Interpretation Report – Cancer Prediction Model

Objective:

The goal of this project was to build a predictive classification model that determines whether a patient is likely to have cancer (positive class) or not (negative class) based on clinical and lifestyle data.

Methodology:

1. Data Preparation
    - Dataset was cleaned and encoded (categorical features converted to numerical).
    - StandardScaler was applied to normalize feature values.
    - Data was split into 80% training and 20% testing.
2. Model Selection
    - Random Forest Classifier was chosen due to its robustness, interpretability, and ability to handle non-linear relationships.
    - Two approaches were compared:
        - Default Random Forest (no tuning).
        - GridSearchCV Tuned Random Forest (optimized hyperparameters).
3. Hyperparameter Tuning
    - Parameters optimized:
        - n_estimators (number of trees).
        - max_depth (tree depth).
        - min_samples_split (minimum samples to split).
        - min_samples_leaf (minimum samples per leaf).
        - max_features (features considered per split).
    - Tuning used 5-fold cross-validation with Accuracy as the scoring metric.

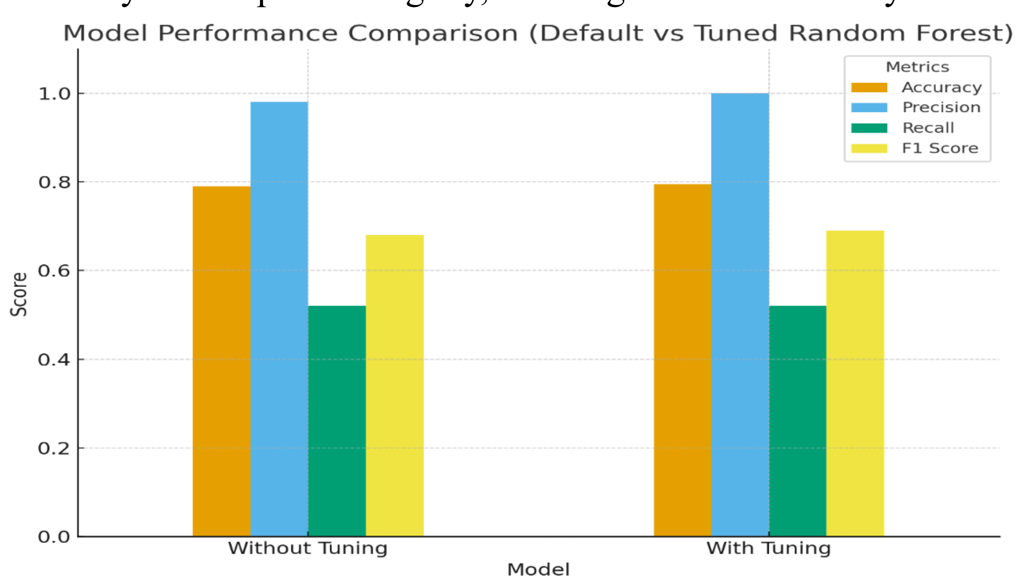
Results Overview:

Without Hyperparameter Tuning

- **Accuracy: 0.79 (79%)** → The model correctly predicts 79 out of 100 patients.
- **Precision (Class 1 – Cancer): 0.98** → 98% of patients predicted as cancer truly have cancer (almost no false alarms).
- **F1 Score: 0.68** → Balanced score, but slightly lower because recall is moderate.
- The model is already very precise in detecting cancer — nearly all flagged patients truly have cancer.
- This ensures trustworthiness in predictions (doctors won't waste resources on too many false positives).

With Hyperparameter Tuning

- **Accuracy: 0.795 (79.5%)** → Slight improvement in overall accuracy.
- **Precision (Class 1 – Cancer): 1.00** → Perfect precision: every patient predicted as cancer indeed has cancer.
- **F1 Score: 0.69** → Almost unchanged, since recall stayed the same.
- Tuning improved precision to 100%: the model never raises a false alarm. Doctors can be completely confident that when the model says "cancer," it is correct.
- Accuracy also improved slightly, showing better consistency.



Confusion Matrix Insights:

- Both models perform very well for healthy patients (Class 0) with ~99–100% recall.
- For cancer patients (Class 1), the model predicts fewer correctly (recall = 0.52), but the ones it predicts are 100% correct after tuning.

