Colorectal Cancer Survival Prediction

A Data Science Report

By Rohit Ashokkumar Tolani

Table of Contents

1. Introduction

Colorectal cancer is a significant global health concern, and early prediction of survival outcomes can assist healthcare professionals in designing better treatment plans. This project analyzes a dataset of **89,945 patient records with 30 features**, aiming to identify key factors influencing survival chances.

2. Objective

The primary goal of this study is to build a predictive model to determine which factors contribute most to **survival chances** among colorectal cancer patients. The analysis includes:

- Identifying the most important predictors of survival.
- Evaluating the impact of demographics, medical history, and lifestyle choices.
- Developing a machine learning model to predict survival probabilities.

3. Dataset Overview

- **File Name**: colorectal_cancer_prediction.csv
- **Format**: CSV (Comma-Separated Values)
- **Number of Rows**: 89,945
- **Number of Columns**: 30
- **Key Features**: Age, Gender, Tumor Aggressiveness, Medical History, Lifestyle Choices, Treatment Access, and Survival Status.

4. Data Preprocessing

4.1 Data Cleaning

- Handled missing values by replacing numerical missing data with **median values** and categorical missing data with **mode values**.
- Removed irrelevant features and redundant columns.

- Standardized and normalized numerical values for better model performance.

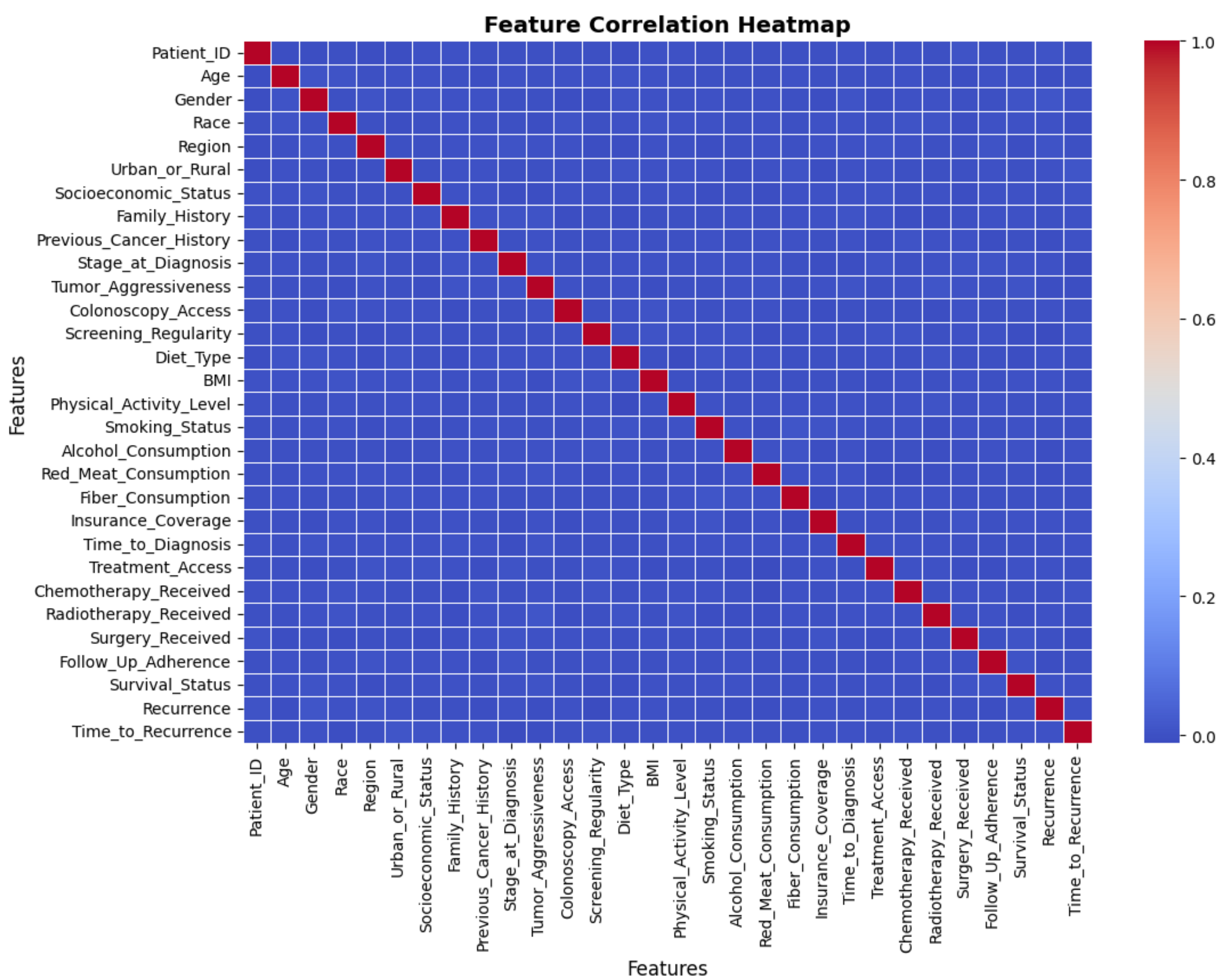
4.2 Feature Encoding

- Converted categorical variables using **Label Encoding** to make them machine-readable.
- Scaled numerical variables using **StandardScaler** to ensure uniformity.

5. Exploratory Data Analysis (EDA)

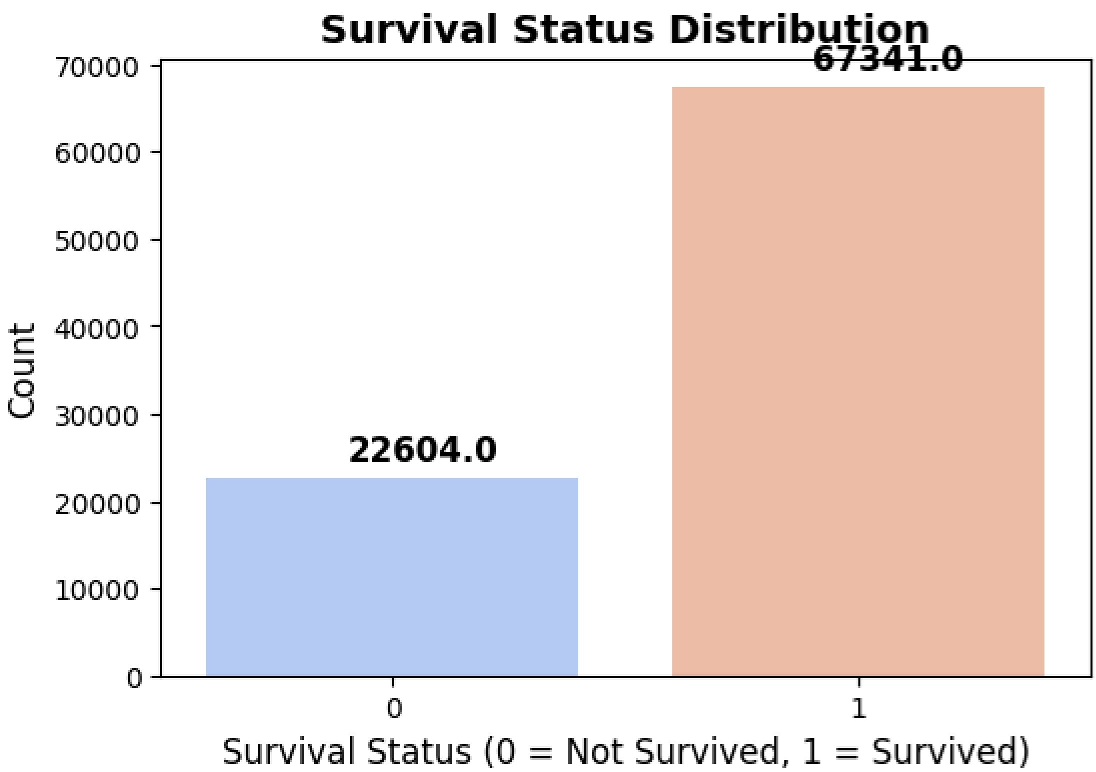
5.1 Correlation Analysis

A **correlation heatmap** was created to visualize relationships between features. This helps in identifying redundant features and understanding interactions between medical and lifestyle factors.



