Eliminating Health Disparities through Predictive Modelling in PySpark

IST 718 Big Data Analytics
Spring 2024 – Final Report

Group Number: 7

Somia Abdelrahman| Thomas Archibald| Rajat Chelani | Immanuel Odarteifio

1. Introduction

Health disparities in the diagnosis and treatment of diseases such as cancer are significant public health concerns. This project leverages Big Data analytics to understand and potentially reduce inequities in healthcare, particularly focusing on the timing of metastatic cancer diagnoses. The data seeks to isolate how demographic and environmental factors influence medical outcomes. Predictive modelling in PySpark offers a promising approach to identifying and addressing these disparities effectively.

2. Project Objectives

The main goal of this project is to develop a predictive model capable of accurately forecasting whether patients will receive a diagnosis of metastatic triple-negative breast cancer within 90 days. This type of cancer is particularly aggressive and demands prompt diagnosis and treatment, making early detection vital for effective intervention. With this analysis we hope to unveil any external factors that influence time to detection.

Secondary Objectives:

Feature Identification: To identify and analyse the key features that contribute to late diagnoses, and we are using 90 days as the window to categorize early or late. Understanding these factors is crucial in developing strategies to prevent delays in diagnosis and treatment.

Impact of Demographics and Environment: Examine how various demographic and environmental factors influence the timing of metastatic detection. This includes evaluating the effects of environmental pollutants on diagnosis and treatment outcomes.

Insurance Analysis: Develop specialized models to predict diagnosis rates among holders of commercial insurance policies, analysing how different types of health coverage affect diagnosis timeliness.

These objectives guide the research methodology and the analytical approach, aiming to not only predict early or late diagnosis but also to understand the underlying causes of delays in the diagnostic process. This dual focus enhances the project's potential to inform policy and healthcare practice improvements.

3. Data Description

The dataset consists of 12,906 patient records from 2015 to 2018, each later diagnosed with metastatic cancer. It includes:

Patient demographics: Age, gender, race, and state.

Health data: Breast cancer diagnosis codes and descriptions, metastatic cancer diagnosis codes.

Socioeconomic factors: Information by patient zip code, including income, education levels, and employment status.

Environmental data: Levels of pollutants such as ozone (O3), PM2.5, and nitrogen dioxide (NO2).

4. Methodology

The project follows a structured process flow, starting from data ingestion in PySpark to extensive data cleaning and transformation. Analytical methods employed include:

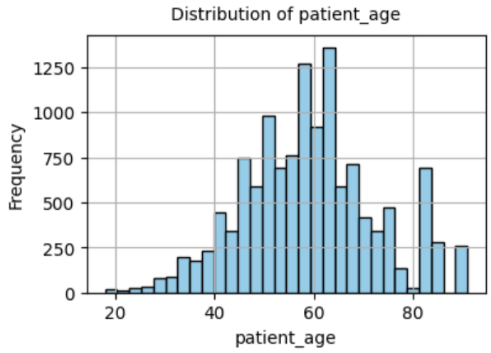
Model Training: Multiple models such as Random Forest, Gradient Boost Classifier, Decision Tree, and Logistic Regression were trained. Feature selection was performed to enhance model performance.

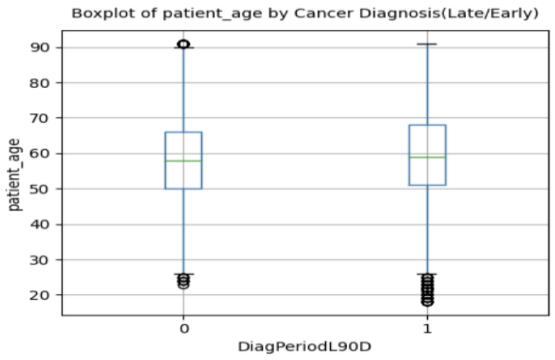
Model Validation: The models were validated using AUC as a performance metric to ensure the robustness of the predictions.