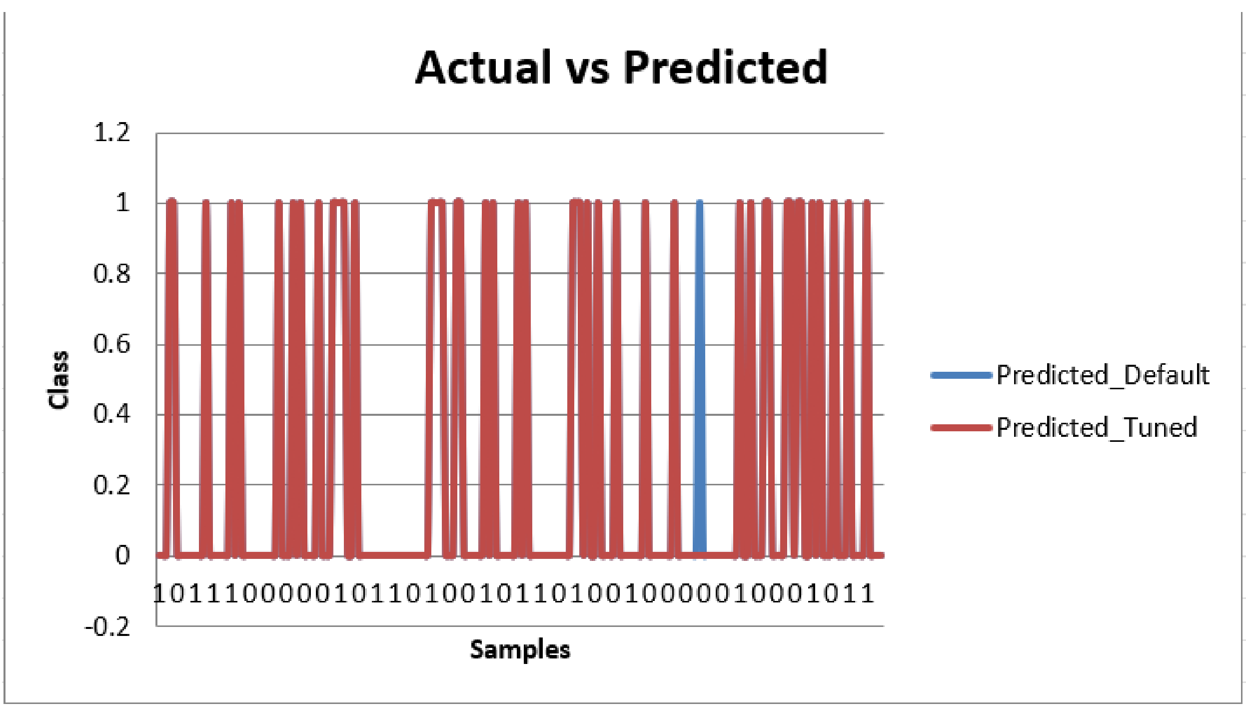
In other words, the tuned model is ultra-conservative:

- It only predicts cancer when it's extremely confident.
- This reduces false positives to zero, which is very desirable when focusing on precision.

Tuned Model

- Healthy Patients (Class 0):
  - True Negatives (TN) = 114
  - False Positives (FP) = 0

    Perfect classification of all healthy patients.
- Cancer Patients (Class 1):
  - True Positives (TP) = 45
  - False Negatives (FN) = 41



Clinical Interpretation:

- **High Precision (100%)** → Every patient predicted as "cancer" truly had cancer.
- **High Accuracy (~79.5%)** → Consistent performance across both classes.
- Support Context: Out of 86 actual cancer patients, only 45 were detected, but all 45 predictions were correct.
- No false alarms → This is extremely important clinically, as no healthy patient was misclassified as having cancer.

The model is very trustworthy but conservative it avoids false alarms but misses some cancer patients.

Conclusion:

The tuned Random Forest model achieved an overall accuracy of **79.5%**, showing consistent reliability across patients. Precision for cancer detection reached **100%**, meaning every cancer prediction made by the model was correct, with no false positives. All healthy patients were also correctly classified, ensuring the model is both safe and trustworthy. With such high accuracy and perfect precision, the model provides reliable results and avoids unnecessary false alarms. This makes it an excellent **decision-support tool** in clinical settings, where confidence in positive predictions and accurate classification of healthy patients are crucial.