**Predicting Metastatic Cancer Diagnosis: Uncovering Healthcare Disparities**

**INTRODUCTION**

Breast cancer is one of the most common tumors affecting females all over the world, and timely diagnosis and treatment are essential for improving patient outcomes. However, disparities in access to and treatment may lead to a delay in diagnosis, especially in aggressive breast cancer types such as metastatic Triple Negative Breast Cancer (TNBC). These delays may have devastating consequences, which is why it is essential to identify and address the elements that contribute to inequalities in healthcare.

This study is based on the analysis of a single dataset provided by Gilead Sciences, which includes demographic, socioeconomic, and other relevant information, to predict whether patients receive a metastatic tumour diagnosis within 90 days of examination. The aim is to uncover patterns and relationships between patient demographics (if any), and the likelihood of timely diagnosis by implementing various classification models. The ultimate objective is to reveal potential disparities in healthcare admission and to understand how socioeconomic status may affect cancer diagnosis and treatment.

**RESEARCH QUESTIONS**

1. **How do patient demographics (e.g., age, race, socioeconomic status) influence the likelihood of receiving a metastatic cancer diagnosis within 90 days of screening?**
2. **Which classification model performs the best in predicting whether a patient will receive a metastatic cancer diagnosis within 90 days of screening?**

**DATASET OVERVIEW**

The dataset consists of 12,906 patient records related to breast cancer diagnoses, enriched with demographic, socioeconomic, geographic, and environmental data. It includes 83 attributes per patient, covering details such as age, race, insurance type, geographic location, income levels, education, and environmental factors like air quality metrics. The target variable, DiagPeriodL90D, indicates whether a patient received a metastatic cancer diagnosis within 90 days of screening. This binary variable plays a key role in recognizing differences in timely diagnosis and treatment. The dataset also contains detailed information about breast cancer diagnosis codes and descriptions, though some key columns, such as BMI and treatment data, have significant missing values.

The dataset covers U.S. states, with the highest concentration of patients in California. It includes socioeconomic indicators like median household income, education levels, and employment status, as well as environmental data such as ozone (O3) and fine particulate matter (PM2.5) levels. However, 49.47% of race data and 69.46% of BMI data are missing, which could limit analyses related to racial and health disparities. Although these obstacles remain, the dataset provides a solid foundation for exploring the factors affecting cancer diagnosis and treatment.

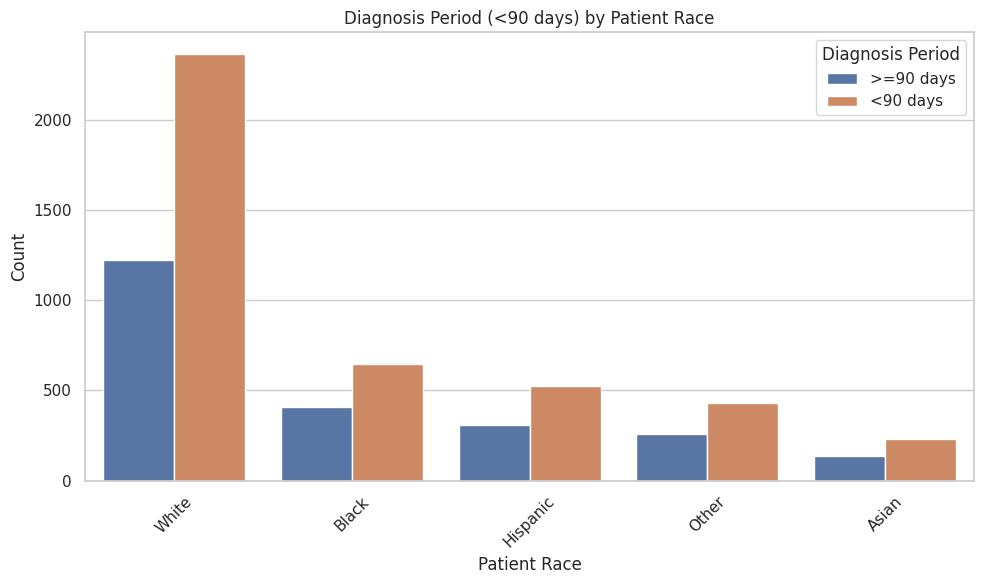
**METHODOLOGY**

All analyses, including visualizations, preprocessing, and model training, were performed on the training data. This approach helps in building models that generalize well to new, unseen data.

