University of Engineering and Technology Lahore
Department of Computer Science
Muhammad Abubakar 2022-CS-188
Abdullah Azher Chaudhary 2022-CS-204

Comparative Analysis of Machine Learning Models for HIV Infection Prediction

HIV remains a global health challenge, affecting approximately 38.4 million individuals in 2022. Early diagnosis plays a critical role in controlling its spread and improving patient outcomes.Traditional diagnostic methods like RT-PCR are accurate but resource-intensive and inaccessible in low-resource settings.Machine Learning (ML) techniques can provide efficient and scalable solutions for early HIV prediction.

Research Objective

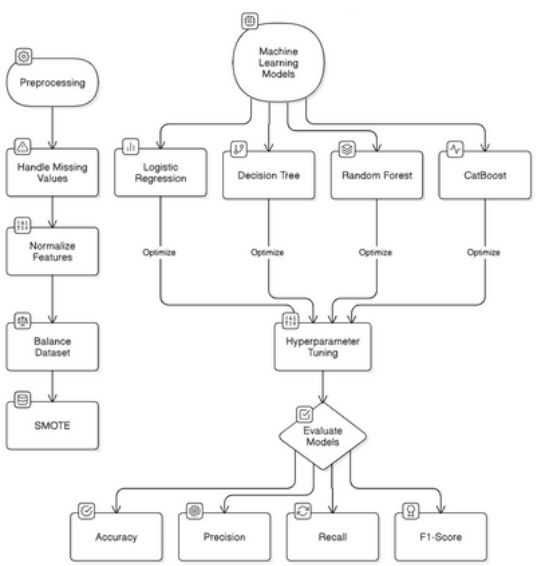
Evaluate the performance of multiple ML models for predicting HIV infection.Develop a robust preprocessing pipeline for handling class imbalance and scaling features.Identify critical factors influencing HIV prediction through feature importance analysis.

Key Concepts

- **HIV Prediction:** ML models offer efficient solutions for early diagnosis.
- **Class Imbalance:** Resolved using SMOTE for balanced training.
- **Critical Features:** Behavioral, demographic, and clinical attributes.
- **Best Model:** CatBoost achieved 90% accuracy with robust performance.
- **Optimization:** Optuna enhanced model tuning and efficiency.

Methodology

- **Preprocessing**:
- Addressed missing values and normalized numerical features using MinMaxScaler.
- Balanced the dataset using Synthetic Minority Oversampling Technique (SMOTE).
- **Machine Learning Models:**
- Logistic Regression, Decision Tree, Random Forest, CatBoost.
- Optimized using the Optuna framework for hyperparameter tuning.
- **Evaluation Metrics**: Accuracy, Precision, Recall, and F1-Score.



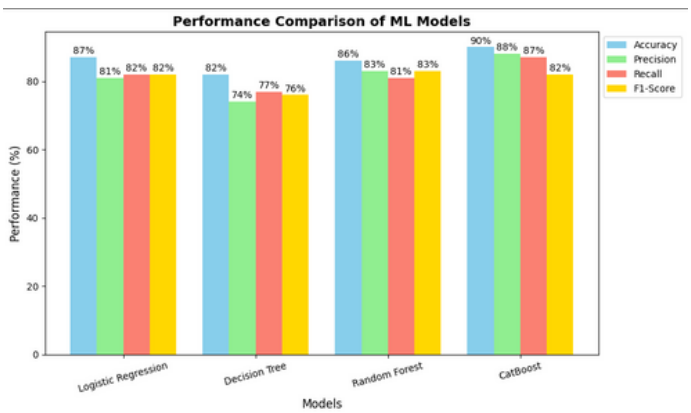
Dataset Details

- **Source:** AIDS Clinical Trials Group Study 175 dataset (1996).
- **Attributes:**
- **Demographic:** Age, weight, gender, race.
- **Behavioral:** IV drug use, sexual activity.
- **Clinical:** CD4/CD8 counts, Karnofsky score.
- **Treatment:** Antiretroviral therapy history.
- **Outcome Variable:** HIV infection status (0 = No, 1 = Yes).

Results

Best Model: CatBoost classifier outperformed all others.

- **Accuracy:** 90%
- **Precision:** 88%
- **Recall:** 87%
- **F1-Score:** 82%



Tools and Technologies

- Frameworks: Python (Pandas, Scikit-learn, CatBoost).
- Optimization: Optuna for hyperparameter tuning.
- Preprocessing: MinMaxScaler, SMOTE for class imbalance resolution.

Future Work

- Incorporate Temporal Data:
- Use time-series data for improved predictions.
- Expand Datasets for Diversity:
- Include diverse populations for broader applicability.
- Apply Advanced Deep Learning Models:
- Enhance accuracy with advanced models.
- Integrate an Empathetic Chatbot:
- Add a chatbot for personalized, empathetic support.
- Enable Real-time Feedback:
- Provide immediate insights with real-time processing.

Conclusion

Machine learning models, particularly CatBoost, demonstrate significant efficacy in predicting HIV infection, offering a powerful tool for early diagnosis and timely intervention. These advanced approaches enhance the accuracy of predictions, enabling healthcare providers to implement targeted public health strategies that reduce transmission rates and improve outcomes. By leveraging data-driven models, healthcare systems can overcome barriers in resource-limited settings, providing more equitable access to diagnosis and treatment. Furthermore, the use of these models supports continuous monitoring and adaptive interventions, ultimately leading to better long-term management of the HIV epidemic and improving overall public health infrastructure.