PREDICTION MODELS IN CYBERSECURITY AND HEALTHCARE

Assignment Project Presentation

By Nikunj Gupta

PRESENTATION AGENDA

- Overview of Prediction Models
  - Healthcare Prediction Models
  - Cybersecurity Prediction Models
  - Algorithm Analysis and Comparison
  - Input/Output Specifications
  - Implementation and Training Process
  - Model Performance and Results
  - Conclusions and Future Work

HEALTHCARE PREDICTION MODELS

Disease Risk Prediction

- Algorithms: Random Forest, Logistic Regression, XGBoost

Patient Readmission Forecasting

- Algorithms: Random Forest, SVM, Neural Networks

Mortality Prediction

- Algorithms: Cox Regression, Random Forest, Deep Learning

Treatment Response Analysis

- Algorithms: Ensemble Methods, Gradient Boosting

CYBERSECURITY PREDICTION MODELS

Intrusion Detection Systems (IDS)

- Algorithms: Random Forest, SVM, Neural Networks, Ensemble Methods

Malware Detection

- Algorithms: CNN, Random Forest, Deep Learning

Network Anomaly Detection

- Algorithms: Isolation Forest, Autoencoders, LSTM, Clustering

Phishing Identification

- Algorithms: Logistic Regression, SVM, Random Forest

COMMON ALGORITHMS ACROSS DOMAINS

Random Forest

- Most widely adopted approach
- Handles missing data, provides feature importance

Neural Networks & Deep Learning

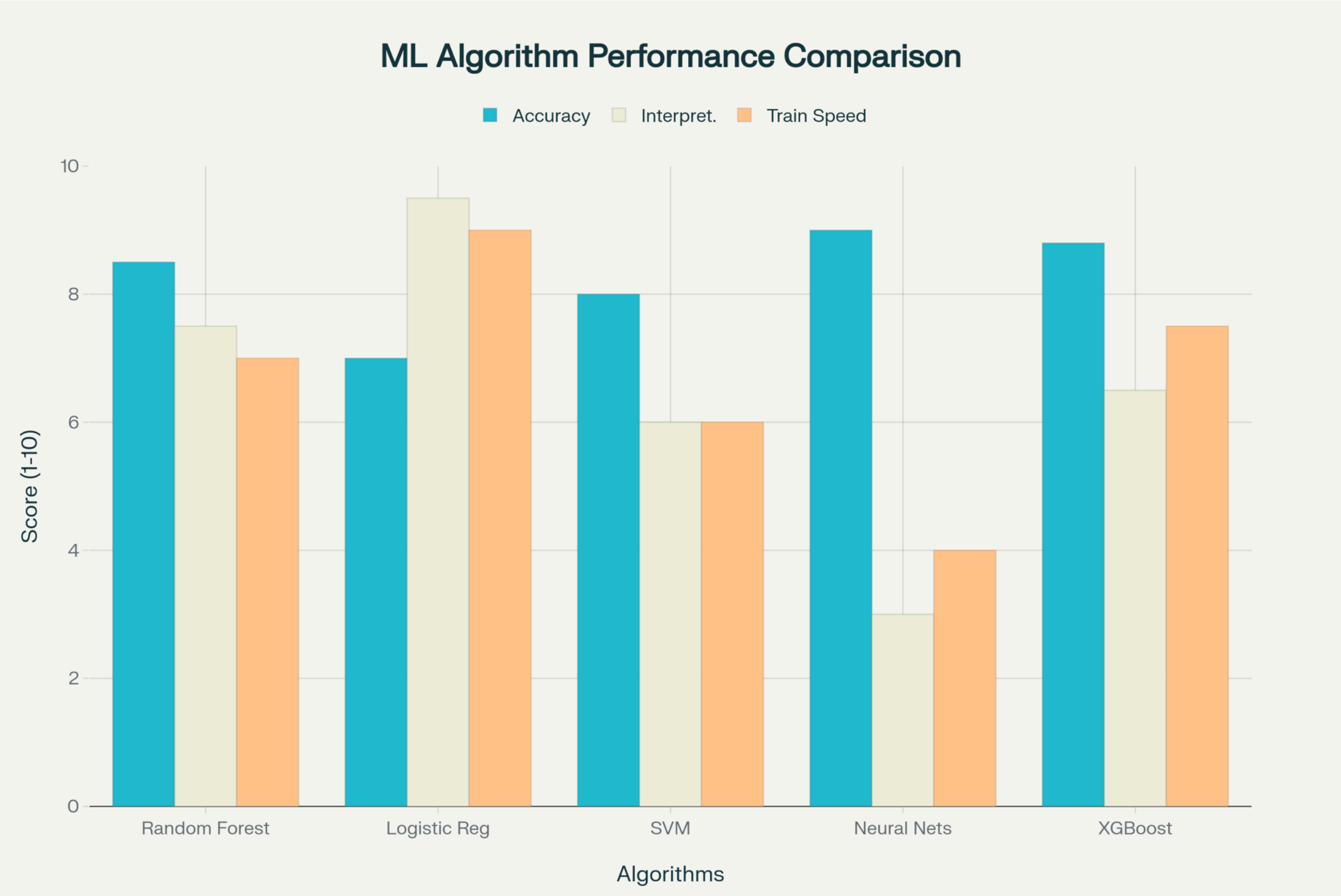
- Exceptional performance in complex pattern recognition

