AI Astra – Prediction Report (Sprint 2)

Project Phase: Sprint 2 Deliverables

1. Executive Summary

In Sprint 2, we developed and evaluated multiple machine learning models to predict cancer risk. The models tested included Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Based on evaluation metrics such as Accuracy, Precision, Recall, F1 Score, and ROC-AUC, the Random Forest model was identified as the best-performing model. Although Decision Tree results appeared similar in many aspects, Random Forest achieved a higher ROC-AUC score, indicating superior ability to distinguish between positive and negative cases. Therefore, Random Forest was chosen as the final model for deployment.

2. Dataset & Preprocessing

The dataset was cleaned and preprocessed before modeling. Key preprocessing steps included:

- Handling missing values and duplicates
- Encoding categorical variables (Gender, Smoking, Alcohol, Insurance Type, etc.)
- Scaling numerical features such as Age, BMI, and Blood Sugar
- Generating a processed dataset (AI_ASTRA_ProcessedDataset.xlsx) used in model training

3. Models Evaluated

The following classification models were trained and compared:

- Logistic Regression – A linear model that outputs probabilities using a sigmoid function. It is simple, interpretable, and often used as a baseline classifier.

- Decision Tree – A tree-based model that splits data based on feature thresholds. It is easy to visualize but prone to overfitting on training data.

- Random Forest – An ensemble method combining multiple decision trees. It reduces overfitting, increases robustness, and generally provides higher AUC compared to a single tree.

- Support Vector Machine (SVM) – A model that finds an optimal hyperplane to separate classes. Effective in high-dimensional spaces but can be computationally expensive.

- K-Nearest Neighbors (KNN) – A non-parametric method that classifies based on the closest neighbors. It is simple but less effective on large datasets or when features are noisy.

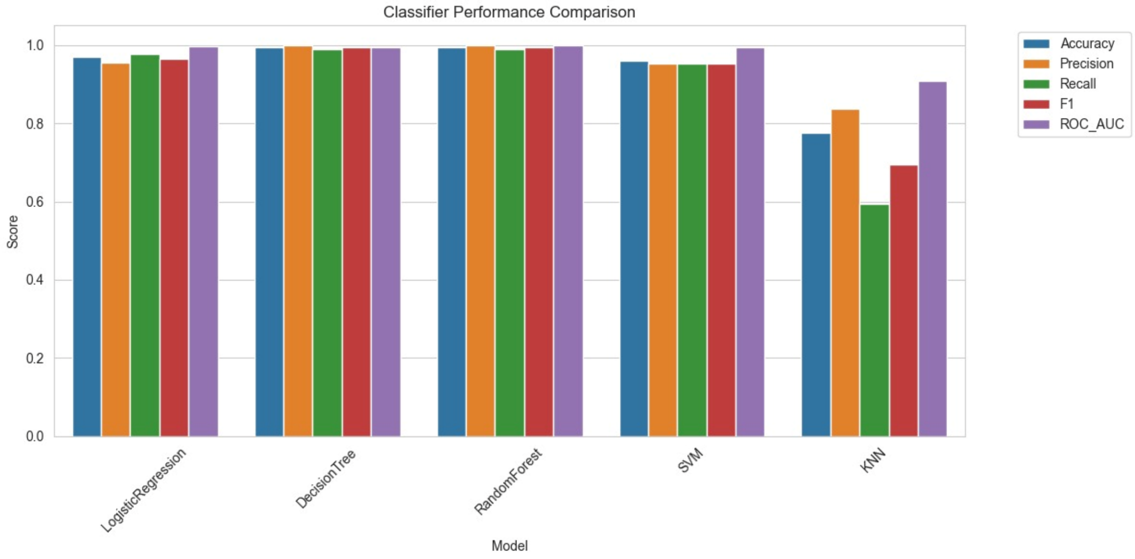
4. Performance Metrics

Model performance was measured using:

- Accuracy – overall proportion of correct predictions
- Precision – proportion of predicted positives that were correct
- Recall – proportion of actual positives correctly identified
- F1 Score – harmonic mean of precision and recall
- ROC-AUC – trade-off between sensitivity (recall) and specificity

5. Model Comparison Results

The chart below shows the performance comparison across all models using multiple evaluation metrics.



6. Key Findings

- Logistic Regression and Decision Tree provided competitive baseline performance.
- Random Forest outperformed Decision Tree with a notably higher ROC-AUC, indicating stronger generalization and reduced risk of overfitting.
- SVM achieved good overall results but was computationally heavier.
- KNN lagged behind in precision and recall, making it less reliable for this dataset.

7. Visual Insights

ROC curves, Precision-Recall curves, and confusion matrices were generated for each model. These visualizations provided insight into sensitivity and trade-offs between false positives and false negatives. The Random Forest model consistently showed superior AUC and balanced performance across metrics.

8. Classification evaluation metrics

To evaluate the performance of machine learning models, we rely on classification metrics derived from the confusion matrix.

Confusion Matrix Terms:

- True Positive (TP): Model correctly predicts a positive outcome.

- True Negative (TN): Model correctly predicts a negative outcome.

- False Positive (FP): Model incorrectly predicts a positive outcome.

- False Negative (FN): Model incorrectly predicts a negative outcome.

From these, the following key metrics are derived:

- **Accuracy**: Overall proportion of correct predictions.

  Formula: (TP + TN) / (TP + TN + FP + FN)

  - Best used when classes are balanced.

- **Precision**: Of all positive predictions, how many were actually correct.

  Formula: TP / (TP + FP)

  - Important when the cost of false positives is high.

- **Recall (Sensitivity)**: Of all actual positives, how many did the model identify.

  Formula: TP / (TP + FN)

- Important when the cost of false negatives is high.

- **F1 Score**: Harmonic mean of precision and recall.

  Formula: 2 * (Precision * Recall) / (Precision + Recall)

  - Useful when dataset is imbalanced and a balance between precision and recall is required.

- **ROC-AUC** (Receiver Operating Characteristic – Area Under Curve):

  - Measures the trade-off between true positive rate and false positive rate across thresholds.

  - A higher AUC indicates stronger model discrimination ability.

9. Conclusion & Recommendation

After evaluating multiple models, the Random Forest classifier was chosen as the best-performing model. Although the Decision Tree showed similar results for some metrics, the Random Forest's higher ROC-AUC demonstrates its superior ability to distinguish between cancer-positive and cancer-negative cases. This advantage comes from Random Forest's ensemble approach, which reduces variance and prevents overfitting compared to a single Decision Tree. Given its robustness, higher recall, and stronger generalization ability, Random Forest is recommended for deployment in cancer risk prediction.