Cancer Cell Prediction Using Machine Learning

An End-to-End ML Lifecycle Project

Project Team:

Project Manager: Rohit Singh

Data Engineer: Dhrithi MV

Data Scientist: Lavanya K Gowda

ML Engineer: Kusumitha KP

Quality Test Engineer: Sourabh Rajendragouda Doddagoudar

ML Ops: Keerthana V & Likhitha D

Executive Summary: A Breakthrough in Cancer Prediction

Our project successfully developed, trained, and evaluated multiple machine learning models to predict cancer risk, culminating in a highly effective solution ready for deployment.

Project Goal

Build a classification model for accurate cancer detection.

Key Outcome

Developed and evaluated robust models for predicting cancer risk.

Top Performer

Random Forest model with superior ROC-AUC score.

Recommendation

Random Forest model is primed for real-world deployment.

Project Goal & Scope: Empowering Early Detection

Our core objective is to enhance early cancer detection, thereby improving treatment outcomes through an automated risk prediction system designed to support medical professionals.

Problem Statement

Early cancer detection is critical for successful treatment. This model provides an automated, AI-driven risk prediction to assist doctors in timely diagnosis and intervention.

Key Requirements

- **Functional:** Data cleaning, robust model training, thorough evaluation, and a prediction API for seamless doctor integration.
- **Non-Functional:** Over 90% recall (minimizing false negatives is paramount), real-time prediction latency under 2 seconds for immediate insights, and stringent HIPAA compliance for patient data security and privacy.

Agile Methodology & Timeline

We adopted an Agile Scrum framework, completing the project in three focused sprints over a six-week period to ensure iterative development and rapid delivery.

01

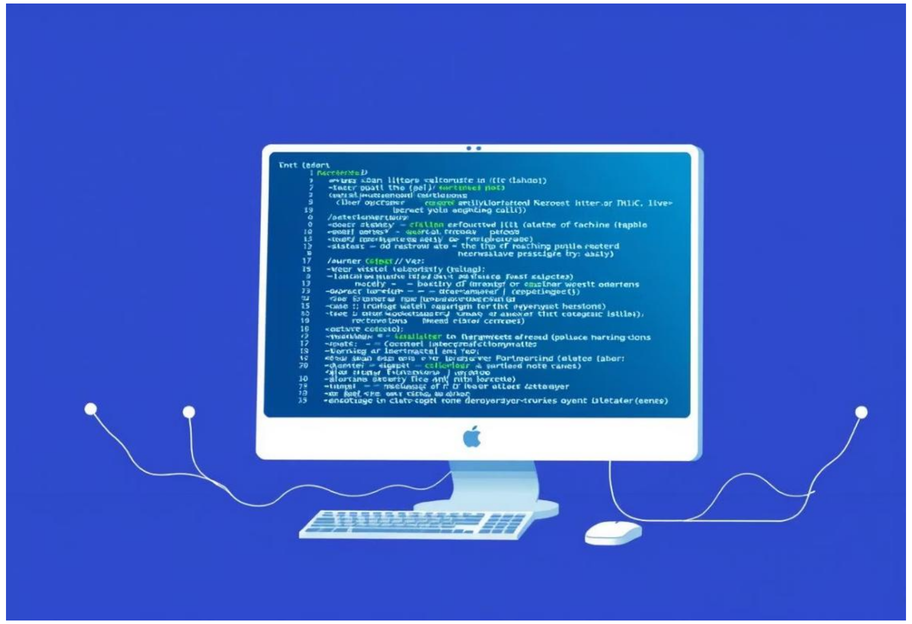
Sprint 1: Data Preparation & Setup

Weeks 1-2: Focused on data acquisition, cleaning, initial exploration, and establishing the project infrastructure.