Support Vector Machines

- Excel in high-dimensional data scenarios

Logistic Regression

- Interpretable with probabilistic outputs


ML Algorithm Performance Comparison
Accuracy
Interpret.
Train Speed
Score (1-10)
10
8
6
4
2
0
Random Forest
Logistic Reg
SVM
Neural Nets
XGBoost
Algorithms

HEALTHCARE MODEL: INPUT & OUTPUT

Input Features:

- Patient Demographics: Age (18-90), Gender, Blood Type
- Medical Data: Diabetes, Hypertension, Asthma, Test Results
- Administrative: Admission Type, Insurance, Billing ($1,000-$50,000)
- Hospital Stay Duration: 1-30 days

Output:

- Binary Classification: High Risk (1) or Low Risk (0)
- Probability Scores: 0.0 to 1.0 likelihood of adverse outcomes

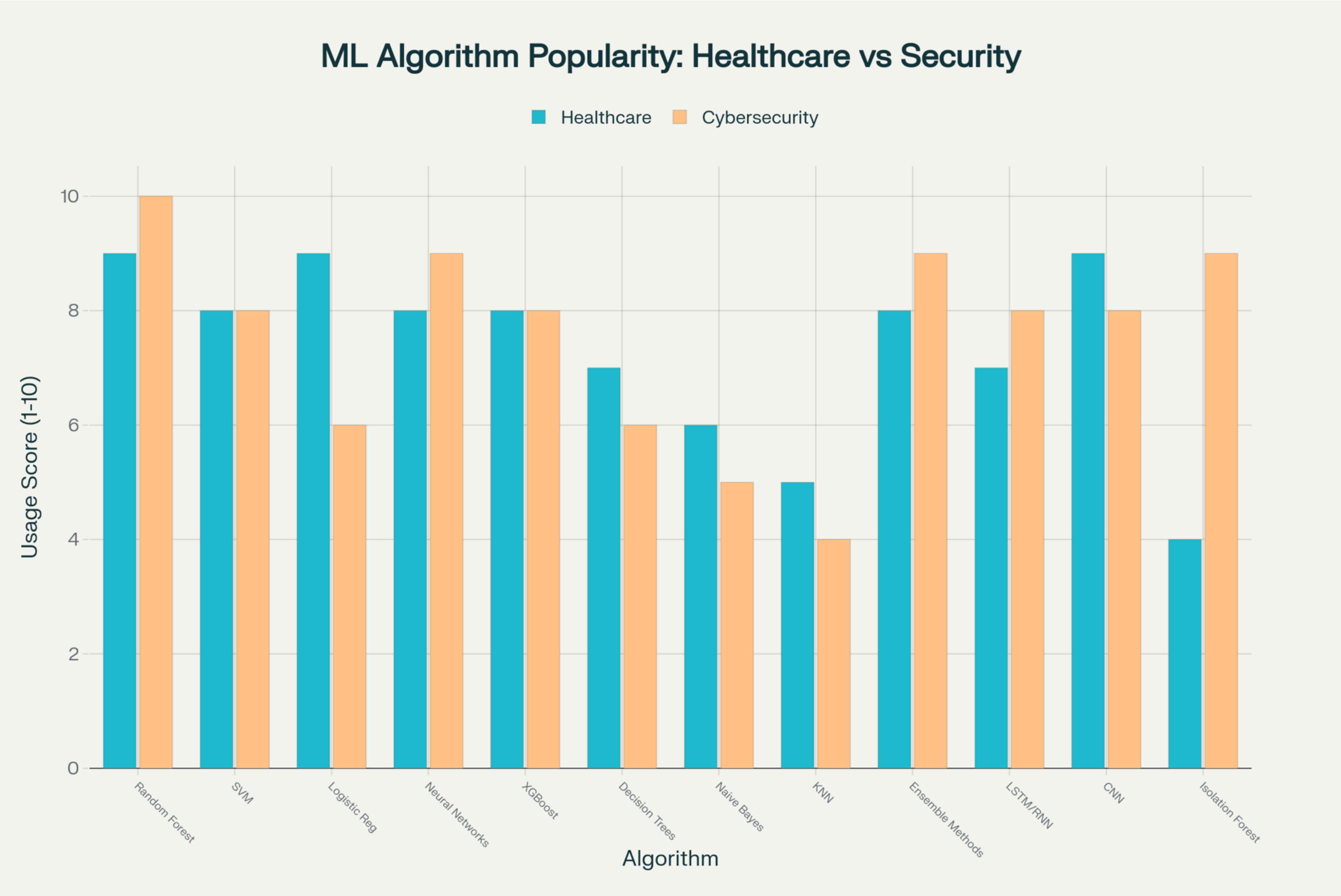
CYBERSECURITY MODEL: INPUT & OUTPUT

Input Features:

- Network Traffic: Flow duration, bytes/second, packets/second
- Flow Characteristics: Forward/backward packet counts, total lengths
- Packet Analysis: IAT means, PSH/URG flag counts, average sizes

Output:

- Binary Classification: Attack (1) or Benign (0)
- Probability Scores: 0.0 to 1.0 likelihood of malicious activity


ML Algorithm Popularity: Healthcare vs Security
Healthcare
Cybersecurity
Usage Score (1-10)
0
2
4
6
8
10
Random Forest
SVM
Logistic Reg
Neural Networks
XGBoost
Decision Trees
Naive Bayes
KNN
Ensemble Methods
LSTM/RNN
CNN
Isolation Forest
Algorithm

IMPLEMENTATION DETAILS

Algorithm Selection:

- Random Forest Classifier with 100 estimators
- Fixed random state (42) for reproducible results

Data Preprocessing Pipeline:

- Healthcare: Label Encoding, StandardScaler, SimpleImputer
- Cybersecurity: StandardScaler normalization only

Feature Engineering:

- Feature importance rankings for interpretability

TRAINING PROCESS & METHODOLOGY

Data Generation:

- Synthetic data with configurable parameters
- Realistic patient profiles and network traffic patterns

Training Workflow:

- 80-20 data split with stratification
- Cross-validation for model stability assessment
- Comprehensive metrics calculation
- Real-time prediction interface

