THE UNIVERSITY OF TEXAS AT ARLINGTON

Data-Driven Insights into Healthcare Inequality and Treatment Disparities

Alain Areeba Siddiqui, Diego Maldonado, Faizah Khan, Mariah N Cornelio, Akari Kojima

Division of Data Science, College of Science, The University of Texas at Arlington

healthverity

Background

- Healthcare inequity remains a major challenge affecting patient outcomes and access to quality care
- Historical and systemic disparities have led to unequal treatment across demographic groups, especially in cancer care

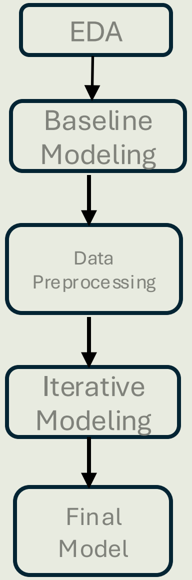
Objectives

- Develop and train machine learning models to predict whether patients receive a metastatic cancer diagnosis within 90 days of screening.
- Inform strategies for improving equity and fairness in diagnostic and treatment practices

Research Question

Can machine learning models identify whether patient demographics influence the likelihood of receiving a timely cancer diagnosis?

Methodology

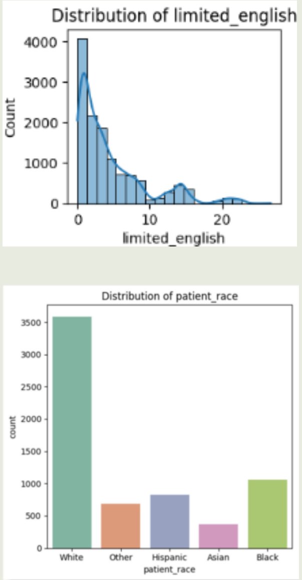

EDA

Baseline Modeling

Data Preprocessing

Iterative Modeling

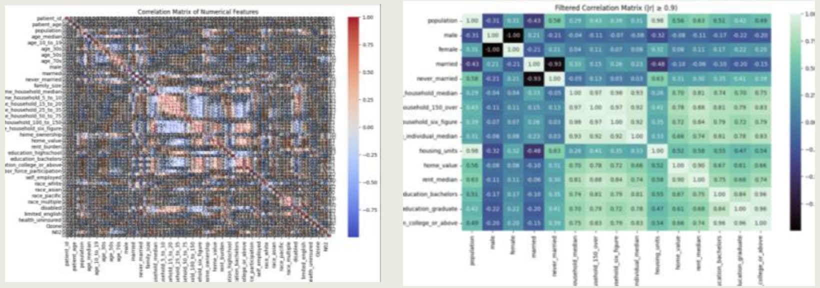
Final Model



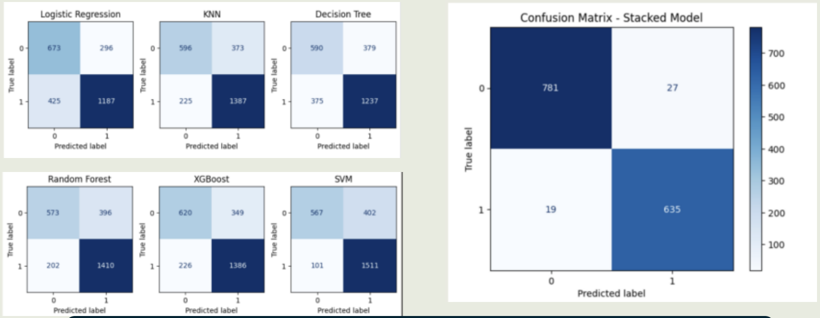
Distribution of limited_english variable and patient race

Results

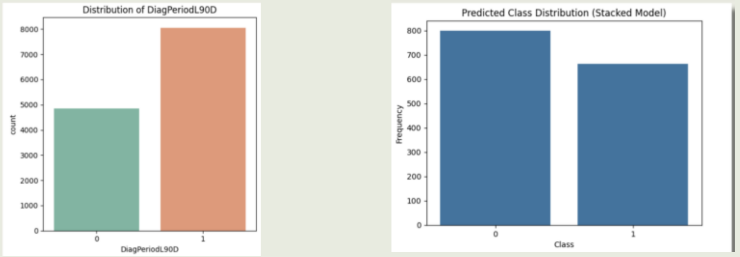
Before and After PCA: 82 features to 42



Baselines Models Ran vs Final Model Confusion Matrices



Before SMOTE vs After



Conclusions

By addressing class imbalance with SMOTE and stacking methods, we substantially enhanced model sensitivity. The final model achieved a 97% recall across both classes, indicating a highly reliable prediction of patient diagnosis outcomes.

Model	Class 0 Recall Score	Class 1 Recall Score
Logistic Regression	0.69	0.74
K-Nearest Neighbors (KNN)	0.62	0.86
Decision Tree	0.61	0.77
Random Forest	0.59	0.87
XGBoost	0.64	0.86
Support Vector Machine	0.59	0.94
Stacked Model	0.97	0.97

Future Works

Some future methods to use would be feature engineering, different versions of synthetic sampling methods, and using one-hot encoding to create more significant features and focus on adaptive learning based on different demographics for model deployment.

Model	Class 0 Recall Score	Class 1 Recall Score
Logistic Regression	0.69	0.74
K-Nearest Neighbors (KNN)	0.62	0.86
Decision Tree	0.61	0.77
Random Forest	0.59	0.87
XGBoost	0.64	0.86
Support Vector Machine	0.59	0.94
Stacked Model	0.97	0.97