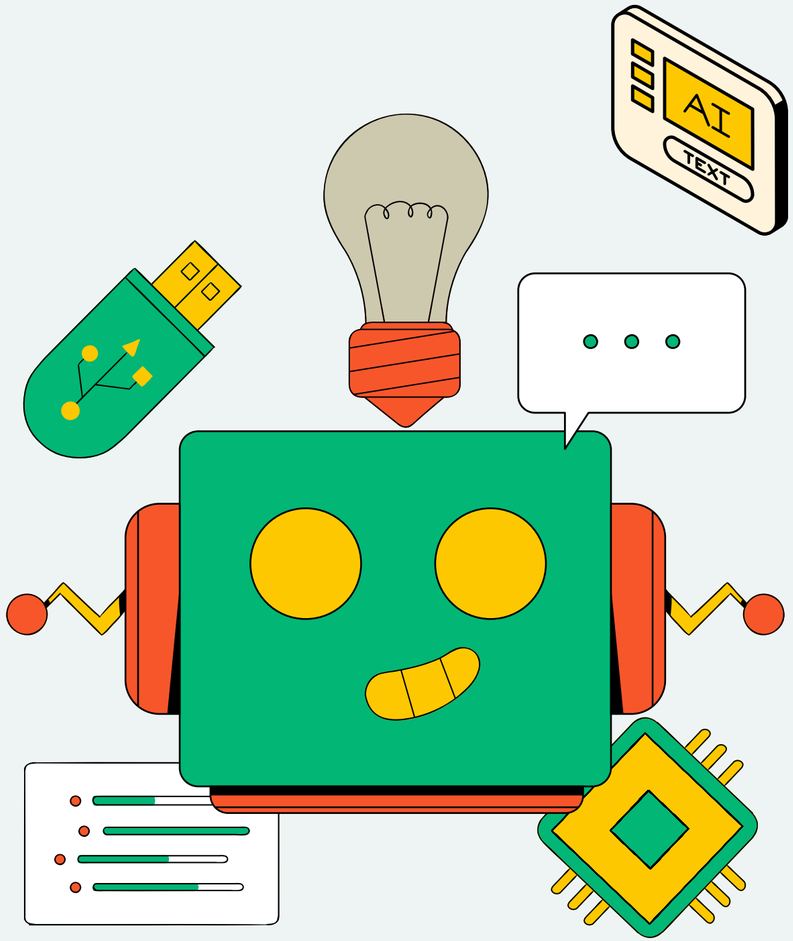


ML-BREAST-CANCER-CLASSIFIER

PRESENTATION

PRESENTED BY:

SABUHI GASIMZADA
SHAMIL HUSEYNOV
FARID AHMADZADEH
TELMAN MAMMADOV
ELGUN PASHAYEV
AYKHAN MOHSUMOV

INTRODUCTION

Objective

- To predict whether a tumor is benign or malignant using machine learning models.

Dataset:

- Source: Breast Cancer Wisconsin Dataset
- 569 samples, 32 columns (30 features, 1 target column, 1 ID column)
- Target: Diagnosis (Benign = 0, Malignant = 1)

Tools Used:

- Python, Scikit-learn, Matplotlib, SHAP, SMOTE

DATA PREPROCESSING

Steps:

1. **Removed irrelevant column: ID**
2. **Encoded categorical target variable: Benign (B) = 0, Malignant (M) = 1**
3. **Handled missing values: None were present.**
4. **Scaled numeric features using StandardScaler.**
5. **Balanced the dataset using SMOTE for minority class oversampling.**

Dataset Overview:

- 30 features: Mean, Standard Error, Worst-case metrics.
- Diagnosis Distribution: Balanced after SMOTE.

FEATURE SELECTION

Techniques Applied:

- Recursive Feature Elimination (RFE): Selected top 10 features.
  - Examples: Perimeter Mean, Area Mean, Concavity Mean.
- Mutual Information Scores: Top features with highest predictive power.

Key Insight: Using fewer features can maintain or improve performance by reducing noise.