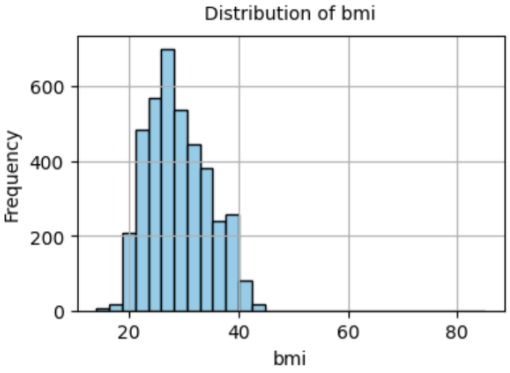
5. Exploratory Data Analysis

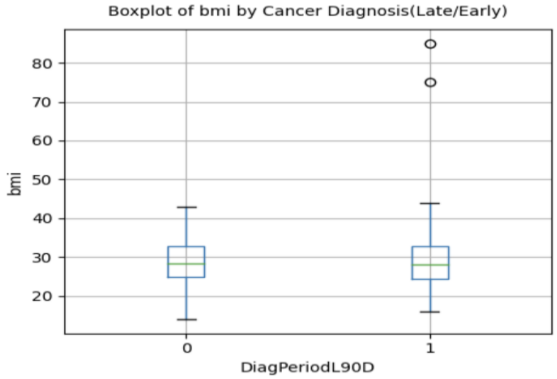
Age and BMI Distribution in Diagnosis Timing

Our analysis began with assessing how age and BMI impact the timing of cancer diagnoses. The data shows that the age distribution of patients diagnosed with metastatic cancer did not significantly differ between early and late diagnosis groups, with most patients falling within the 50-60 age range. Similarly, BMI distributions were consistently centred around the 20s to 30s range across both groups, indicating that neither age nor BMI significantly influences diagnosis timing on their own.





