Predictive Modeling of HIV Diagnosis Rates Using Machine Learning Techniques

Rahul Jarupula, B. Pharm ♦ rjarupul@oswego.edu; Isabelle Bichindaritz, Ph. D ♦ ibichind@oswego.edu
Department of Biomedical & Health Informatics, SUNY Oswego, NY, USA.

Abstract

This project presents a comprehensive data preprocessing and regression modeling pipeline to predict the rate of new diagnoses using a large epidemiological dataset. The workflow involves data cleaning, transformation, encoding categorical features, and handling missing values. Exploratory data analysis techniques, including distribution plots and correlation heatmaps, were applied to understand feature relationships.

Several regression models—Linear Regression, Random Forest, Gradient Boosting, Decision Tree, K-Nearest Neighbors, and Polynomial Regression—were trained and evaluated based on MAE, RMSE, and R^2 scores. The results provide insights into the effectiveness of different models in predicting health-related trends and support informed decision-making in public health analytics.

Introduction

- According to AIDSVu, HIV continues to pose a serious public health concern in the United States, with thousands of new diagnoses reported annually, disproportionately affecting certain communities.[1]
- Traditional methods of assessing HIV risk often fail to capture complex patterns in demographic and socioeconomic factors that vary across regions.[2]
- By leveraging machine learning techniques on AIDSVu's publicly available datasets, we can build predictive models that identify populations at elevated risk, enabling more targeted and data-driven public health interventions.

Project Focus

Mission Statement

- **Enhancing prediction accuracy and decision-making through machine learning models applied to structured dataset analysis.**
- Traditional prediction systems often rely on limited features or static rule-based algorithms that may not generalize well across diverse datasets. In this study, we applied advanced machine learning techniques to predict a continuous best outcome.
- This model-driven approach demonstrates the potential of machine learning to enhance predictive analytics and support data-driven decisions in real-world applications such as resource planning, performance monitoring, and personalized recommendations.

Learning Objectives

- Given an HIV affected data by County but particularly predicting New Diagnosis Rate, using an algorithm generated by machine learning.
- **To evaluate the effectiveness of advanced algorithms** in improving predictive accuracy and supporting data-driven decision-making processes.

