

CANCER PREDICTIVE MODEL

By Haya Hadaya

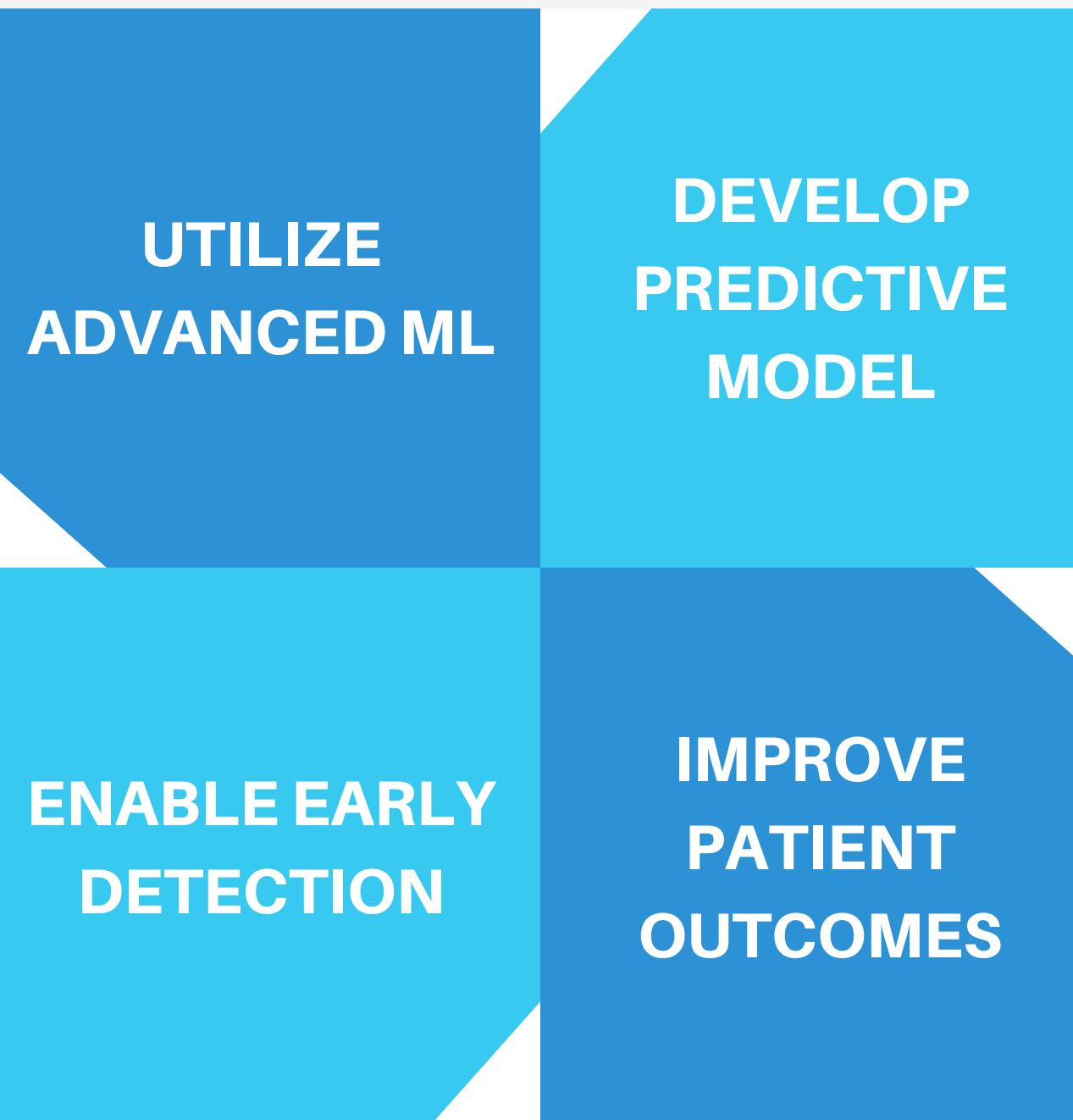
Problem Statement Overview

 Late Diagnosis: Impairs treatment, worsens outcomes.

