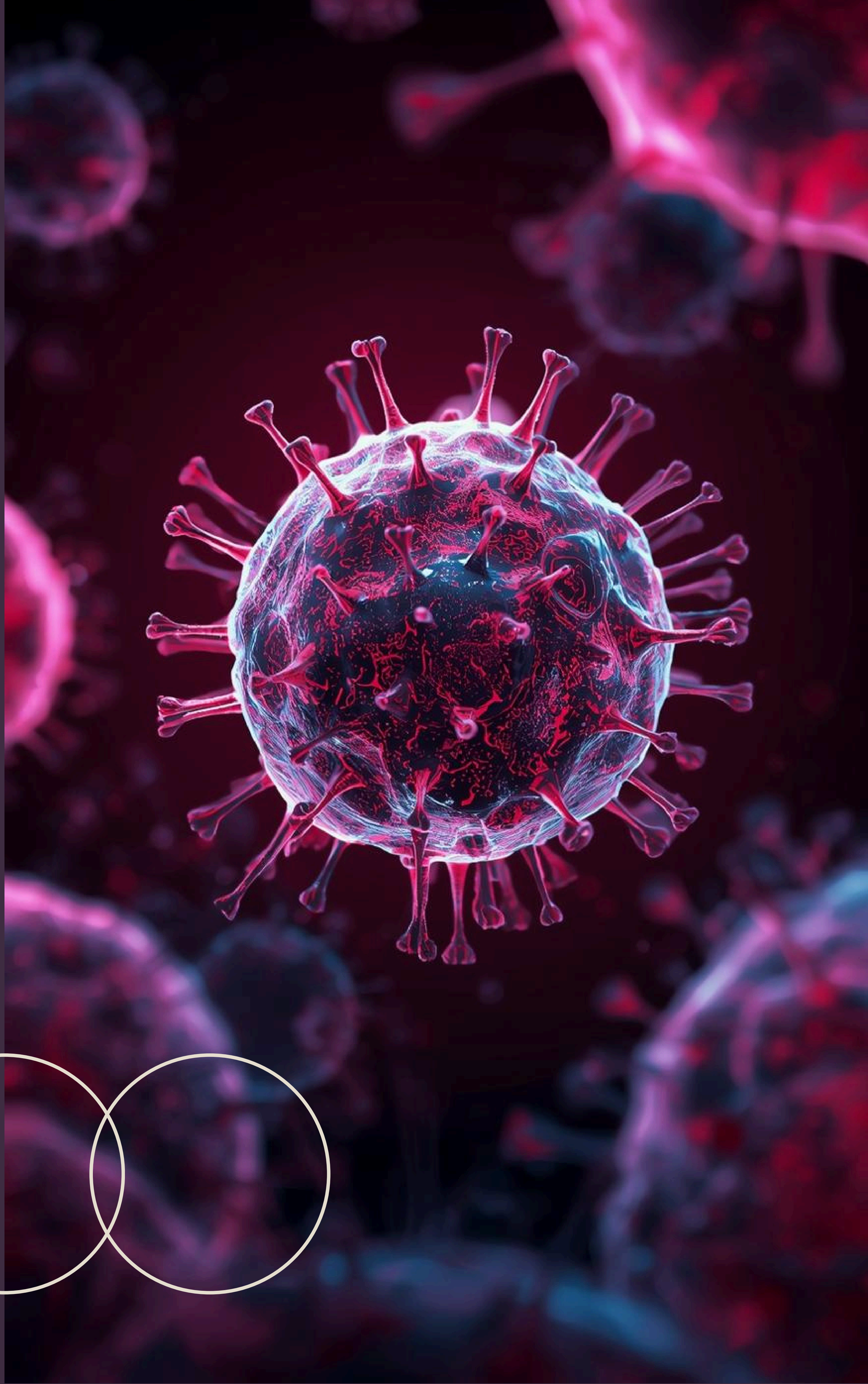
FIGHTING BREAST CANCER WITH DATA

By : Meshack Mboya



BUSINESS UNDERSTANDING

- Breast cancer has different types that affect how it’s treated. Genetic tests to identify these types aren't always easily accessible.
- This project uses machine learning to predict breast cancer molecular subtype and survival status, helping guide treatment decisions without relying on costly tests

PROBLEM STATEMENT

- The lack of access of genetic test like PAM50 in most health centres always leave the patients uninformed about their tumor aggressiveness
- This project predicts breast cancer molecular subtype and survival status using machine learning, helping guide treatment.