#### **Visualizations to Understand the Data**

To better understand the dataset, several visualizations were created. These helped identify patterns, trends, and disparities in the data. Some of the important visualizations include:

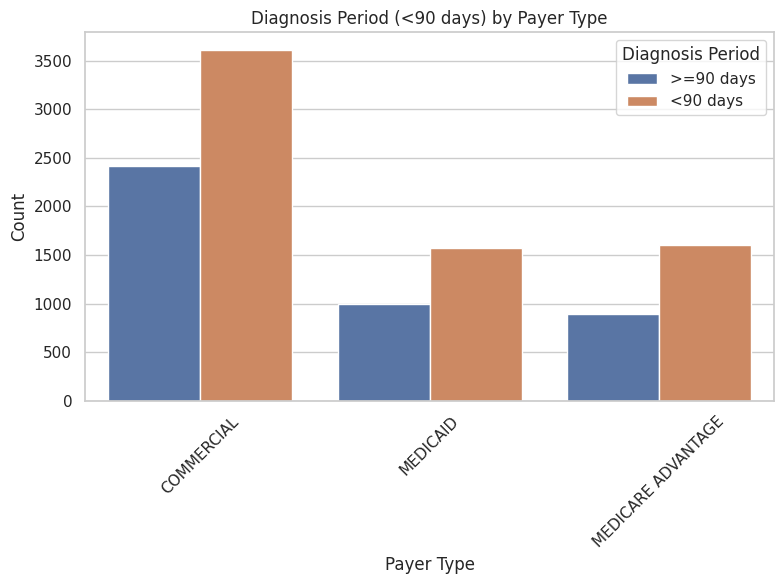
1. **Diagnosis Period by Patient Race**:



**Fig 1:** Diagnosis Period by patient race

This bar chart compares the number of patients diagnosed within 90 days versus those diagnosed later, across different racial groups. It shows that White patients form the largest group, with more timely diagnoses compared to other racial groups like Black, Hispanic, and Asian. It should be noted that, the dataset had large number of entries for White Patients compared to other races.

1. **Diagnosis Period by Payer Type**:



**Fig 2:** Diagnosis Period by payer type

This visualization breaks down the diagnosis periods by the type of health insurance (e.g., Commercial, Medicaid, Medicare Advantage). It reveals that patients with Commercial insurance are more likely to receive timely diagnoses compared to those on Medicaid or Medicare Advantage. It can be seen that, the number of Commercial Payers are significantly higher in numbers than the other 2 types.

These visualizations helped identify important trends and disparities in the data, guiding the subsequent analysis and modeling steps.

#### **Data Preprocessing Steps**

Before building any models, several preprocessing steps were implemented to ensure dataset was clean and ready for analysis:

1. **Handling Missing Values:**  
   Columns with more than 65% missing data (e.g., BMI, metastatic treatment details) were dropped. For the remaining columns, missing numeric values were filled with the median, and categorical values were filled with the most frequent category (mode).
2. **Dropping Irrelevant Columns:**  
   Columns like patient\_gender (all patients were female) and patient\_id (unique identifiers) were removed as they added no value to the analysis.
3. **Encoding Categorical Variables:**  
   Categorical columns (e.g., patient race, payer type) were converted into numerical format using label encoding or one-hot encoding, depending on the model requirements.
4. **Splitting the Data:**  
   The dataset was split into training and testing sets, with 80% of the data used for training the models and 20% reserved for evaluation. This ensured that the models were tested on unseen data to assess their performance accurately.

#### **Approach 1: Simple Models**

In the first approach, simple machine learning models were trained and evaluated to establish a baseline performance. The models used included:

* **Logistic Regression**: A basic model that predicts the probability of the target variable.
* **Random Forest**: A tree-based model that handles complex relationships in the data.
* **Gradient Boosting**: An advanced tree-based model that builds trees sequentially to improve performance.
* **Support Vector Machine (SVM)**: A model that finds the best boundary to separate the two classes.

These models were trained on the training data and evaluated using metrics like ROC AUC and classification reports.

#### **Approach 2: Advanced Models with Hyperparameter Tuning**

In the second approach, in addition to the first approach, more advanced models were introduced, and hyperparameter tuning was performed to optimize their performance. The models included:

* **XGBoost**: A powerful gradient boosting algorithm known for its speed and accuracy.
* **CatBoost**: A gradient boosting algorithm designed to handle categorical data efficiently.
* **LightGBM**: Another gradient boosting algorithm optimized for large datasets.
* **AdaBoost**: An ensemble method that combines multiple weak models to create a strong model.

Each model was fine-tuned using GridSearchCV, which tests different combinations of hyperparameters to find the best settings.