 Early Detection Challenge: Delays care, raises expenses.

 Importance: Timely detection important for better outcomes, cost-efficiency.

Proposed Solution



Early Diagnosis Impact



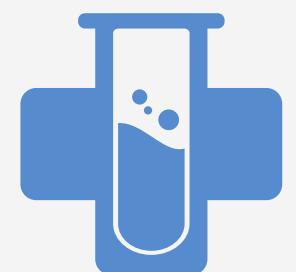
Resource Shift: Early intervention reduces costs



ML Detection: Slows cancer, betters outcomes.



Risk Awareness: Informs healthier choices.



Personalized Treatment: Early detection tailors therapies.

ANALYSIS

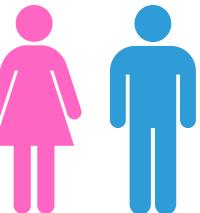
Dataset Overview

Features and Target Variable

1- Lifestyle Factors :

- Health Conditions 
- Healthcare Practices 
- General Lifestyle Choices 

2-Demographic Factors



Target Variable : Cancer (Cancer Risk) 

Dataset source and Description

- The Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) for the year 2021.

- The dataset has 438,693 rows and 303 columns , and it was condensed down to just 18 columns

Preprocessing Procedures

1

MISSING VALUES

Imputation and Removal

2

ONE-HOT ENCODING

Dummy Variables

3

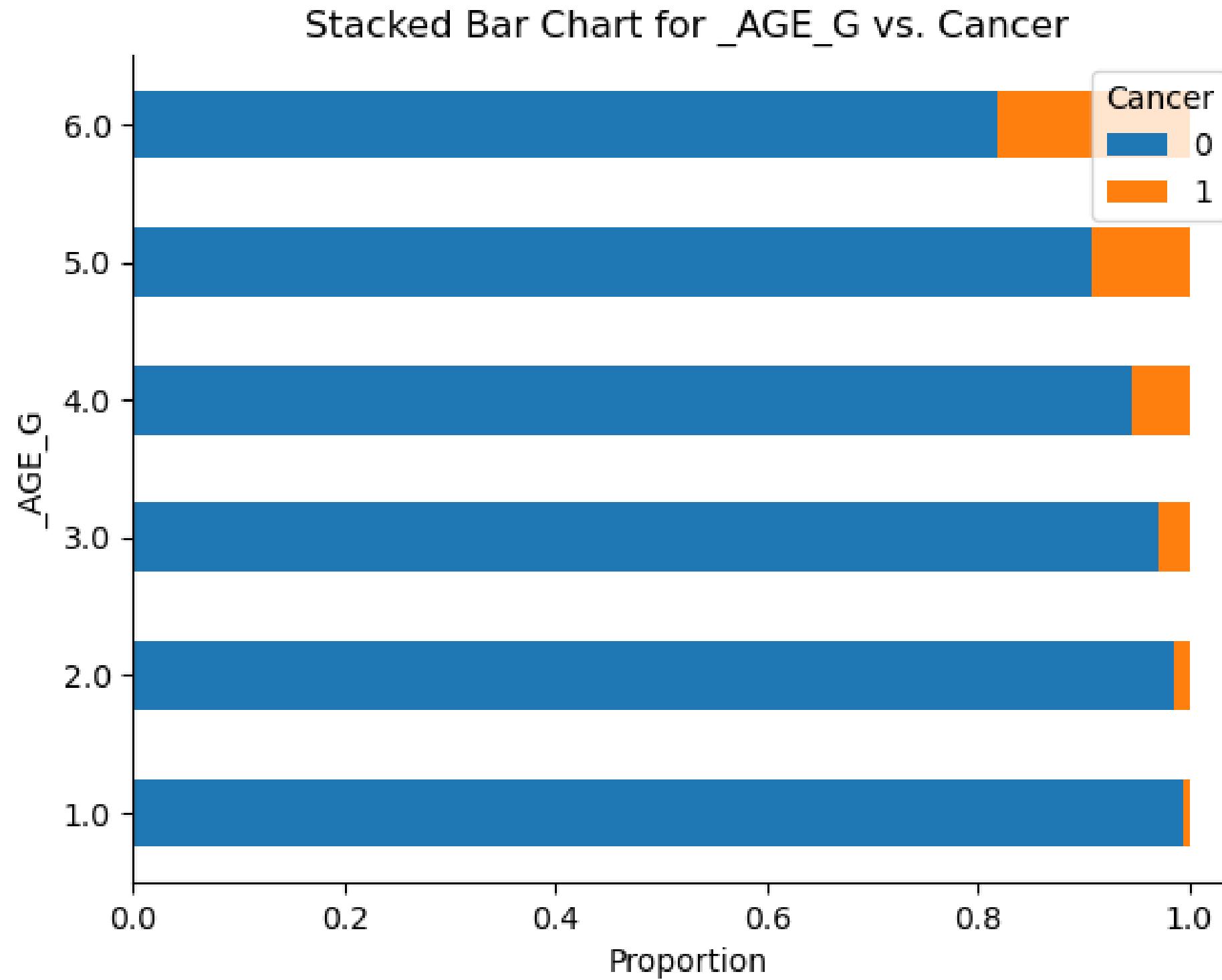
FEATURE ENGINEERING

Grouping Infrequent Values

4

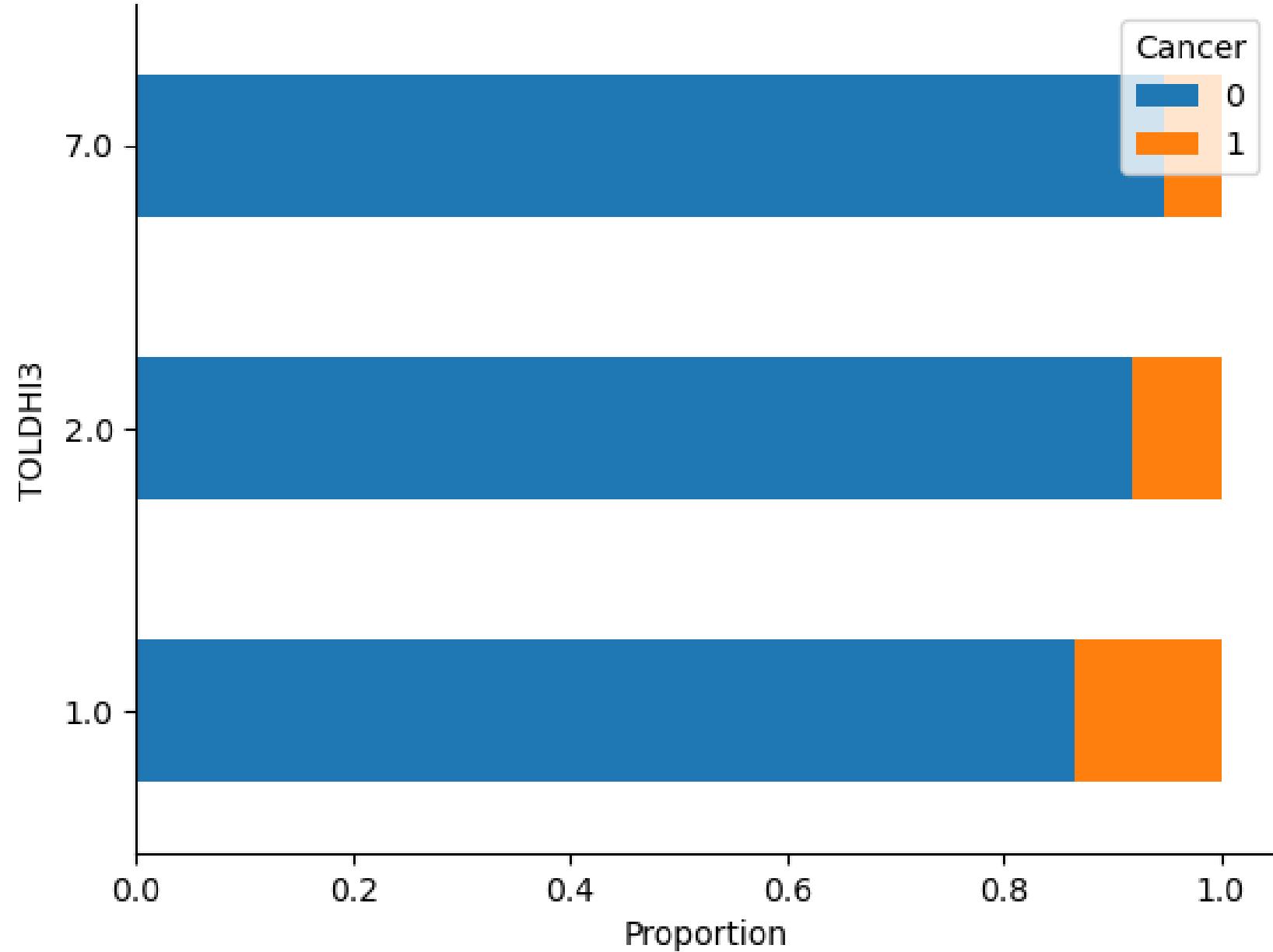
PRINCIPAL COMPONENT ANALYSIS (PCA)