PROJECT OBJECTIVE

This project uses machine learning to:

- Predict breast cancer molecular subtype (digital alternative to PAM50 tests for personalized treatment)
- Predict breast cancer survival status to identify high-risk patients early
- Predict a patient’s vital status by identifying whether death is due to cancer or other causes

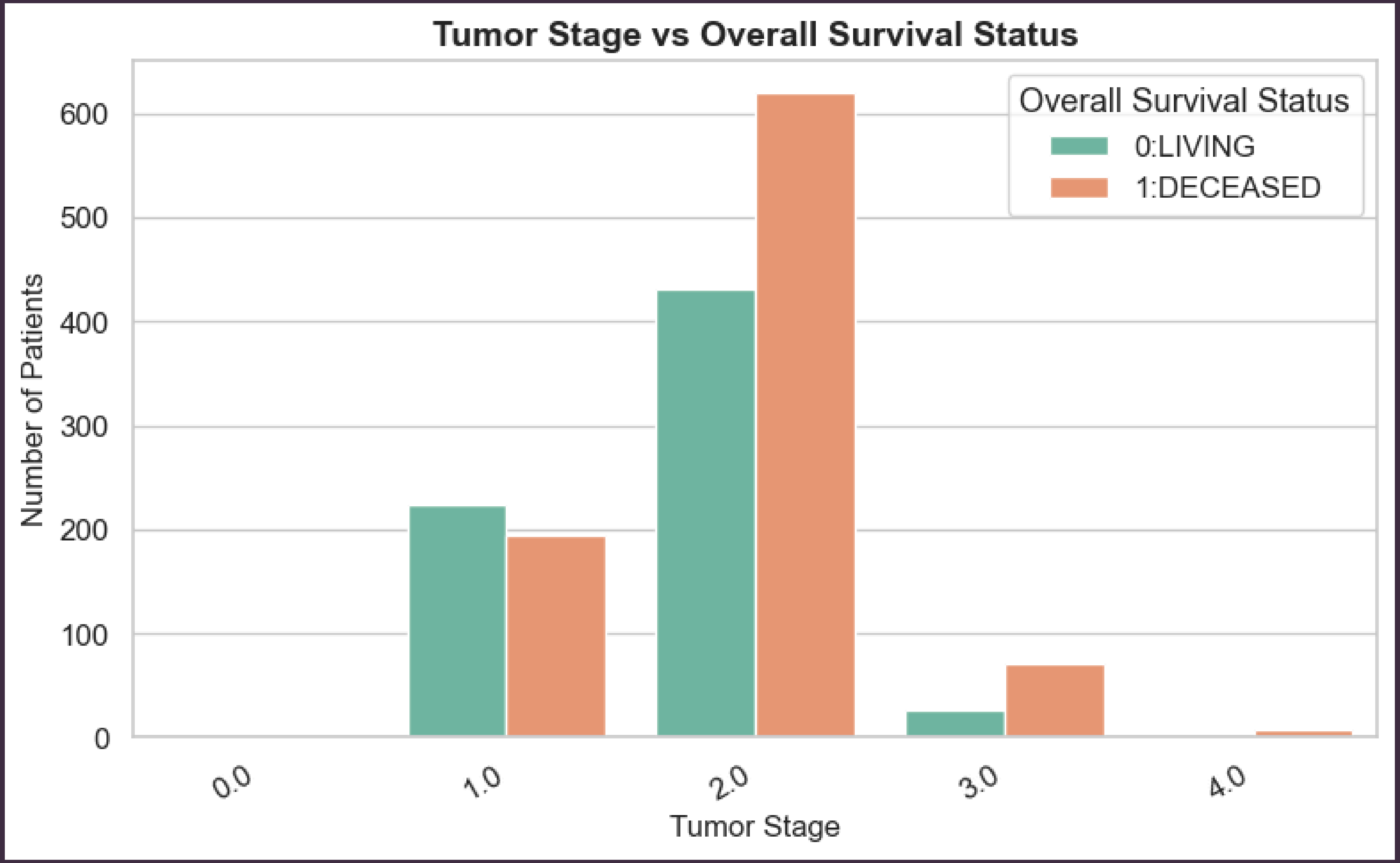
DATA UNDERSTANDING