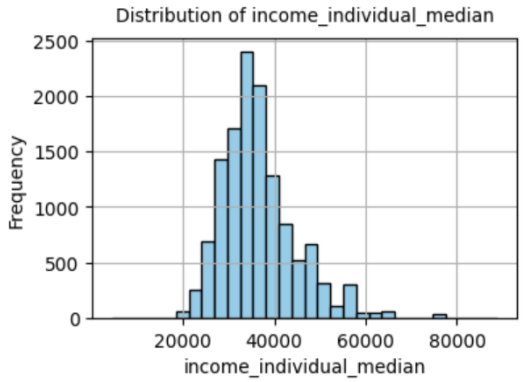


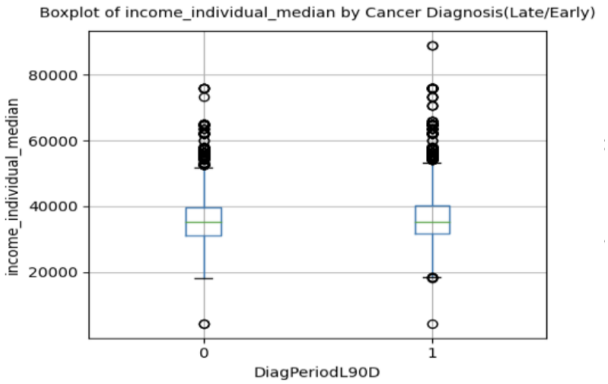


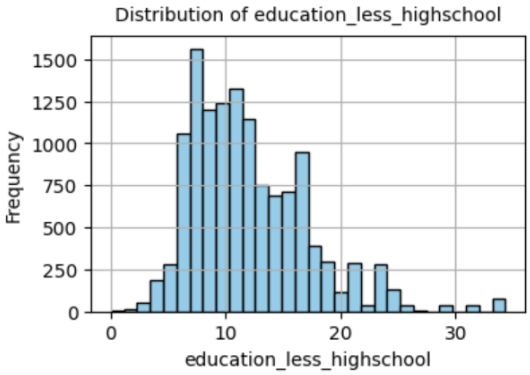
Impact of Socio-Economic Factors

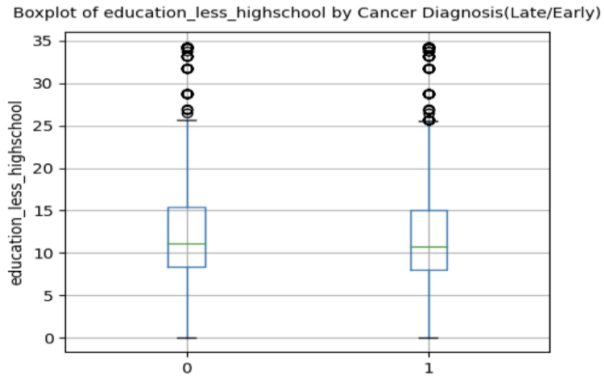
The analysis explored how socio-economic factors, such as income levels, educational attainment, and employment status, influence the timing of cancer diagnoses. Data grouped by zip code revealed that lower socio-economic status might be associated with delayed diagnoses, suggesting that limited access to healthcare resources and lower health literacy could hinder timely medical intervention. This segment of the study underscores the importance of socio-economic conditions in health equity,

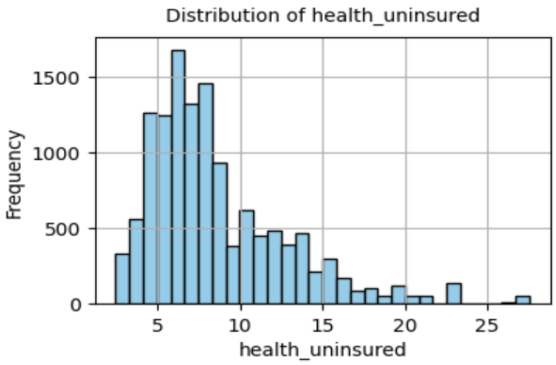
pointing towards the need for targeted healthcare policies and interventions to improve early cancer detection in underprivileged areas.