Top Key Findings from EDA



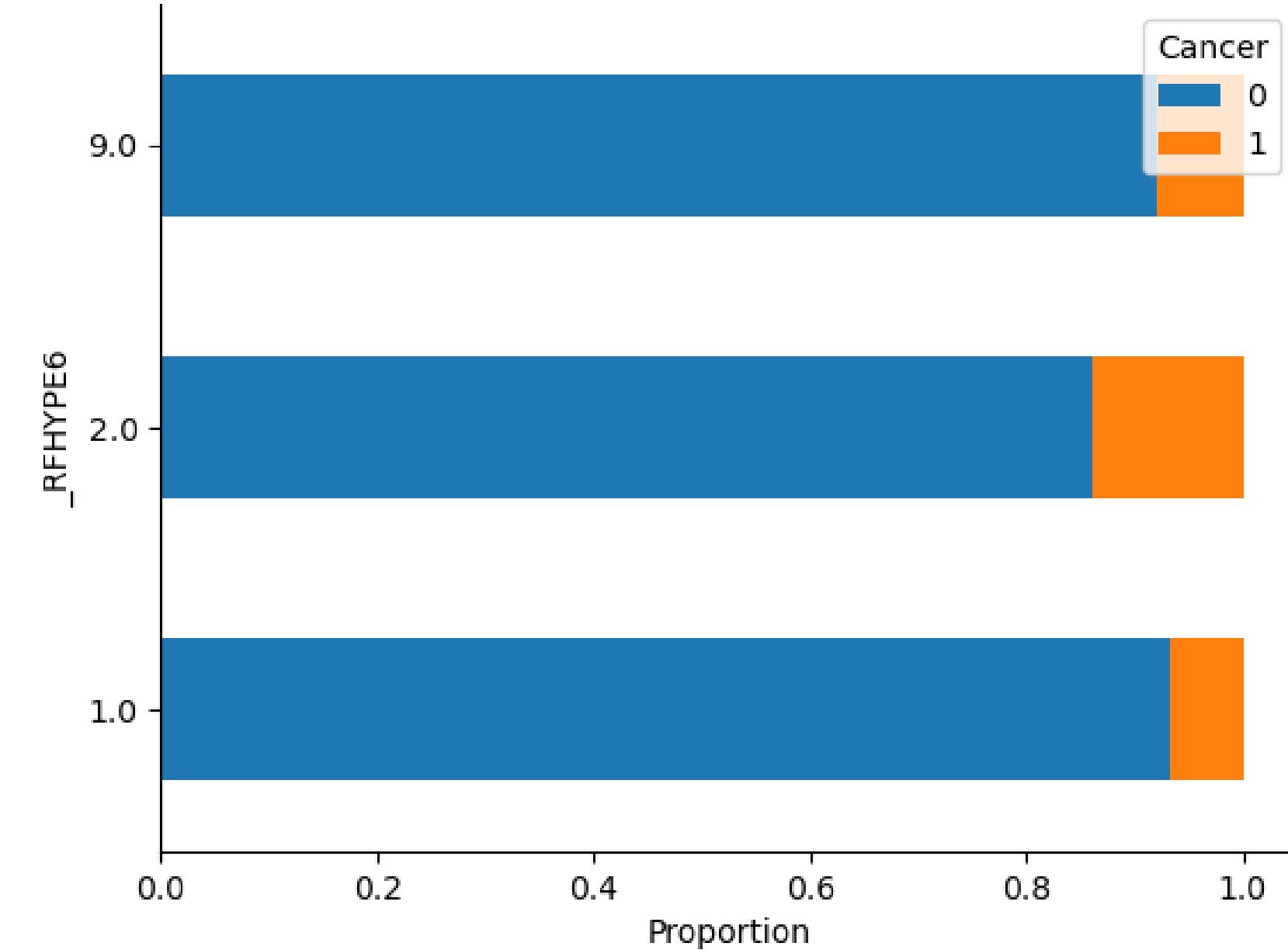
There is a higher association between older age groups and cancer presence.