- The data used is a METABRIC dataset that combines clinical and genomic data to explore links between tumor characteristics, treatment response, and survival outcomes.
- It consist of 2,509 rows and 39 features columns

METHODOLOGY

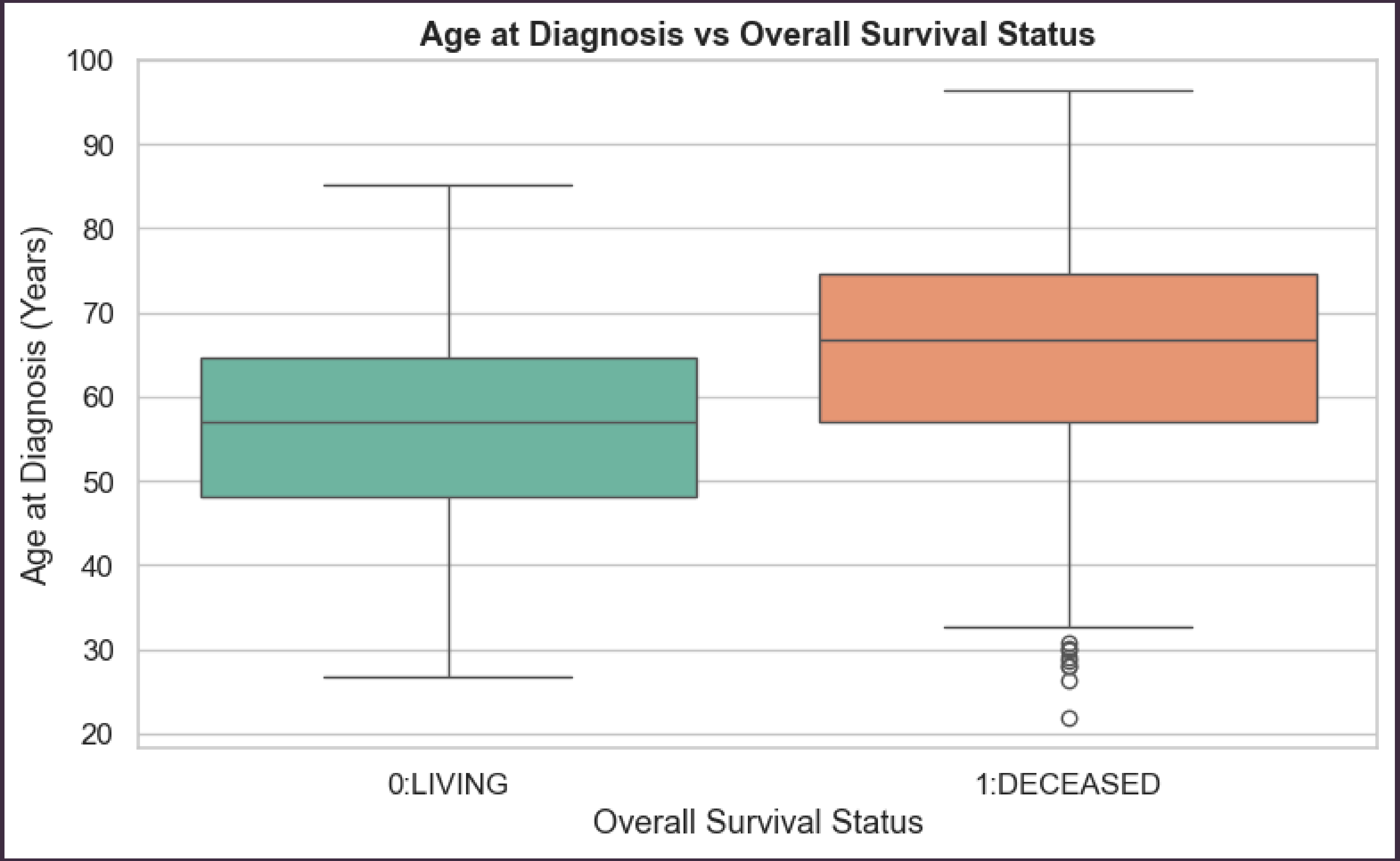
- Data preparation (handling missing values, duplicates)
- EDA and visualization
- Predictive Modeling
- Model Interpretation using Feature Importance Analysis
- Model Evaluation (Using Macro F1 score, accuracy, confusion matrix)

