Imarticus

alva.healthtech@gmail.com

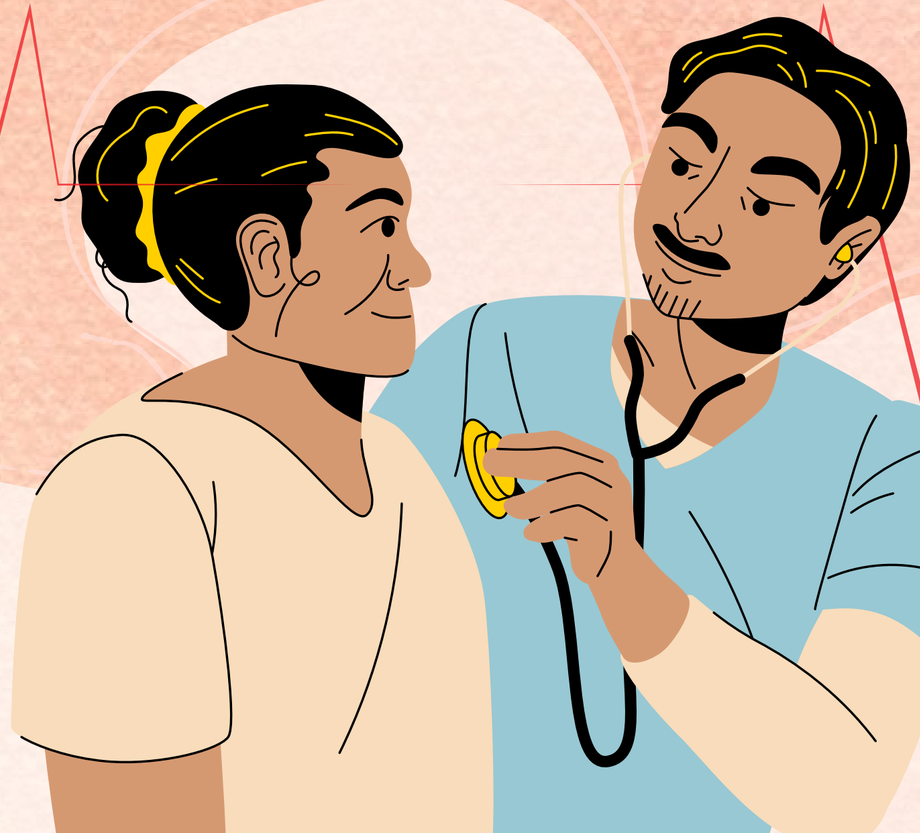
123-456-7890

ALVA

LATIN FOR "LIFE"

EMPOWERING BREAST CANCER PREDICTION

Through cell Nuclei detection and classification



Presented By :

Ishika Bhatia

Imarticus
alva.healthtech@gmail.com
123-456-7890

MOTIVATION & PROBLEM STATEMENT

EARLY DETECTION OF BREAST CANCER WITH MACHINE LEARNING

- Breast Cancer is the most commonly diagnosed cancer among women globally.

- Early and accurate diagnosis significantly improves survival rates
- However, interpreting cytological data (eg. cell nuclei measurement) can be time-consuming and error-prone

- Goal: Build a machine learning-powered diagnostic tool that predicts whether a breast mass is benign or malignant, based on cell nucleus features.

Imarticus
alva.healthtech@gmail.com
123-456-7890

DATA & PREPROCESSING

- Source: Wisconsin Breast Cancer Dataset

- Total Samples: 569
- Features : 30 numeric predictors + 1 target

- Target Classes: Malignant (M) = 1,
  Benign (B) = 0

- Features grouped into:
  - Mean values (eg - radius_mean)
  - Standard Error (eg - perimeter_se)
  - Worst values (eg - smoothness_worst)

- Preprocessing Steps:
  - Dropped irrevelant columns
    - id, unnamed index
  - Label encoding for diagnosis column
  - StandardScaler applied for model compatibility

alva.healthtech@gmail.com

123-456-7890

MODEL ARCHITECTURE & TRAINING

LOGISTIC REGRESSION MODEL

- Model: Logistic Regression
  - Chosen for interpretability and performance on the classification
- Scaled Features using StandardScaler
- Data split: 80% train, 20% test
- Model serialized with Pickle for reuse in app/web
    - Tried different models

- Performance:
  - Accuracy: ~99% on test set
  - Classification Report:
    - Precision, Recall
    - F1-score > 0.98 for both classes