Stacked Bar Chart for TOLDHI3 vs. Cancer

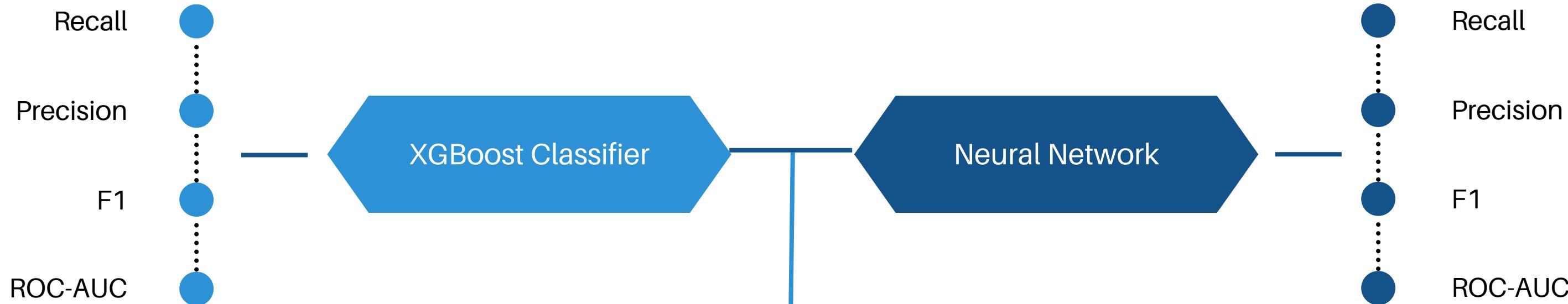


High cholesterol levels among cancer patients

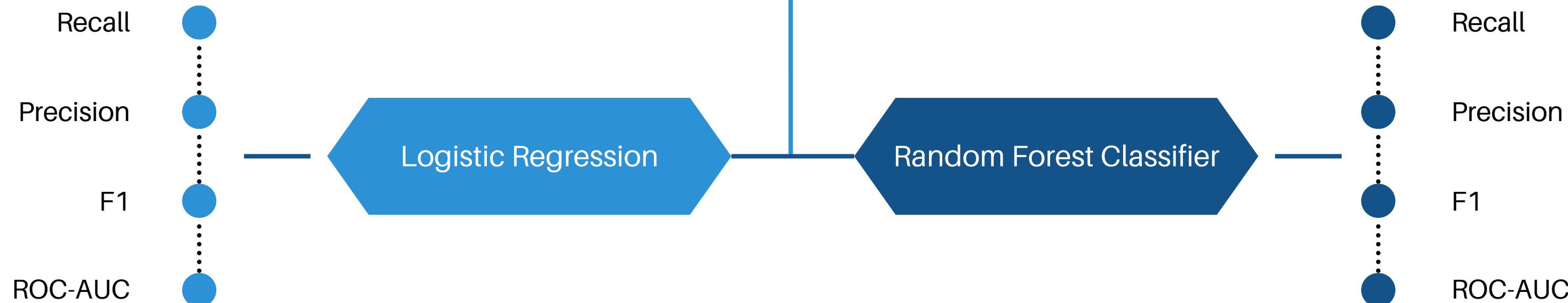
Stacked Bar Chart for _RFHYPE6 vs. Cancer