##### **CatBoost Model: With and Without One-Hot Encoding**

To better understand feature importance, the CatBoost model was trained in two ways:

1. **Without One-Hot Encoding**: CatBoost's built-in ability to handle categorical data directly was used to extract feature importance in the original format.
2. **With One-Hot Encoding**: Categorical variables were converted into numerical format using one-hot encoding before training the model.

Both methods were used to analyze which features contributed most to the predictions.

##### **Ensemble Learning: Voting Classifier**

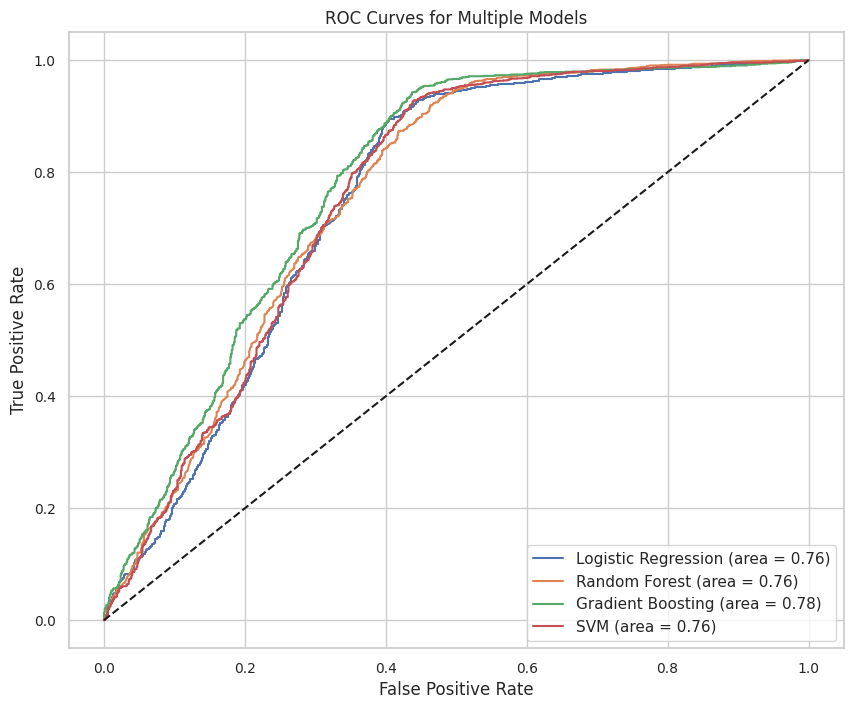
To further improve performance, an ensemble learning technique called the Voting Classifier was used. This approach combines the predictions of multiple models (e.g., CatBoost, XGBoost, LightGBM) using soft voting, which averages the predicted probabilities from each model to make the final prediction.

**RESULTS**

#### **Results from Approach 1: Simple Models**

In the first approach, simple machine learning models were trained and evaluated to establish a baseline performance. The models included Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). The performance of these models was measured using the ROC AUC score, which indicates how well the model distinguishes between patients diagnosed within 90 days and those diagnosed later.

* **Logistic Regression**: Achieved an AUC of **0.76**.
* **Random Forest**: Achieved an AUC of **0.76**.
* **Gradient Boosting**: Achieved an AUC of **0.78**.
* **SVM**: Achieved an AUC of **0.76**.

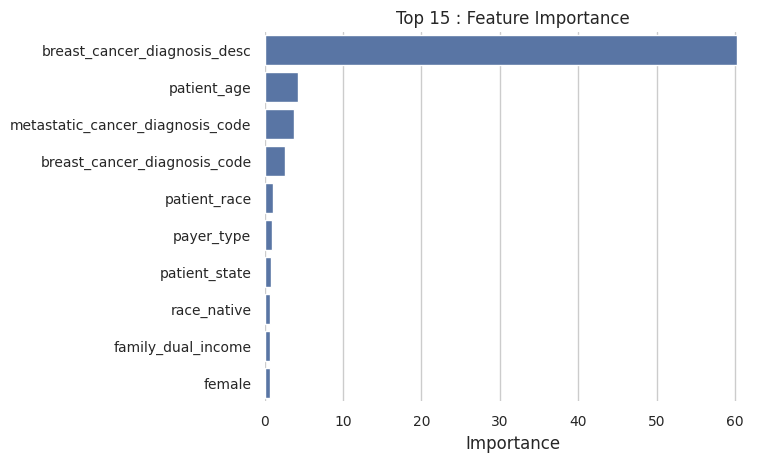
**Fig 3:** Approach 1 ROC Curves

Among these models, Gradient Boosting performed the best, with an AUC of 0.78, making it the top performer in Approach 1.