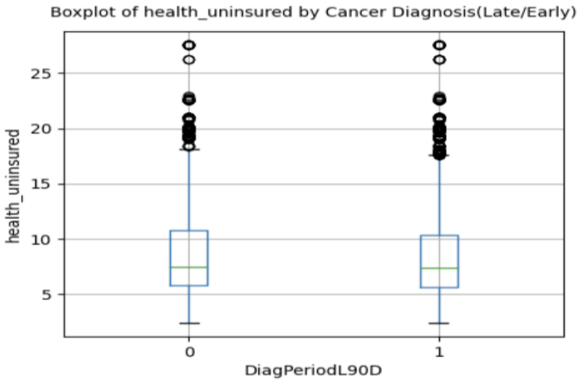


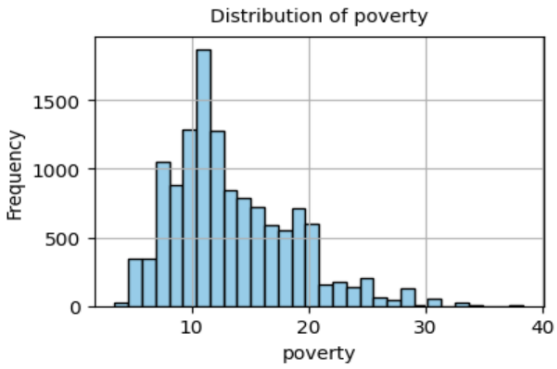


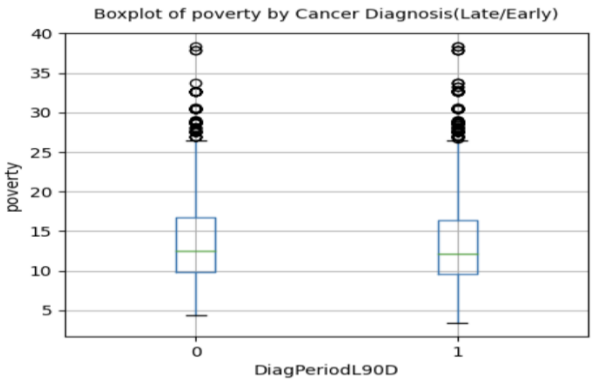






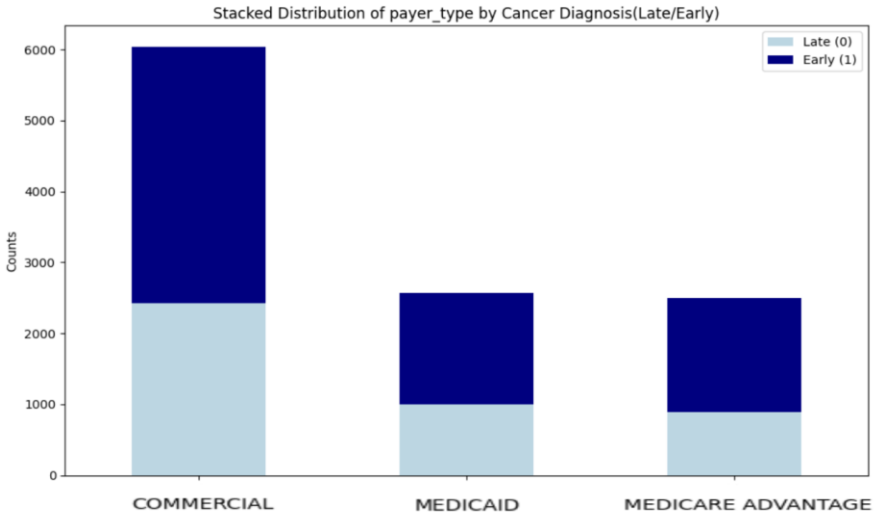






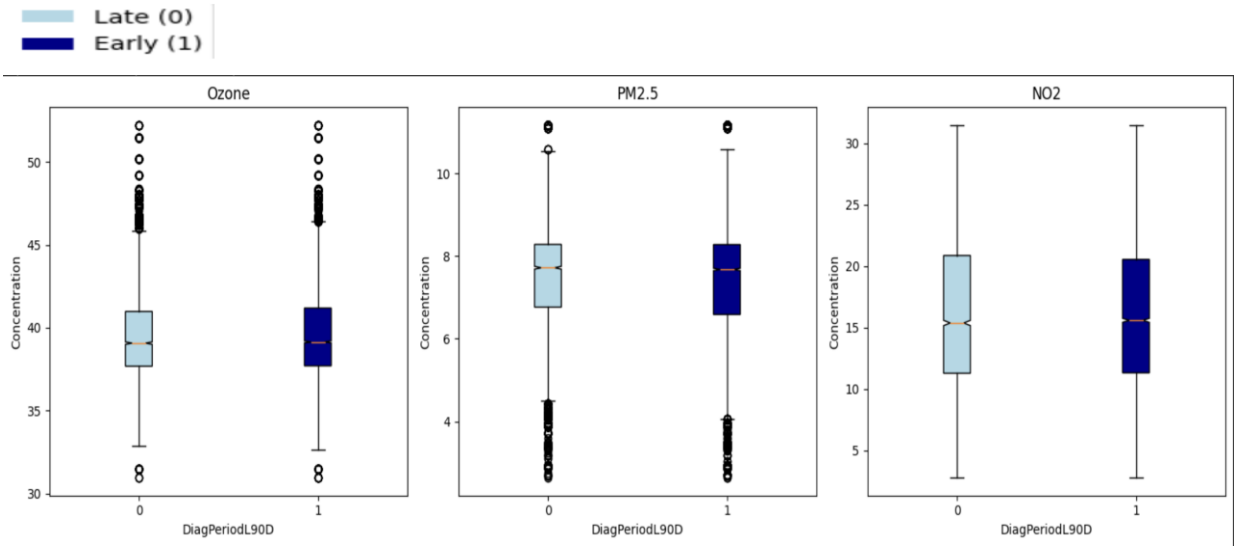
Analysis of Healthcare Coverage

We examined the role of different types of health insurance in the timing of cancer diagnoses. However, our current dataset did not show a clear differentiation in the timing of diagnoses across patients with commercial insurance, Medicare, or Medicaid. This observation suggests that further detailed analysis could be needed, possibly with additional data, to determine if and how insurance type affects the speed and timing of receiving a cancer diagnosis. Future studies could also explore the quality of insurance coverage and access to healthcare services as potential factors influencing diagnosis timing.