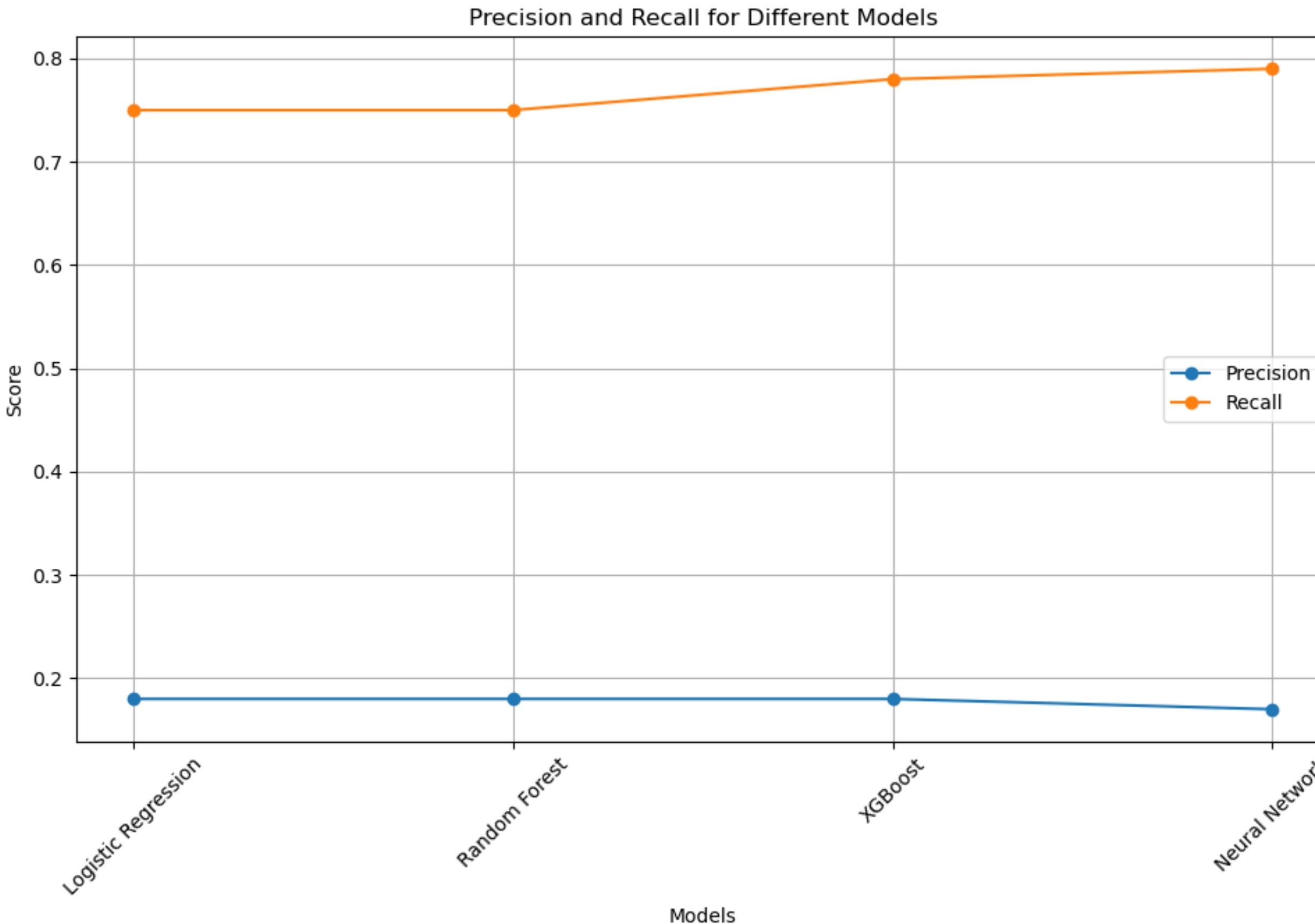
High blood pressure among cancer patients