5.2 Survival Status Distribution

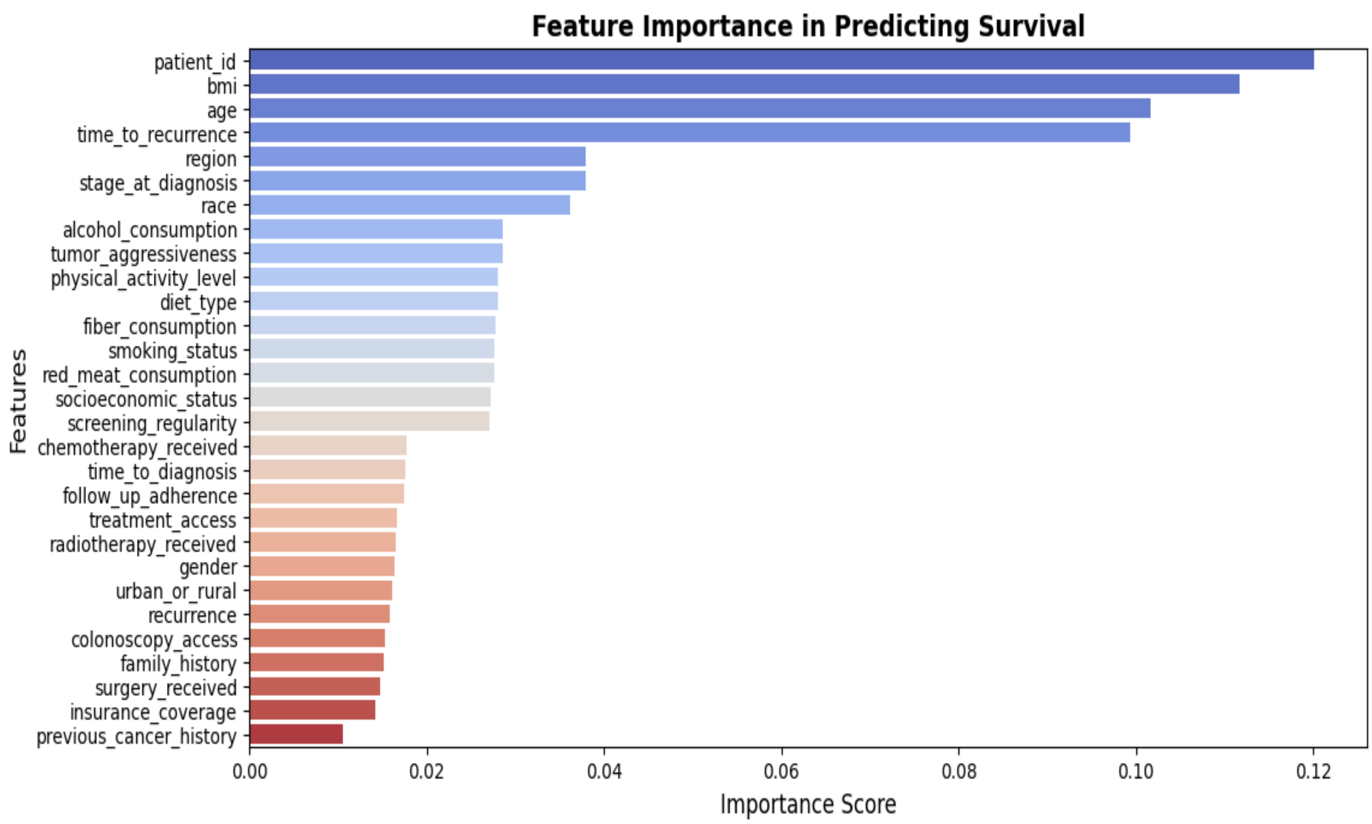
A **bar chart** was used to visualize the proportion of patients who survived versus those who did not. The dataset exhibits an imbalance, with **67,341 survived and 22,604 not survived**.



5.3 Feature Importance

Feature importance analysis highlights key variables influencing survival. Key predictors include:

- **BMI**: Affects survival outcomes significantly.
- **Stage at Diagnosis**: Early detection increases survival probability.
- **Previous Cancer History**: Impacts long-term survival rates.
- **Treatment Access & Follow-up Adherence**: Plays a crucial role in post-treatment recovery.



6. Machine Learning Model

6.1 Model Selection

A **Random Forest Classifier** was chosen due to its robustness in handling large datasets with mixed data types. The model was trained on an **80-20 train-test split**.

6.2 Model Performance

The model was evaluated using:

- **Confusion Matrix** for classification accuracy.
- **ROC-AUC Score**: Demonstrated a strong ability to distinguish between survival and non-survival cases.
- **Precision, Recall, and F1-Score**: Ensured balanced performance.

7. Key Insights & Conclusion

1. **Early detection significantly improves survival rates**, highlighting the importance of regular screening.
2. **BMI, diet, and physical activity play a role in cancer prognosis**, indicating the need for lifestyle interventions.
3. **Access to healthcare and treatment adherence are crucial factors**, emphasizing healthcare accessibility in improving survival outcomes.
4. **The model performed well but could be further improved using ensemble methods (e.g., XGBoost) and hyperparameter tuning.**

8. Future Scope

- Integrating additional medical imaging data for **better feature extraction**.
- Testing **deep learning approaches** for improved accuracy.
- Conducting **real-world validation** with healthcare professionals.