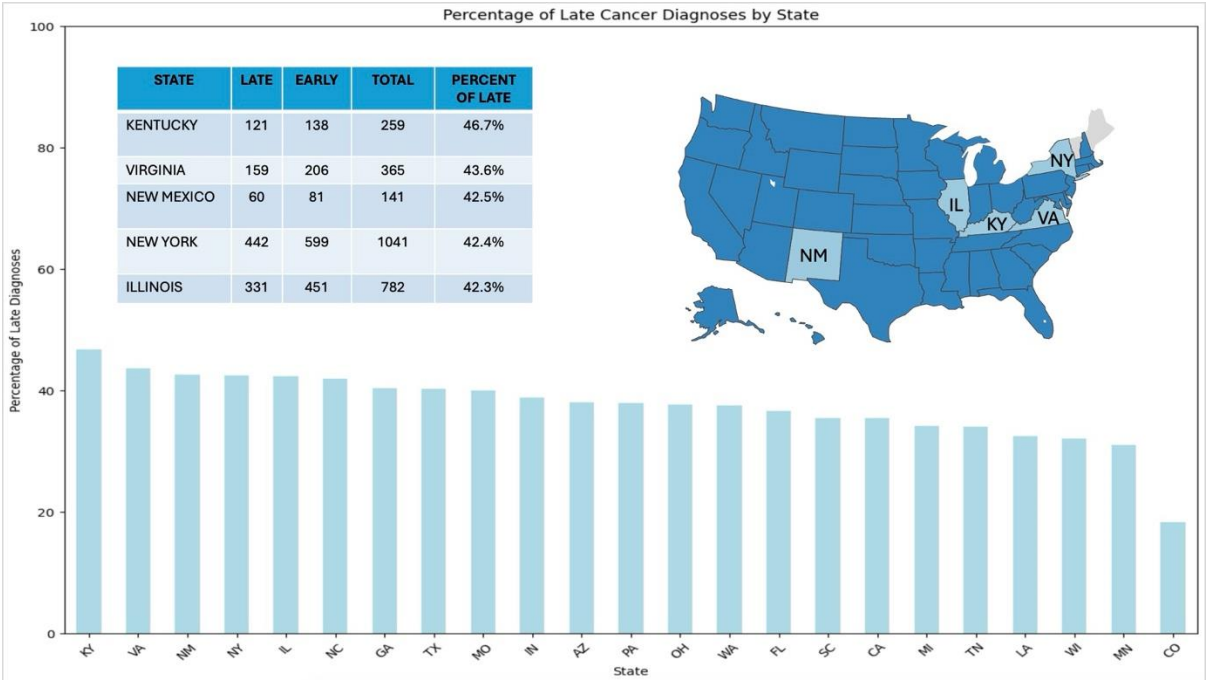
Environmental Pollutants and Health Outcomes

Environmental factors, including levels of ozone, PM2.5, and NO2, were analyzed to determine their impact on diagnosis timing. Although there was no clear correlation found between overall pollutant levels and the timing of diagnoses, areas with consistently high pollutant levels showed a slight trend towards later diagnoses, warranting further investigation into environmental health impacts.