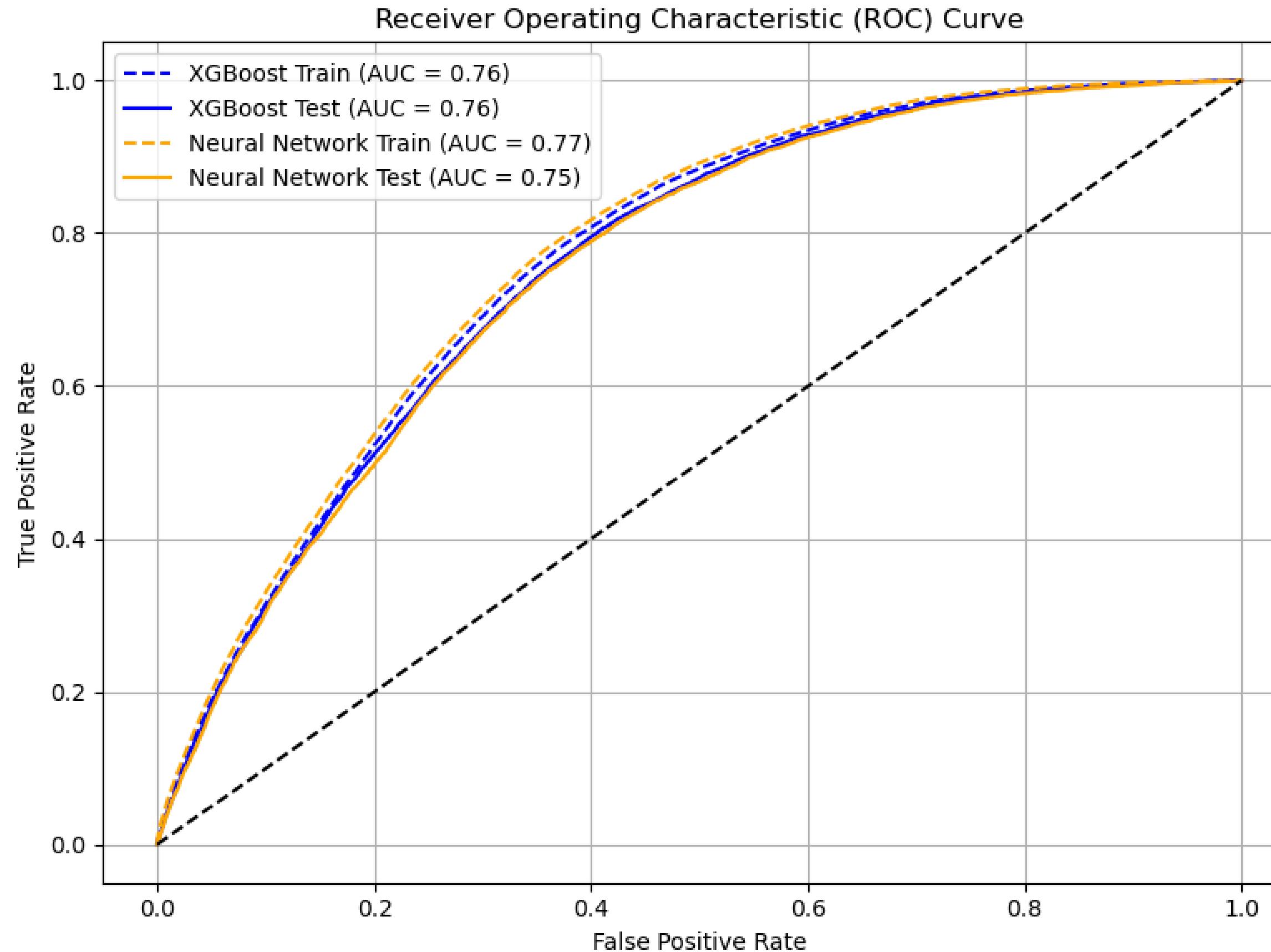
MODELLING APPROACHES AND EVALUATION METRICS