TECHNICAL IMPLEMENTATION

- The project includes a comprehensive Streamlit web application featuring:
- User Interface: Intuitive usage pages for both healthcare and cybersecurity models
- Data Synthesis: Configurable custom synthetic data generation for training and testing
- Model Training: Automated training pipeline with one click training
- Results Visualization: Performance metrics, confusion matrices, and feature importance plots
- Prediction Interface: Real-time prediction capabilities with user input forms

MODEL PERFORMANCE & EVALUATION

- Evaluation Metrics Framework:
  - Accuracy: Overall correctness of predictions
  - AUC: Discrimination ability across thresholds
  - Precision: Quality of positive predictions
  - Recall: Completeness of positive case identification
  - F1-Score: Balanced measure of precision and recall
  - Confusion Matrix: Detailed classification breakdown

Metric	Description	Healthcare Application	Cybersecurity Application
Accuracy	Percentage of correct predictions out of total predictions	Overall accuracy in predicting patient risk levels	Overall accuracy in detecting attacks vs benign traffic
AUC Score	Area Under ROC Curve - measures discrimination ability	How well the model distinguishes high-risk from low-risk patients	How well the model distinguishes attacks from normal traffic
Precision	True Positives / (True Positives + False Positives)	Of patients predicted as high-risk, how many actually are high-risk	Of traffic predicted as attacks, how much is actually malicious
Recall	True Positives / (True Positives + False Negatives)	Of all actual high-risk patients, how many are correctly identified	Of all actual attacks, how many are correctly detected
F1-Score	Harmonic mean of Precision and Recall	Balance between identifying high-risk patients and avoiding false alarms	Balance between detecting attacks and minimizing false positives
Confusion Matrix	2x2 matrix showing TP, TN, FP, FN counts	Shows correct vs incorrect risk classifications	Shows correct vs incorrect attack/benign classifications
Classification Report	Detailed per-class precision, recall, F1-score	Detailed performance for high-risk and low-risk classes	Detailed performance for attack and benign classes
Feature Importance	Ranking of features by their contribution to predictions	Which patient factors most influence risk predictions	Which network features most indicate malicious activity
Cross-Validation	K-fold cross-validation capability built-in	Validates model stability across different patient populations	Validates model stability across different network conditions
Model Interpretability	Feature importance provides model explainability	Clinicians can understand why a patient is flagged as high-risk	Security analysts can understand why traffic is flagged as malicious

RESULTS & CODE EXECUTION

Implementation Architecture:

- Streamlit web application with intuitive interfaces
- Professional-grade visualization components

Performance Outcomes:

- High accuracy suitable for critical decision-making
- Clear feature importance rankings for interpretability
- Scalable architecture for large-scale data processing
- Real-time prediction capabilities

Component	Description
Healthcare Prediction Model	Random Forest model for healthcare risk prediction
Cybersecurity Prediction Model	Random Forest model for cybersecurity threat detection
Features(columns) - Healthcare	Age, Gender, Medical Conditions, Billing, Hospital Days, Test Results, etc.
Features(columns) - Cybersecurity	Flow Duration, Packet Counts, Bytes, Network Traffic Features, etc.
Evaluation Metrics	Accuracy, AUC Score, Confusion Matrix, Classification Report
User Interface	Streamlit web application with multiple pages for easy usage
Training Process	Train-test split (80-20), StandardScaler, Label Encoding

KEY FINDINGS & CONCLUSIONS

Random Forest provides optimal balance:

- Accuracy, interpretability, computational efficiency
- Cross-domain applicability demonstrated by using same model for different purposes and getting optimal results

Critical Success Factors:

- Comprehensive evaluation metrics beyond accuracy
- Feature importance for model explainability
- User-friendly interfaces

THANK YOU

GitHub link to project file: https://github.com/Nikunj-Gupta-1/Model_trainer.git

To view the Web-app : https://nikunj-s--ml-trainermodel-final.streamlit.app