Demographic Influences on Diagnosis Timing

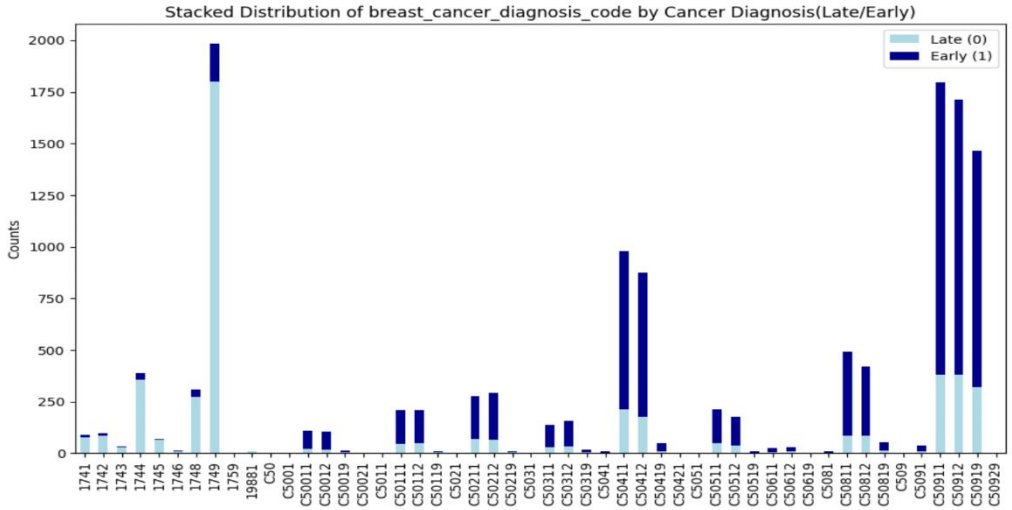
In assessing the impact of demographics on the timing of cancer diagnoses, our data revealed significant state-by-state disparities. The top five states with the highest rates of late diagnoses were Kentucky (46.7%), Virginia (43.6%), New Mexico (42.5%), New York (42.4%), and Illinois (42.3%). This data suggests a geographical influence on diagnosis timing, with these states exhibiting a notably higher percentage of late-stage diagnoses upon initial presentation. Such findings underscore the critical need for targeted interventions to improve early detection rates, which may include enhancing public awareness, screening accessibility, and healthcare infrastructure in these regions.



STATE	LATE	EARLY	TOTAL	PERCENT OF LATE
KENTUCKY	121	138	259	46.7%
VIRGINIA	159	206	365	43.6%
NEW MEXICO	60	81	141	42.5%
NEW YORK	442	599	1041	42.4%
ILLINOIS	331	451	782	42.3%

Technological and Coding Changes

The transition from ICD-9 to ICD-10 coding systems offered a unique insight into how technological and procedural updates in healthcare can impact diagnostic practices. Our study found a significant improvement in the timing of diagnoses following the adoption of ICD-10, likely due to the enhanced specificity and accuracy in disease coding.



6. Pre-Modelling (Data Preprocessing)

Based on the results obtained from the exploratory analysis the following steps were taken to prepare the data for training:

Step	Description
Data cleaning	• Dropping unnecessary columns that has more than 90% missing values. • Ensure all columns have the correct data types
Missing values Imputation	• The technique used to impute the missing column such as Payer types and BMI was **mode/median imputation**. The categorical columns were imputed using the mode and the numerical columns were imputed using the median values. • The imputation of the training dataset was done separately from the testing dataset. The training dataset values were used to impute both training and testing dataset to avoid information leakage.
Features Engineering	• Two new columns were generated based on the results of the exploratory analysis. • The first column was a **binary column** that reflect the type of the **ICD.** This column is derived from the Breast cancer diagnosis code. ICD-9 starts with number while ICD-10 starts with character. • The second feature generated was the metastatic cancer diagnosis category. This feature was obtained by extracting the first 3 characters from the metastatic diagnosis code.

By the end of this step, the dataset was cleaned and ready for training. The total number of features was 79 + 1 label column with 12,906 observations.