EFFECT OF TUMOR STAGE ON SURVIVAL



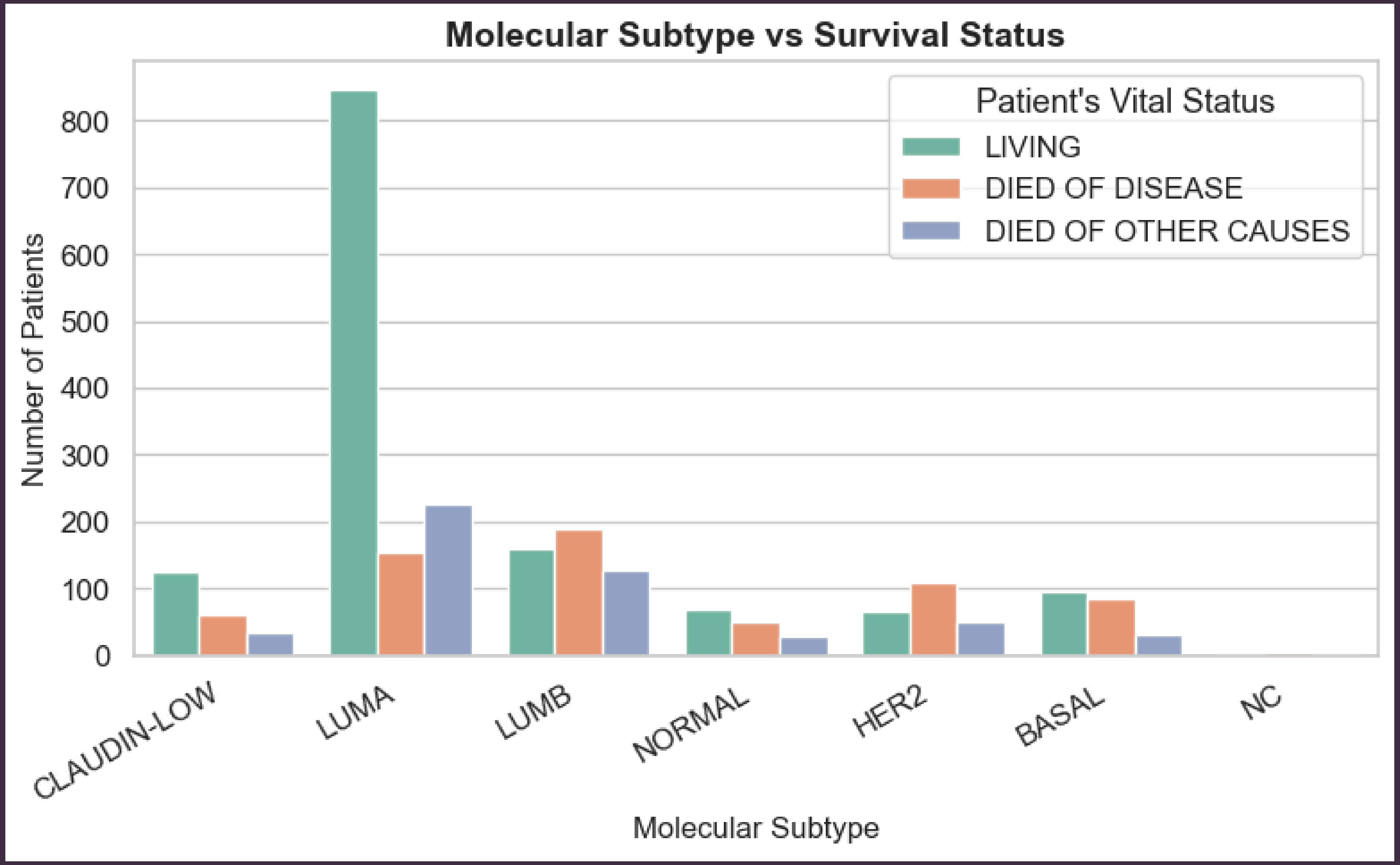
- Lower Tumor stages (0 and 1) have better survival rates compared to higher stages (2 and 3)- bar for deceased patients is higher than that of living patients in stages (2 ,3 & 4)