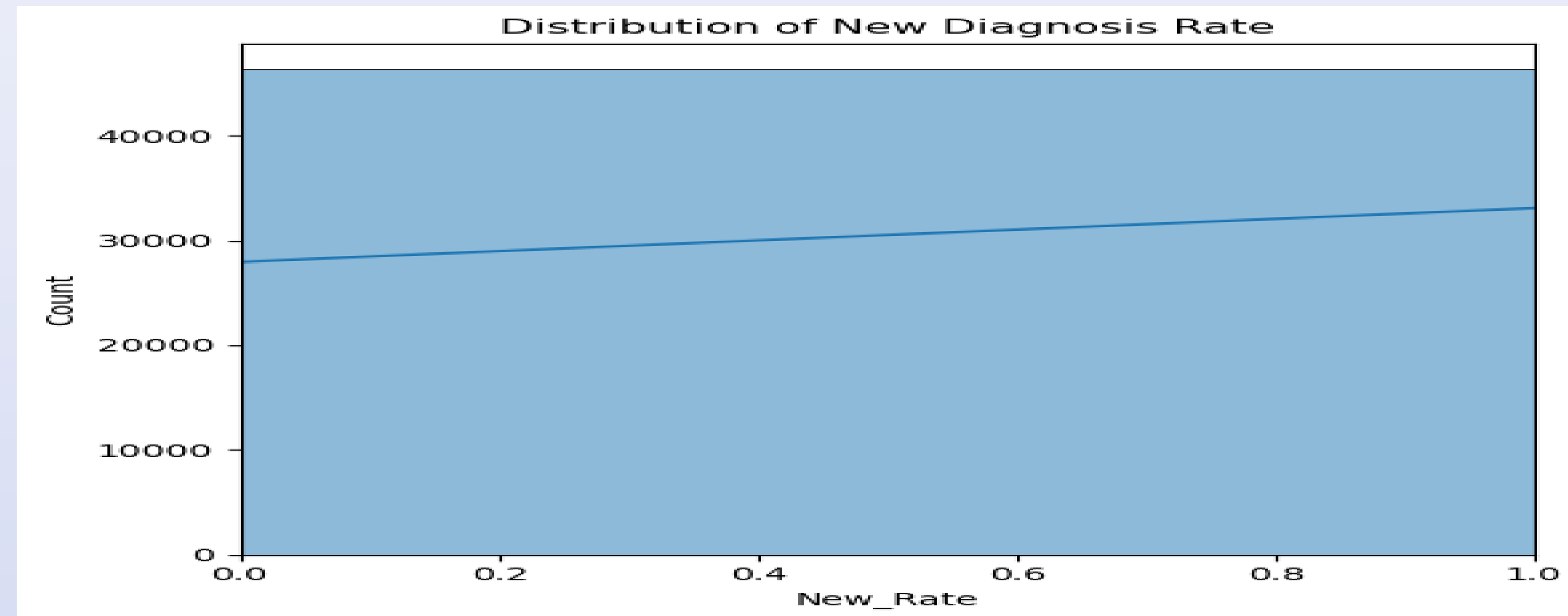
Methods

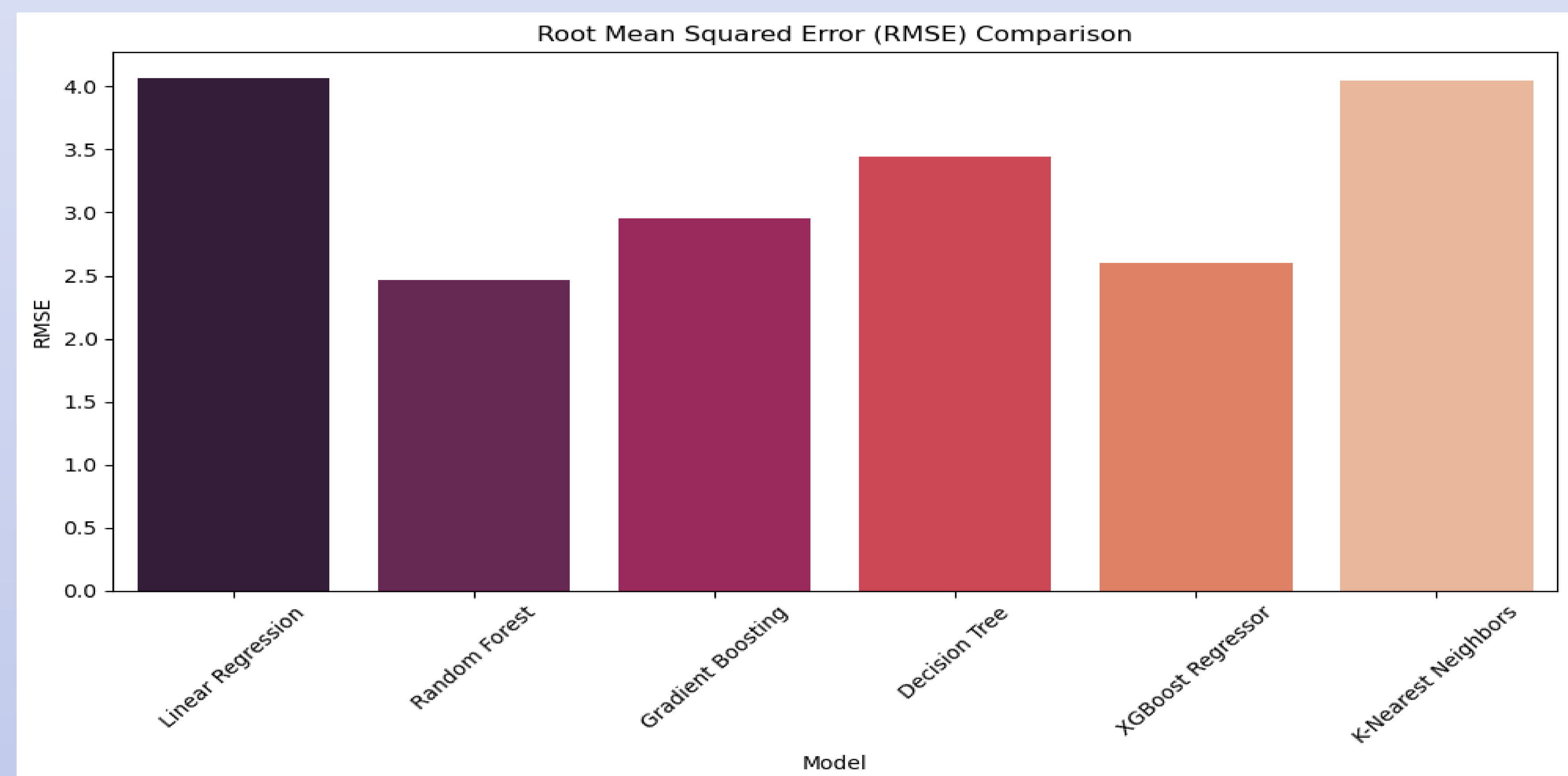
•A dataset containing multiple features relevant to the prediction task was used for this study.
•The target variable was continuous in nature, making this a regression problem.
•Data preprocessing was performed, including missing value handling, outlier removal using z-score methods, and feature scaling where necessary.
Feature selection and engineering were carried out to optimize model input and performance.
Models Implemented:
•Random Forest, Gradient Boost, Linear Regression, Decision Tree, XGBoost, and K Nearest Neighbors regressor Models.
•Each model was trained on the training set and evaluated on the test set.
•Evaluation was based on three metrics: **Mean Absolute Error**, **Root Mean Squared Error**, and **R-squared Score**.
•**Random Forest Regressor outperformed other models**.