02

Sprint 2: Model Development & Training

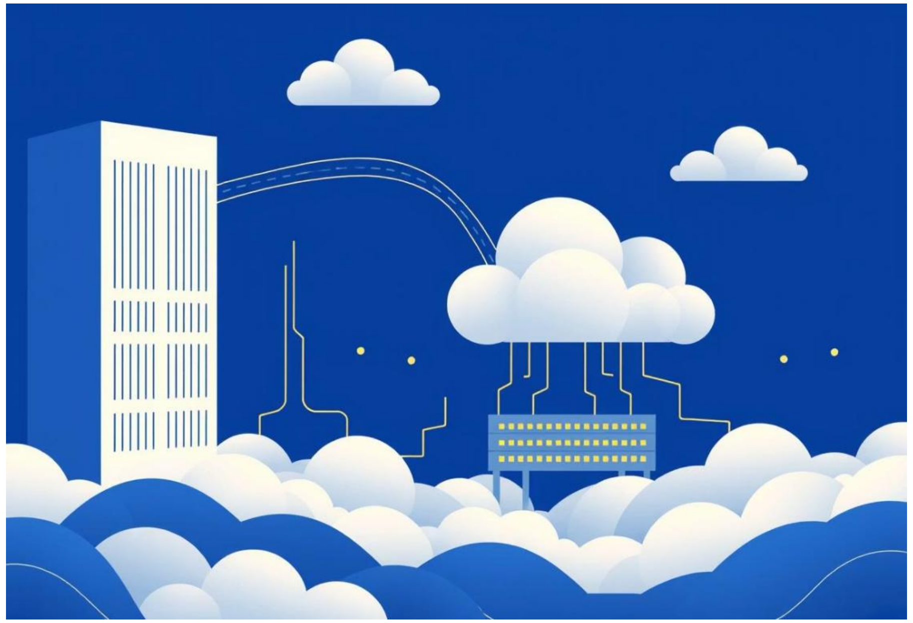
Weeks 3-4: Concentrated on selecting algorithms, model training, hyperparameter tuning, and preliminary evaluation.



03

Sprint 3: Deployment & Reporting

Weeks 5-6: Finalizing the best model, preparing for deployment, and compiling comprehensive project reports.



Data Preparation: Laying the Foundation for Accuracy

A meticulously processed dataset is crucial for the success of any machine learning model. Our efforts focused on ensuring data quality and optimal feature representation.

Dataset Overview

Our model leverages a comprehensive dataset incorporating key physiological and medical indicators. Features include:

- **Age:** Patient's age.
- **BMI:** Body Mass Index.
- **Systolic_BP:** Systolic Blood Pressure.
- **Blood_Sugar:** Blood Sugar levels.
- And other relevant health markers.

Key Preprocessing Steps

- **Handling Missing Values:** Imputation strategies to address gaps in data.
- **Duplicate Removal:** Ensuring data integrity by eliminating redundant entries.
- **BMI Recalculation:** Standardizing BMI values for consistency and accuracy.
- **Categorical Encoding:** Converting non-numerical data into a machine-readable format.
- **Feature Scaling:** Normalizing numerical features to prevent dominance by large-scale values.

Model Development & Evaluation Strategy

To identify the most effective cancer prediction model, we rigorously developed and evaluated five distinct classification algorithms using key metrics tailored to medical applications.

Models Evaluated

Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

Accuracy

Measures the proportion of overall correct predictions, indicating general model performance.

Recall (Sensitivity)

Critically important for not missing actual cancer cases (minimizing false negatives).

ROC-AUC Score

Evaluates the model's ability to discriminate between positive and negative classes.