7. Modelling and Results

❖ General Model Results:

The modelling process followed the below procedure:

1- Dividing dataset 80% for training and 20% for testing
2- Constructing models using all available features, we regard these models as our **baseline models.**
3- We then explore various feature selection techniques to identify the most effective feature sets.
4- We used AUC (Area under the Curve) as our evaluation metric, since our final goal was to find the most important features that are associated with the most accurate model.

Baseline Model Results:

Features engineering	Logistic regression	GBT	Decision Tree	Random Forest	SVM
Yes	79.49%	80%	77.85%	79.93%	78.82%
No	56.00%	80.85%	77.88%	75.47%	55.22%

Features Selection:

To improve the model performance, we tried different features selection techniques such as the Variance Threshold Selector and the Univariate Features Selector, the results obtained by the Variance threshold will not be reported here as they did not improve the AUC, but they can be found in the notebook code.

Univariate Features Selector assesses the relationship between each feature and the target variable independently, without considering the interaction or correlation between features. It uses chi-square test for categorical features and ANOVA test for numerical features.

Based on the result of this test, the highly related features were as follows:

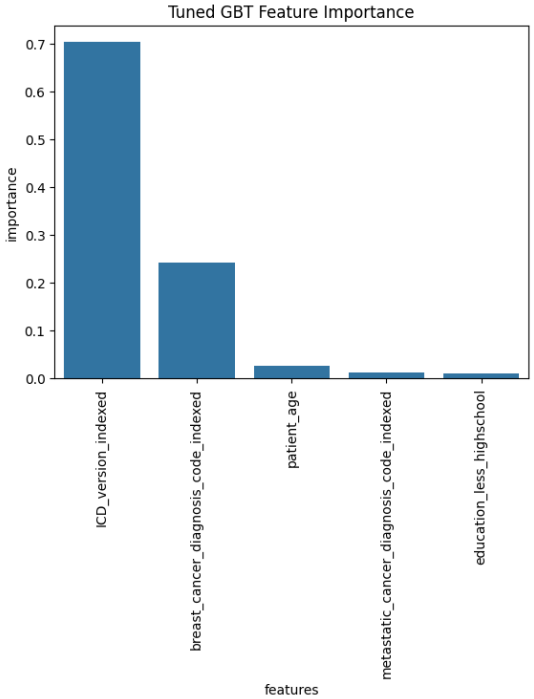
Numerical:	'patient_age' 'education_less_highschool'
Categorical: OBJ	'breast_cancer_diagnosis_code_indexed', 'ICD_version_indexed','diagnosis_category_indexed 'metastatic_cancer_diagnosis_code_indexed',

A subset of the data containing only these features was created, and the modelling results are as follows.

Features engineering	Logistic regression	GBT	Decision Tree	Random Forest	SVM
Yes	80.05%	81.06%	77.69%	81.35%	79.10%

- ✓ It can be observed that including only these features results in improving models AUC by approximately +1%.

In conclusion, the best performing model was a **Random Forest** with an AUC of **81.3%**, highlighting key factors such as the coding system, patient demographics, and zip code related socioeconomic factors as shown below:



❖ Commercial Plan Model Results:

Commercial plans are health services provided by companies for profit. The alternative being government or publicly funded health plans (Medicare, Medicaid, VA Health Care). This investigation was to see how accurately we could predict a patient's chance of receiving a diagnosis within 90 days by just looking at patient who had commercial insurance plans with the assumption that these plans provided better services to their clients. This will also help show disparities between commercial and publicly funded health plans.

The modelling process followed the same steps as the general model, with commercial insurance patients being filtered out.

Baseline Model Results:

Logistic Regression	GBT	Decision Tree	Random Forest	SVM
73.83%	72.44%	66.62%	74.48%	71.22%

In general, all the models had a lower AUC as compared to the general base models. The random forest had the highest AUC among all other commercial plan models.