HOW AGE AFFECT SURVIVAL



- Most deceased patients were diagnosed at an older age . The presence of outliers suggest that there are some younger patients who also experienced poor survival
- Most of middle age patients are living suggesting better survival outcomes for these groups

SURVIVAL OUTCOMES ACROSS MOLECULAR SUBTYPES



- Luminal A is associated with the best survival outcomes compared to other subtypes.

KEY INSIGHTS

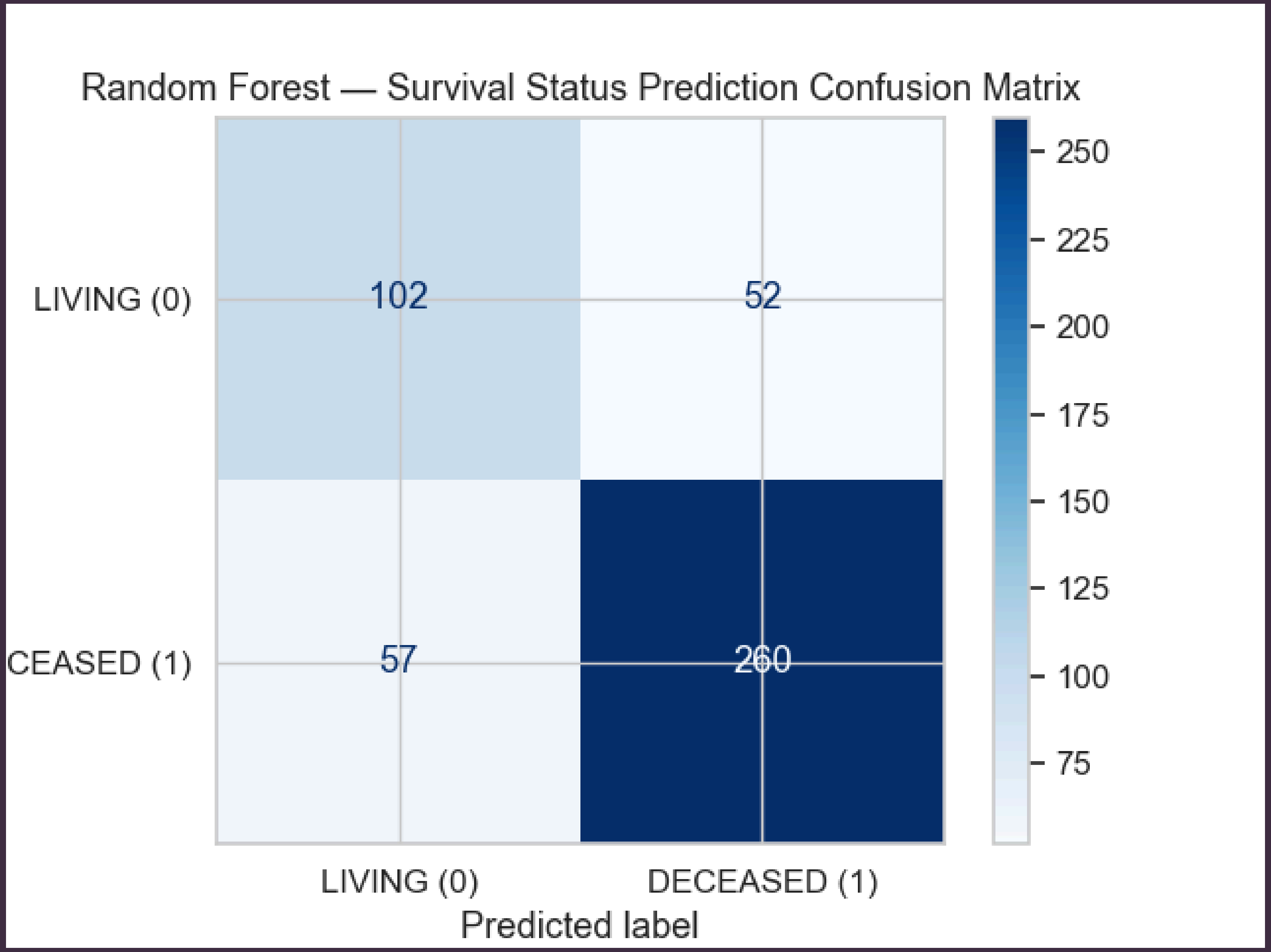
1. Lower Tumor stages (0&1) have better survival rates compared to higher stages
2. Older patients at diagnosis have poorer survival outcomes.
3. Patients with hormone receptor–positive tumors (ER+ and PR+) generally respond well to hormonal therapy and tend to have better survival.
4. Chemotherapy has a positive impact on survival
5. Luminal A subtype respond better to hormone therapy

MODELING

Molecular Subtype Prediction

- Model evaluation metrics: F1-macro of 61% and Accuracy of 70%
- Label encode multi-class target column- convert categorical data into numerical format
- Standard scale numeric features and one-hot encode categorical features
- Applied Smote for class imbalance
- GridsearchCV with best n_estimators of 300 rep total number of trees and best max_depth of 5 to control overfitting.

Survival Status Prediction



- Model evaluation metric: Accuracy of 77%
- Out of 471 patients the model correctly predicts 260 deceased patients and 102 living patients.

CONCLUSION

1. Predictive models can help identify high-risk patients and improve clinical decision-making.
2. Treatment can be adjusted based on the type of breast cancer for the best outcomes.
3. Tumor stage, age at diagnosis, type of treatment, Lymph nodes examined positive, ER/PR/HER2 status strongly influence survival outcomes

RECOMMENDATIONS

1. For health centres: Embed these predictive tools in clinical workflows to support decisions where genomic testing (e.g., PAM50 assay) is unavailable.
2. For Personalised care: Tailor treatment based on individual characteristic
3. Focus follow-up and resources on highest-risk patients: Older patients, Large & High grade tumors, more positive lymph nodes
4. Prioritize public awareness, screening programs & routine check-ups for women to promote early stage detection.

NEXT STEPS

- **Model optimization:** Fine tune patient vital status model so that it correctly predicts Died of Disease and Died of other causes classes
- **Validation:** Test the models on clinical settings for reliability
- **Data Expansion:** Incorporate imaging and gene expression data to enhance analytical insights

Thank you