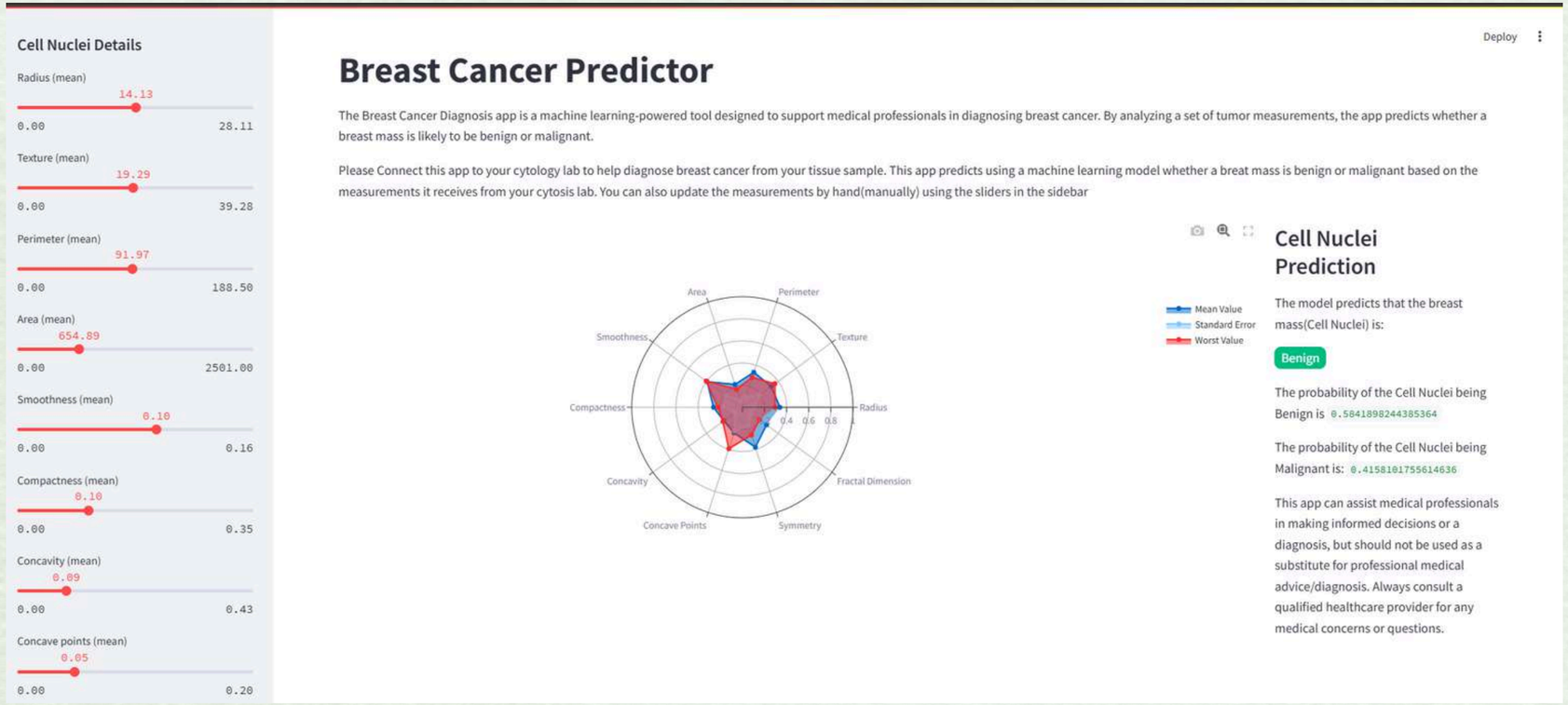
```
Model Comparison:
Logistic Regression: 0.9807 (+/- 0.0065)
SVM: 0.9736 (+/- 0.0147)
Random Forest: 0.9596 (+/- 0.0189)
KNN: 0.9649 (+/- 0.0096)
XGBoost: 0.9701 (+/- 0.0119)

Tuning Logistic Regression...
Tuned Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99        71
           1       1.00      0.98      0.99        43

    accuracy                           0.99       114
   macro avg       0.99      0.99      0.99       114
weighted avg       0.99      0.99      0.99       114

Accuracy: 0.9912280701754386
```

Imarticus

alva.healthtech@gmail.com

123-456-7890

STREAMLIT APP OVERVIEW



alva.healthtech@gmail.com

123-456-7890

MODELS EVALUATED

- Logistic Regression – Ideal for binary classification, interpretable, and fast.
- Support Vector Machine (SVM) – Effective in high-dimensional spaces.
- Random Forest – Ensemble of decision trees, good at handling complex patterns.
- K-Nearest Neighbors (KNN) – Predicts based on closest neighbors, simple but effective.
- XGBoost – Advanced boosting method, learns from past errors for high accuracy.

alva.healthtech@gmail.com

123-456-7890

CROSS-VALIDATION RESULTS

- Logistic Regression: 0.9807
- SVM: 0.9736
- Random Forest: 0.9596
- KNN: 0.9649
- XGBoost: 0.9701

Model Tuning:

- Used GridSearchCV to optimize Logistic Regression.

Post-Tuning Performance (on test data):

- Accuracy: 99.12%
- Precision, Recall, F1-score: All ≈ 0.99 for both classes (malignant & benign)

Imarticus
alva.healthtech@gmail.com
123-456-7890

FINAL CONCLUSION

- Several models were tested, but Logistic Regression gave the best combination of performance, simplicity, and interpretability.
- After tuning, it achieved ~99% accuracy, which is highly acceptable in the healthcare domain.
- Almost perfect balance between catching all malignant cases (recall) and avoiding false positives (precision).

Imarticus

alva.healthtech@gmail.com

123-456-7890

THANK YOU

Q AND A SESSION

Presented By :

Ishika Bhatia