Features Selection:

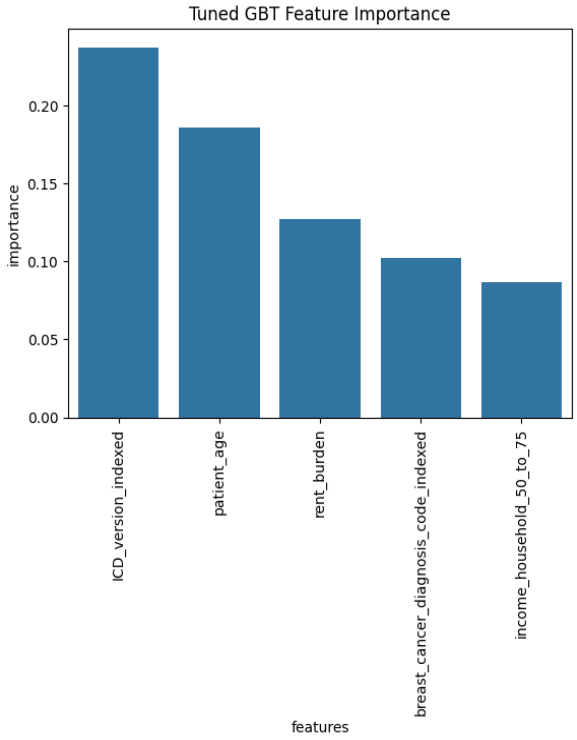
Using Univariate Feature Selection, we came up with categorical and numerical features that we believed to play a large role in predicting patient diagnosis within 90 days.

Numerical:	'patient_age', 'income_household_50_to_75', ' rent_burden', 'education_bachelors', 'self_employed'
Categorical:	'breast_cancer_diagnosis_code_indexed', 'ICD_version_indexed','diagnosis_category_indexed 'metastatic_cancer_diagnosis_code_indexed', 'patient_state_indexed'

Univariate Features Model Results:

Logistic Regression	GBT	Decision Tree	Random Forest	SVM
72.52%	72.41%	67.17%	72.58%	71.23%

In conclusion, the random forest base model gave us the best AUC proving that it was the best model to use in determining diagnosis within 90 days of patients with commercial insurance plans.



8. Discussion of Findings - Insights

The findings suggest that while environmental pollutants did not have a marked impact, socioeconomic factors such as education play critical roles in the timing of cancer diagnoses. The switch from ICD9 to ICD10 coding significantly reduced the rates of late diagnosis, indicating that coding practices can influence diagnostic accuracy and timeliness. This was by far the strongest predictor of diagnosis. The ICD9 system was particularly ineffective for patients over the age of 60. With the new ICD10 system, patients under 45 years old tended to be more at risk.

For demographic factors, high school education completion level was a prominent predictor. Insurance policies also appear to have little effect on time to diagnosis.

In terms of insurance type, we could not get a strong enough model like what we had with the combined data. In addition to that, feature selection giving us similar features as the general model tell us that there might not be a difference in diagnosis time per insurance type. This might be because of the different packages that patient's opt-in for. Since we had no idea what kind of package the patient had there is a likelihood that the quality of the package might be like other public or government funded insurance schemes.

9. Challenges and Limitations

The project faced challenges such as missing data, particularly in BMI and payer types, and limitations inherent to the PySpark functions, affecting the depth of analysis possible within this framework.

10. Conclusion

This project underscores the potential of Big Data and predictive analytics in addressing health inequities. By highlighting key factors influencing the timing of cancer diagnoses, the findings can inform policy changes and targeted interventions designed to reduce disparities in healthcare outcomes.

Recommended future steps for healthcare policy and practice include:

- Develop targeted screening programs for the high-risk demographics identified.
- Research further the success of the ICD10 coding system and how its successes may be applied to other areas.
- Take measures to ensure that age discrimination is kept to a minimum.

11. References and Acknowledgments

PySpark Documentation

- https://spark.apache.org/docs/latest/api/python/reference/index.html

Healthcare datasets and coding systems

- https://www.kaggle.com/competitions/widsdatathon2024-challenge1/data
- http://www.icd10charts.com/chartbuilder.php?chart=555f9168e5

Contributions from all team members and project advisors are gratefully acknowledged.