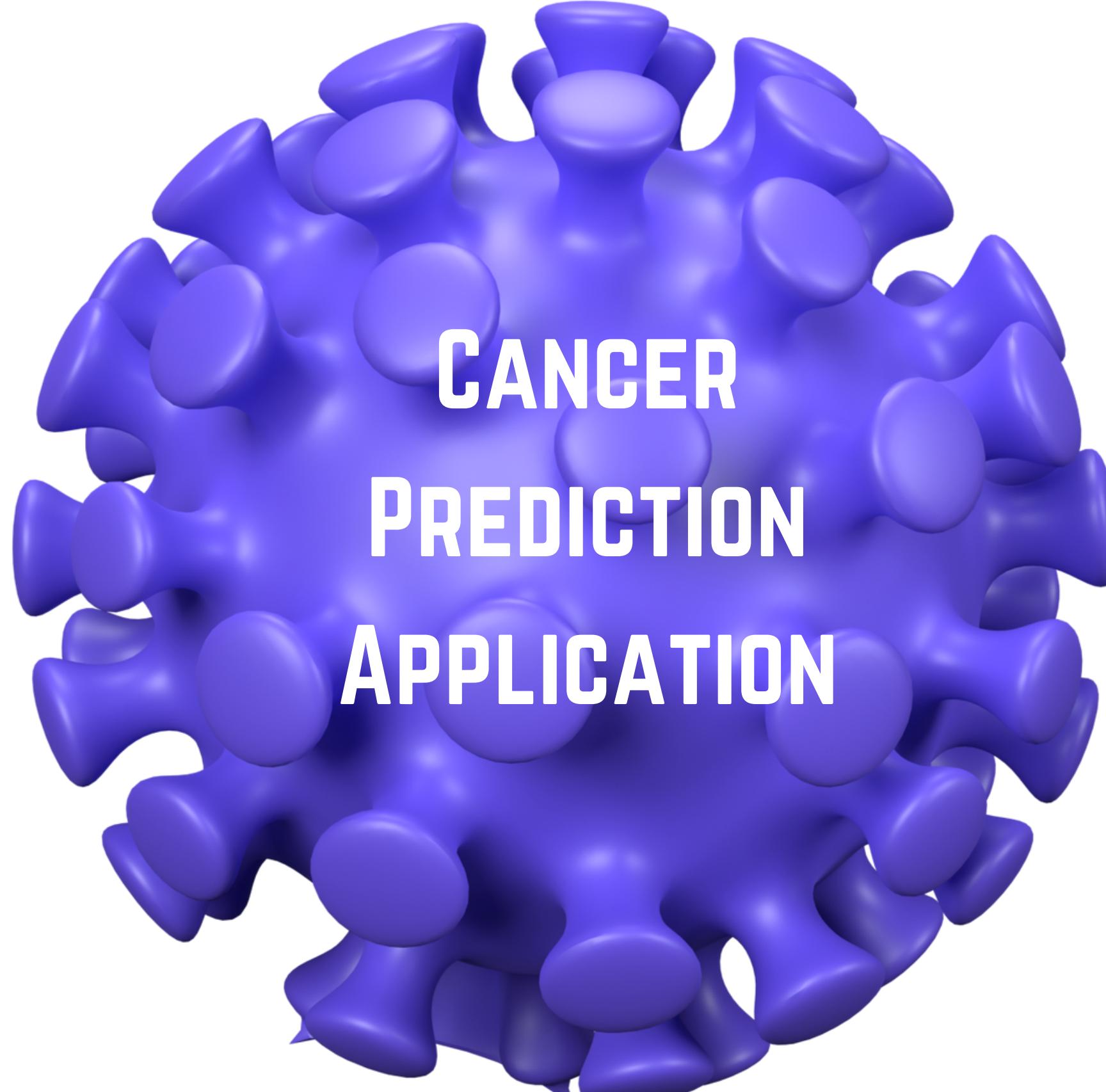


Recall and Precision Scores for Different Models



Comparison of (ROC) Curves for XGBoost and Neural Network Models





CANCER PREDICTION APPLICATION

NEXT STEPS FOR FUTURE WORKFLOW





Thank You