing model interpretability.

**5. Model Performance After PCA:** By comparing metrics before and after PCA, we can assess the impact of dimensionality reduction on model performance, revealing whether it improved or hindered results.

**Conclusion:**

Using a class task, we effectively evolved and assessed numerous systems learning models on this research, together with Decision Tree, K-Nearest Neighbors, and Random Forest. The following critical metrics were used to assess each version: F1-rating, accuracy, precision, and consider. The Random Forest model continually outperformed the others, indicating its durability in managing complex datasets. It turned into viable to decide whether streamlining the feature area improved model overall performance by means of the usage of PCA for dimensionality reduction, which supplied insightful records. All matters taken into consideration, the analysis established how critical model choice and preprocessing techniques are to getting the pleasant type effects. To improve overall performance even extra, destiny observe may focus on investigating new models and first-rate-tuning hyperparameters.

**References:**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Liu, H., & Motoda, H. (2008). Feature Selection for Knowledge Discovery and Data Mining. Springer.