Key Findings: Random Forest Dominates

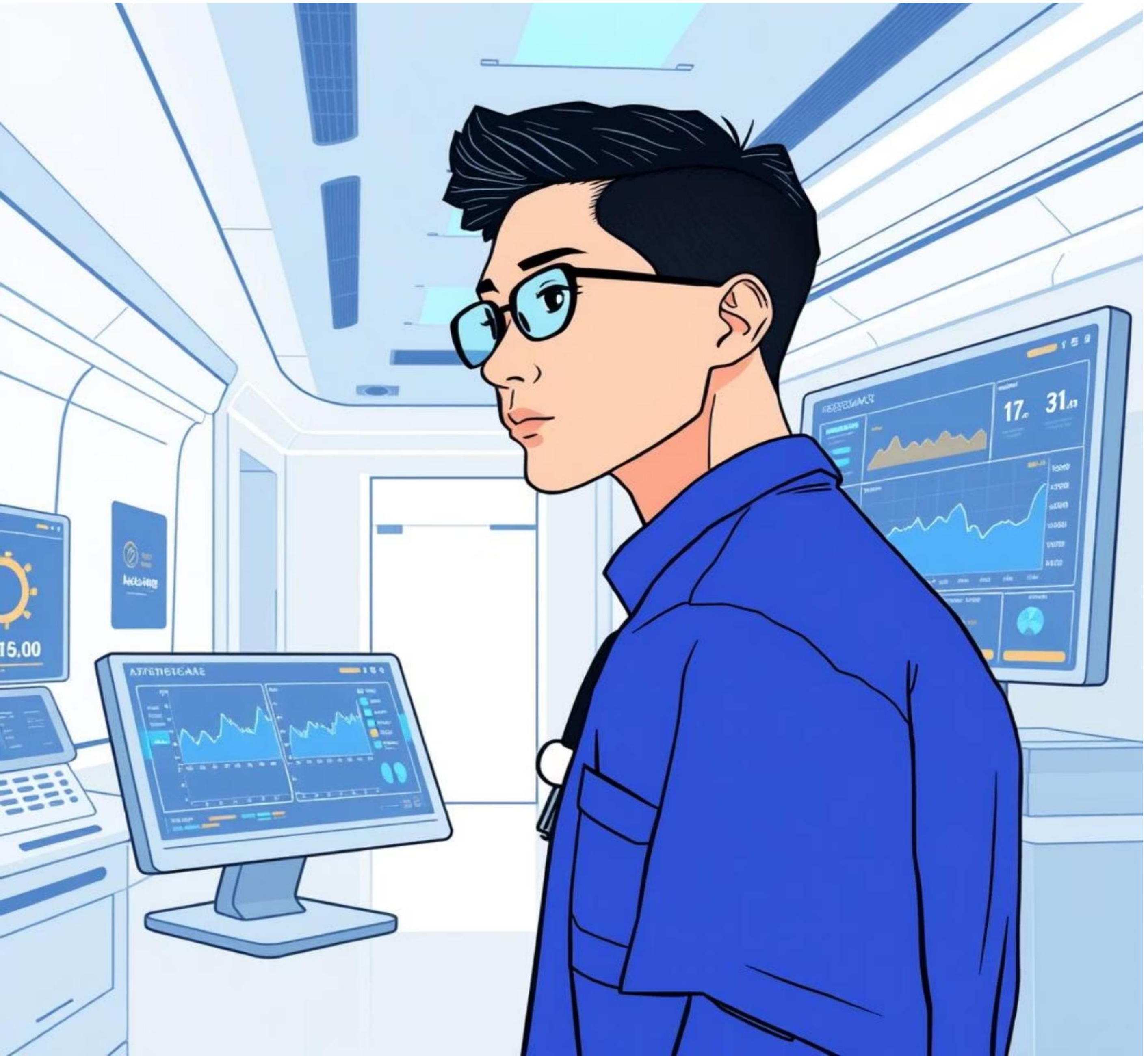
Our comparative analysis clearly shows the superior performance of the Random Forest model, especially concerning its ability to correctly identify true positive cases.

Random Forest	0.995	1.000	0.988	0.994	1.000
Logistic Regression	0.970	0.955	0.977	0.966	0.998
SVM	0.960	0.953	0.953	0.953	0.994
KNN	0.775	0.836	0.593	0.694	0.908

Final Model Selection: The Random Forest model was chosen as the optimal solution due to its perfect ROC-AUC score of 1.0. This indicates an exceptional ability to differentiate between cancer-positive and cancer-negative cases, minimizing critical misclassifications.

Conclusion & Future Work

This project has successfully delivered a highly accurate and reliable model for cancer prediction, with the Random Forest classifier standing out as the top performer. Our next steps focus on real-world implementation and continuous improvement.



Project Conclusion

We have validated the effectiveness of machine learning in early cancer detection. The selected Random Forest model offers a robust and accurate tool, paving the way for improved patient outcomes.

Potential Future Work

- **Deployment:** Transition the model into a production environment via an accessible API or dedicated web application for clinicians.
- **Continuous Monitoring:** Implement automated systems to track model performance, data drift, and prediction accuracy in real-time, ensuring long-term reliability.
- **Feedback Loop Integration:** Establish channels for direct feedback from medical professionals to refine the model, incorporate new data, and enhance prediction capabilities.

Questions & Discussion

We appreciate your attention to this critical initiative. We are now open for any questions or discussions you may have regarding our project, methodology, or findings.

Thank You

Thank you for your time and attention to our presentation on Cancer Detection Using Machine Learning. Your engagement is vital as we push the boundaries of medical AI.