#### **Results from Approach 2: Advanced Models with Hyperparameter Tuning**

##### **CatBoost Model Without One-Hot Encoding**

The CatBoost model was first trained without one-hot encoding to leverage its ability to handle categorical data directly. This approach allowed for a clear interpretation of feature importance in the original format of the data. The model achieved an AUC of 0.79, making it the best-performing model so far.



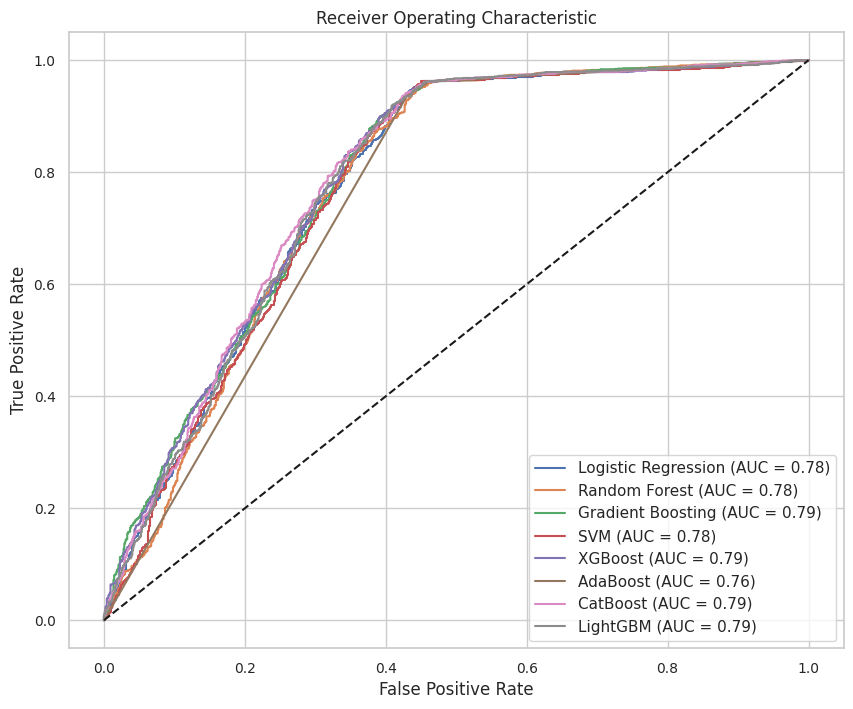
**Fig 4:** Feature Importance (Catboost, without one hot encoding)

* **Feature Importance Plot**: The feature importance plot revealed that breast cancer diagnosis descriptions (e.g., "Malignant neoplasm of breast, unspecified") and patient age were the most influential factors in predicting timely diagnosis. Other important features included patient race, payer type, state etc. Although they were not that important compared to the top variable. This analysis provided valuable insights into which factors contribute most to the model's predictions.

##### **Catboost and all Other Models with One Hot Encoding**

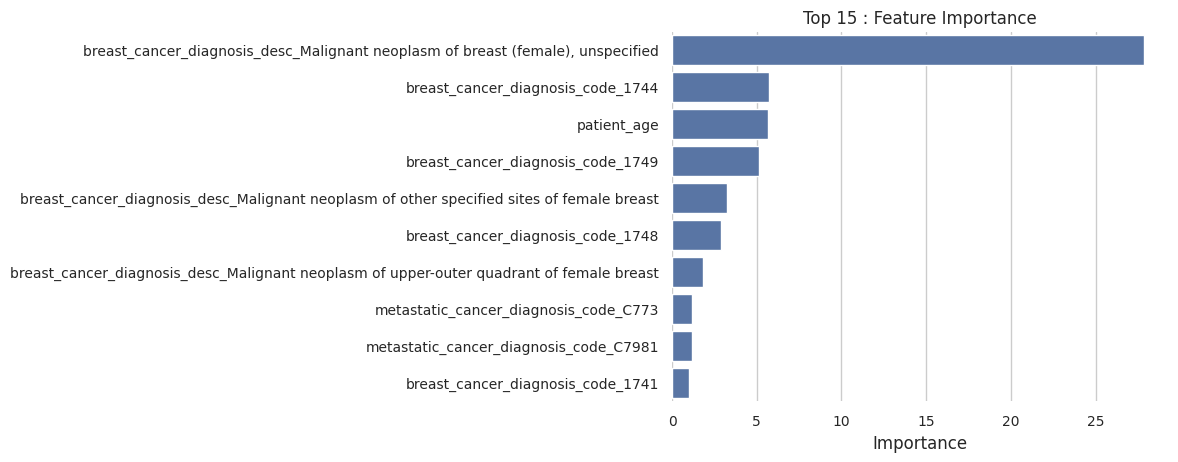
In addition to CatBoost, other advanced models were trained and fine-tuned using hyperparameter optimization and one hot encoding for categorical variables. Their performance was as follows:

* **XGBoost**: Achieved an AUC of **0.79**.
* **LightGBM**: Achieved an AUC of **0.79**.
* **AdaBoost**: Achieved an AUC of **0.76**.
* **Logistic Regression**: Achieved an AUC of **0.78**
* **SVM**: Achieved an AUC of **0.78**
* **Random Forest**: Achieved an AUC of **0.78**
* **Gradient Boosting**: Achieved an AUC of **0.79**
* **Catboost**: Achieved an AUC of **0.79**



**Fig 5:** Approach 2 ROC Curves (With One Hot Encoding)

All models performed well, but CatBoost, XGBoost, and LightGBM stood out with the highest AUC scores of 0.79.



**Fig 6:** Feature Importance (Catboost, with one hot encoding)

* **Feature Importance Plot (With One-Hot Encoding)**:  
  When CatBoost was trained with one-hot encoding, the feature importance plot showed similar trends, with breast cancer diagnosis descriptions and patient age remaining the most influential factors. However, the encoded categorical variables (e.g., specific diagnosis codes and payer types) were also highlighted as significant contributors. This confirmed the consistency of the findings across both approaches.

##### **Ensemble Learning: Voting Classifier**

To further improve performance, a Voting Classifier was used to combine the predictions of the top performing models (CatBoost, XGBoost, and LightGBM). The ensemble model achieved an AUC of 0.79, matching the performance of the individual models but providing more robust and reliable predictions.

#### **Final Model Selection**

Based on the results, both CatBoost and the Ensemble Voting Classifier performed well, achieving an AUC of 0.79. However, due to the simplicity and interpretability of the CatBoost model (Without One Hot Encoding), it was chosen as the final model over the ensemble approach.

##### **Performance Metrics for CatBoost**:

* **ROC AUC**: **0.79**
* **Precision**: 0.78 (for class 1) and 0.90 (for class 0)
* **Recall**: 0.96 (for class 1) and 0.55 (for class 0)
* **F1-Score**: 0.86 (for class 1) and 0.68 (for class 0)
* **Accuracy**: 81%

##### **Reason for Selecting CatBoost**: The CatBoost model without one-hot encoding was chosen as the final model because it is well-suited for this scenario with a highest prediction accuracy of 81% and AUC of 0.79. Unlike other models that require categorical data to be converted into numerical format (like one-hot encoding), CatBoost can directly handle categorical data. This makes the process simpler and faster, as it avoids the need for extra preprocessing steps. Additionally, CatBoost provides clear and interpretable feature importance, which helps us understand which factors like breast cancer diagnosis descriptions and patient age are most important in predicting timely diagnosis. This interpretability is crucial for making actionable recommendations in healthcare.

While the Ensemble Voting Classifier also achieved similar results, its complexity and lack of interpretability made CatBoost the preferred choice.

**CONCLUSION**

This project focused on predicting whether patients received a metastatic cancer diagnosis within 90 days of screening, using a dataset enriched with demographic, socioeconomic, geographic, and environmental data. The analysis was conducted in two main approaches. In the first approach, simple machine learning models like Logistic Regression, Random Forest, Gradient Boosting, and SVM were trained and evaluated. Gradient Boosting emerged as the best-performing model in this approach, achieving an AUC of 0.78. In the second approach, more advanced models like XGBoost, CatBoost, LightGBM, and AdaBoost were included, and hyperparameter tuning was performed to optimize their performance.

The analysis addressed two key research questions. First, it explored how patient demographics influence the likelihood of receiving a metastatic cancer diagnosis within 90 days of screening. The results showed that patient age is one of the most influential factors. Additionally, payer type (e.g., Commercial insurance) and race (e.g., White patients) also played a role. These findings show that demographic factors, particularly age, race and payer type, influence the likelihood of timely diagnosis. Second, the project evaluated which classification model performs the best in predicting whether a patient will receive a metastatic cancer diagnosis within 90 days of screening. Among all the models evaluated, CatBoost performed the best, achieving an AUC of 0.79. It outperformed other models due to its ability to handle categorical data directly, its interpretability, and its consistent performance across all metrics. The model's feature importance analysis also provided valuable insights into the important factors of timely diagnosis of Metastatic Cancer, making it the most suitable choice for this prediction task.