MODELS IMPLEMENTED

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost
- LightGBM
- Neural Network (MLP)

Evaluation Metrics:

Accuracy

Precision, Recall, F1-score

ROC-AUC Score

LOGISTIC REGRESSION RESULTS

Performance:

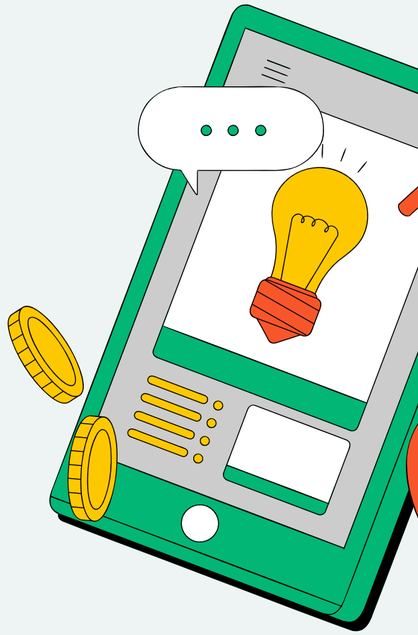
- **Accuracy: 0.99**
- **Precision: 0.99 (Benign), 1.00 (Malignant)**
- **ROC-AUC: 0.985**

Hyperparameters Tuned:

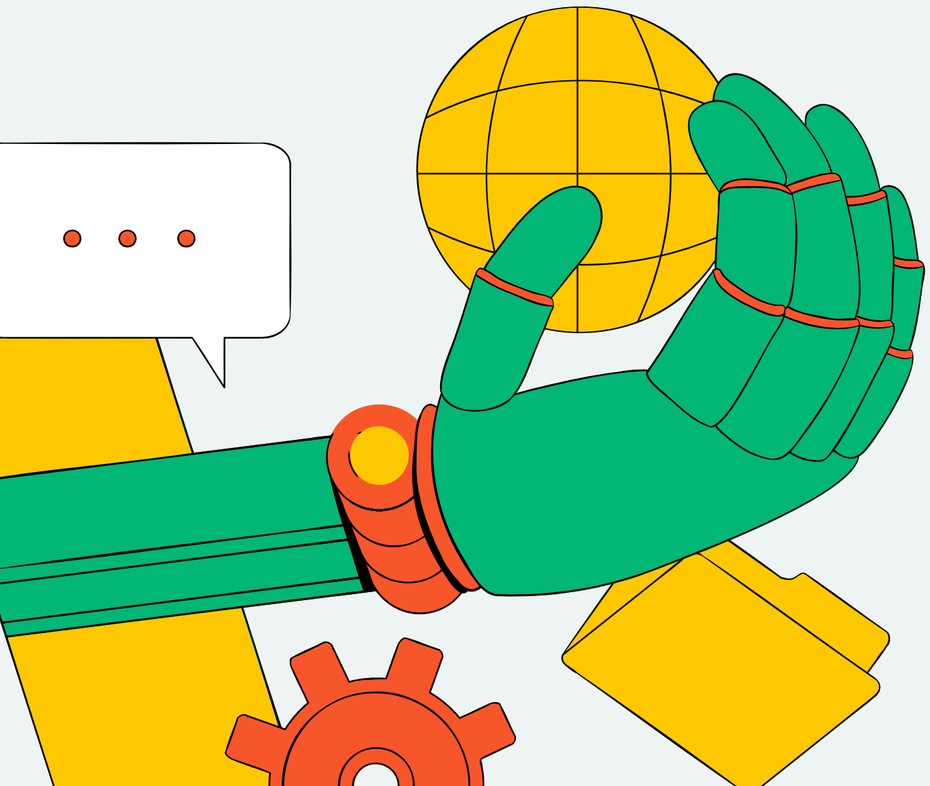
- **Regularization Strength (C): 0.1**
- **Solver: liblinear**

Visualization:

- **Confusion Matrix**
- **ROC Curve**



DECISION TREE RESULTS



Hyperparameters Tuned:

- Max Depth: 4
- CCP Alpha: 0.01

Performance:

- Accuracy: 0.95
- Precision: 0.93 (Malignant)
- ROC-AUC: 0.94

Visualization:

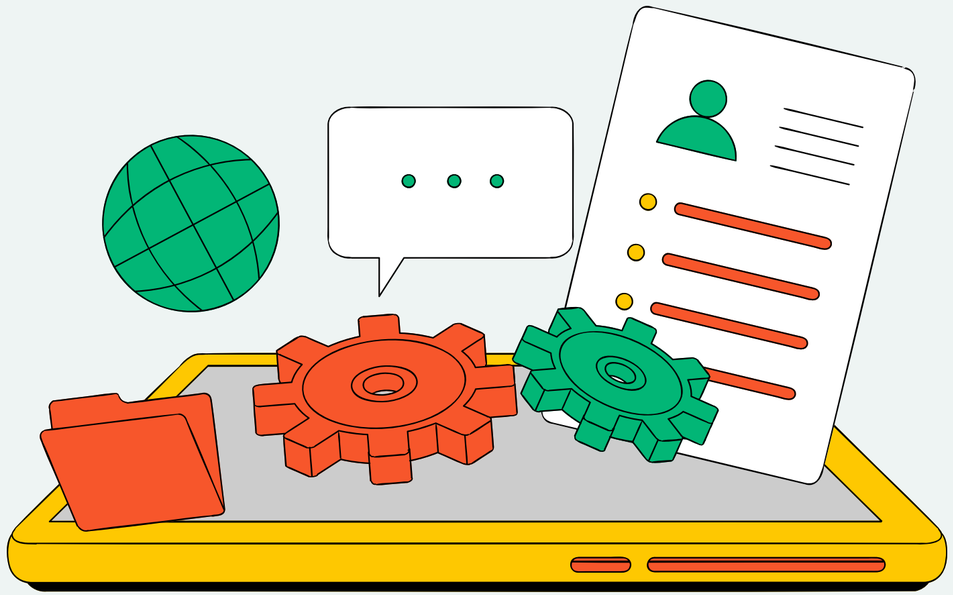
- Confusion Matrix
- Feature Importance Bar Chart

RANDOM FOREST WITH SMOTE

Balanced Dataset: Applied SMOTE to address class imbalance.

Performance:
Accuracy: 0.98
Precision: 0.98
Recall: 0.97

Visualization:
Confusion Matrix
ROC Curve



XGBOOST AND LIGHTGBM RESULTS

Hyperparameters Tuned (XGBoost):

- **Max Depth: 5**
- **Learning Rate: 0.1**

Performance (XGBoost):

- **Accuracy: 0.97**
- **ROC-AUC: 0.98**

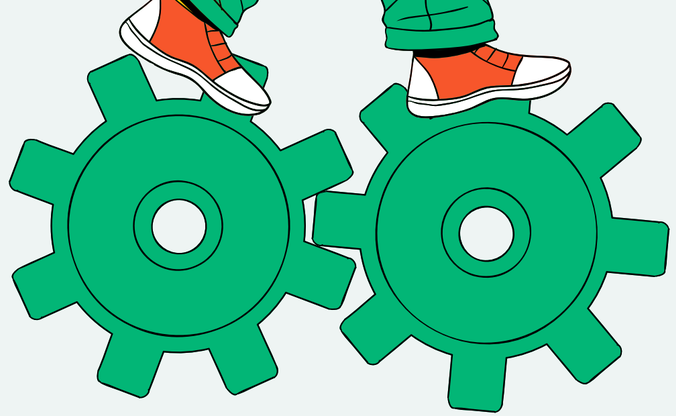
LightGBM:

- **Similar performance, faster training.**

Visualization:

- **Confusion Matrix**
- **SHAP Summary Plot**

NEURAL NETWORK RESULTS



ARCHITECTURE:

Hidden Layers: (128, 64, 32)
Activation: ReLU
Optimizer: Adam

PERFORMANCE:

Accuracy: 0.98
Precision: 0.98
Recall: 0.97

Neural Network Results:

	Precision	Recall	F1-Score	Support
Class 0 (Benign)	0.98	0.98	0.98	108
Class 1 (Malignant)	0.97	0.97	0.97	63
Accuracy:			0.98	171
Macro Average:	0.97	0.97	0.97	171
Weighted Average:	0.98	0.98	0.98	171

COMPARATIVE ANALYSIS AND KEY TAKEAWAYS

Model	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	0.99	0.99	0.98	0.99
Decision Tree	0.95	0.94	0.93	0.94
Random Forest	0.98	0.98	0.97	0.98
XGBoost	0.97	0.97	0.96	0.98
LightGBM	0.97	0.97	0.96	0.97
Neural Network	0.98	0.98	0.97	0.98

Key Takeaways

1. **Logistic Regression: Best for interpretability and simplicity.**
2. **Random Forest & XGBoost: Excellent balance of accuracy and robustness.**
3. **Neural Networks: Slightly better but computationally expensive.**

Future Work:

- **Explore more complex ensembles.**
- **Incorporate external datasets.**

QUESTIONS AND ANSWERS

For questions or suggestions about the project, you can send feedback directly to the project itself from github and also via email.

https://github.com/s3bu7i/ML-Breast-Cancer-Classifier

+994(50)8816613

sabuhi.gasimzada@gmail.com