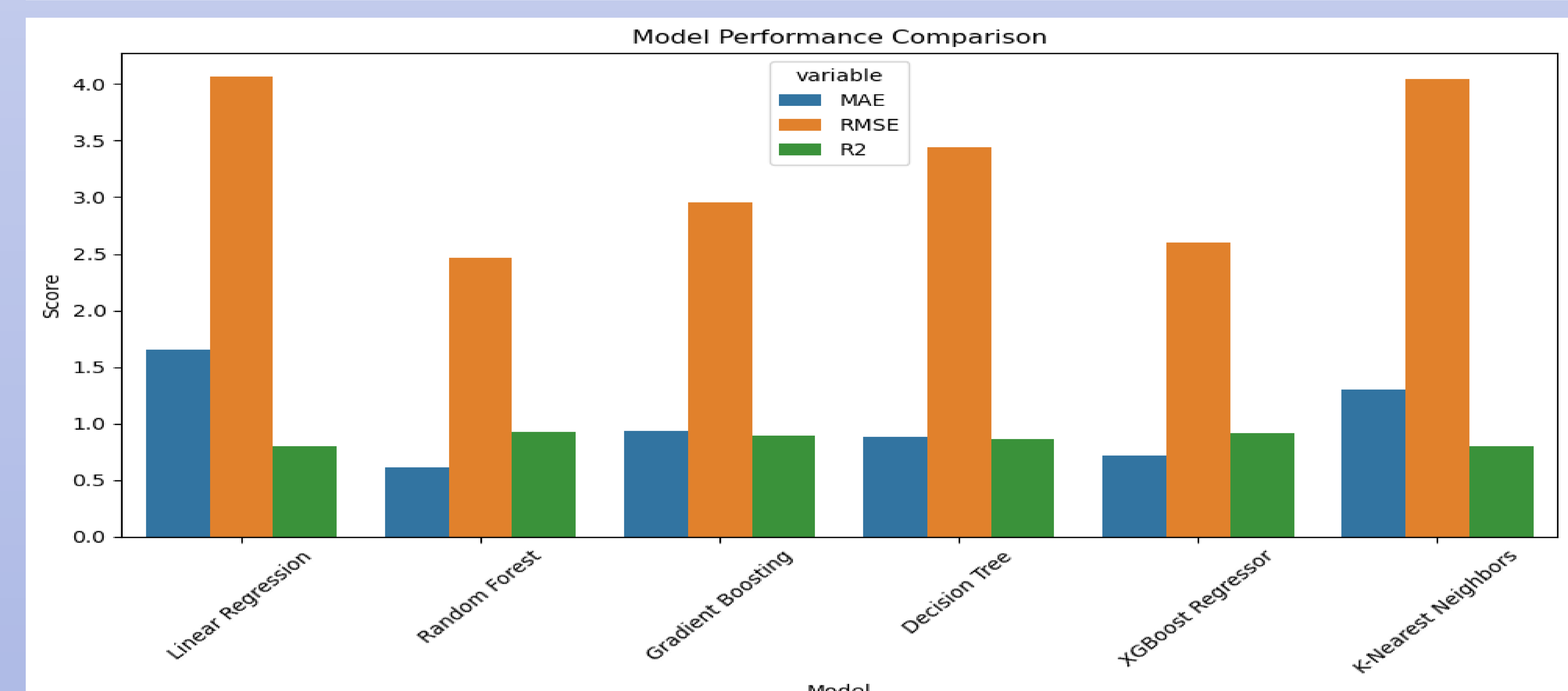
Future Direction

- According to AIDSVu, HIV continues to pose a serious public health concern in the United States, with thousands of new diagnoses reported annually, disproportionately affecting certain communities.[1]
- Traditional methods of assessing HIV risk often fail to capture complex patterns in demographic and socioeconomic factors that vary across regions.[2]
- By leveraging machine learning techniques on AIDSVu's publicly available datasets, we can build predictive models that identify populations at elevated risk, enabling more targeted and data-driven public health interventions.

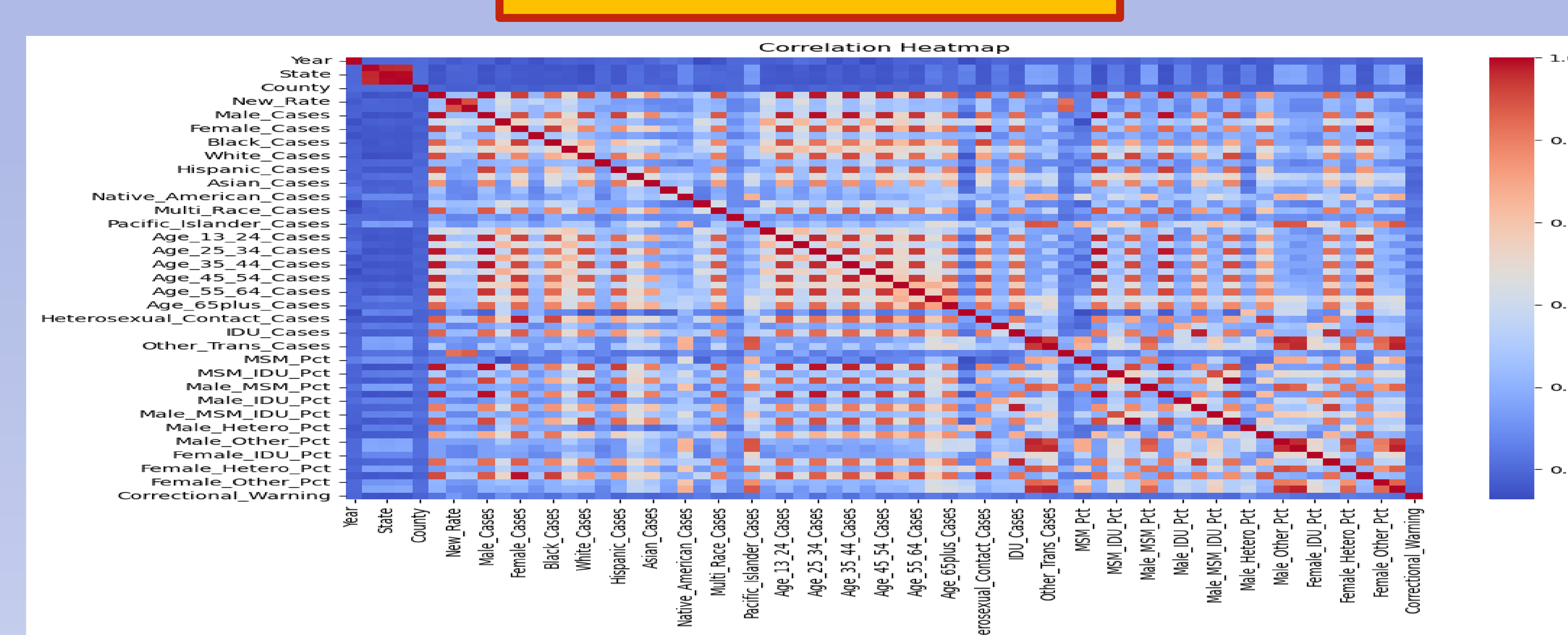
Results







Correlation Matrix



Results

- The Six machine learning models were evaluated to predict HIV diagnosis rates using 18 demographic and healthcare-related features.
- The **Random Forest** model achieved the best performance with an MAE of **0.61**, RMSE of **2.46**, and R^2 of **92.6%**, followed closely by XGBoost and Gradient Boosting with R^2 values above 89%.
- In contrast, **Linear Regression** and **K-Nearest Neighbors** showed higher error rates, with MAE values of 1.64 and 1.29 respectively, and R^2 scores around **80%**, indicating less predictive power.

Discussion

The Random Forest and XGBoost models demonstrated superior predictive accuracy, highlighting their effectiveness in modeling complex patterns in HIV diagnosis rates. The relatively lower performance of Linear Regression and KNN suggests these models may be less suitable for capturing nonlinear relationships in the dataset.

Conclusions

Based on the evaluation of multiple machine learning models, Random Forest Regressor emerged as the most effective technique for predicting HIV diagnosis rates, achieving high R^2 scores and low error metrics (MAE and RMSE).

These results indicate their strong capability in capturing the complex patterns and relationships within the data. Overall, advanced ensemble methods provide a promising approach for enhancing the accuracy of HIV-related public health forecasting and decision-making.

References

1) The Suthaharan S. Machine Learning Models for Predicting HIV/AIDS Risk Based on Socio-Demographic Data. In: *Machine Learning Models and Algorithms for Big Data Classification*. Springer; 2016. doi:10.1007/978-1-4899-7641-3_8

2) Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*. 2016;375:1216-1219. doi:10.1056/NEJMp1606181

3) Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 2016;785–794. doi:10.1145/2939672.2939785

4) Shi, M., Lin, J., Wei, W., Qin, Y., Meng, S., Chen, X., ... & Jiang, J. (2022). Machine learning-based in-hospital mortality prediction of HIV/AIDS patients with Talaromyces marneffei infection in Guangxi, China. *PLoS neglected tropical diseases*, *16*(5), e0010388.https://doi.org/10.1371/journal.pntd.0010388

Acknowledgement

- I thank Dr. Isabelle for her constant support for this project.