Colorectal Cancer Survival and Risk Analysis

Pavan Kumar Sanjay

Introduction/Objective

- Colorectal cancer is a type of cancer that affects the colon(Large intestine) or rectum.
- It starts due to abnormal growth or polyps in the colon or rectum
- Diagnosis of this disease involves the use of CT scans of this specific region
- **Objective:** The objective of this project is to answer the following questions:
  a. Determining the survival status of the individual
  b. Choosing the most significant causes of colorectal cancer

Dataset

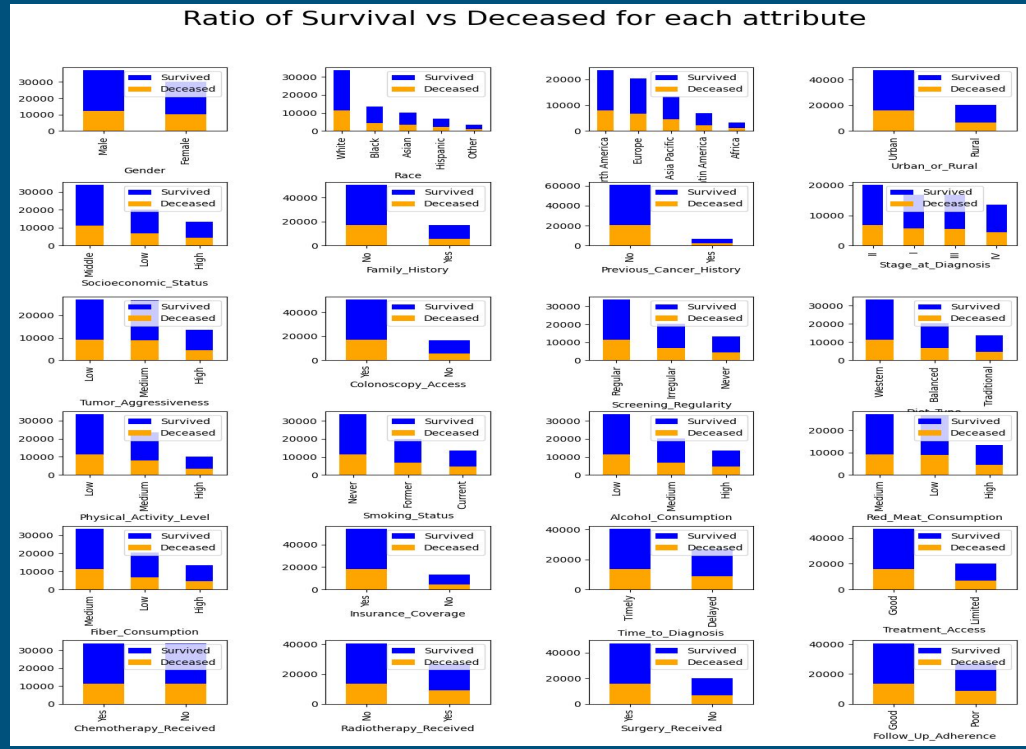
- The dataset used in this case was the Colorectal Cancer Risk and Survival Data which is a dataset available on Kaggle
- It has 89,945 rows and 30 columns with various attributes like patient id, race, gender and so on.

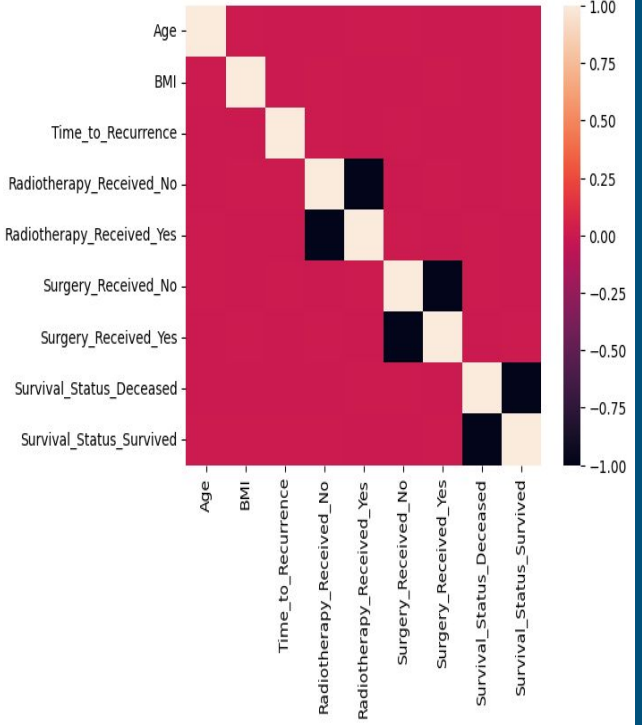
EDA(Exploratory Data Analysis)

- The list of different variables in this dataset is displayed on the right hand side.
- Most of these variables are categorical variables so for the purpose of analysis and modeling one-hot encoding will applied for the variables.
- On displaying the correlation matrix no significant correlation is observed between these variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 89945 entries, 0 to 89944
Data columns (total 30 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Patient_ID                89945 non-null  int64
 1   Age                       89945 non-null  int64
 2   Gender                    89945 non-null  object
 3   Race                      89945 non-null  object
 4   Region                    89945 non-null  object
 5   Urban_or_Rural            89945 non-null  object
 6   Socioeconomic_Status      89945 non-null  object
 7   Family_History            89945 non-null  object
 8   Previous_Cancer_History   89945 non-null  object
 9   Stage_at_Diagnosis        89945 non-null  object
 10  Tumor_Aggressiveness      89945 non-null  object
 11  Colonoscopy_Access        89945 non-null  object
 12  Screening_Regularity      89945 non-null  object
 13  Diet_Type                 89945 non-null  object
 14  BMI                       89945 non-null  float64
 15  Physical_Activity_Level   89945 non-null  object
 16  Smoking_Status            89945 non-null  object
 17  Alcohol_Consumption       89945 non-null  object
 18  Red_Meat_Consumption      89945 non-null  object
 19  Fiber_Consumption         89945 non-null  object
 20  Insurance_Coverage        89945 non-null  object
 21  Time_to_Diagnosis         89945 non-null  object
 22  Treatment_Access          89945 non-null  object
 23  Chemotherapy_Received     89945 non-null  object
 24  Radiotherapy_Received     89945 non-null  object
 25  Surgery_Received          89945 non-null  object
 26  Follow_Up_Adherence       89945 non-null  object
 27  Survival_Status           89945 non-null  object
 28  Recurrence                89945 non-null  object
 29  Time_to_Recurrence        89945 non-null  int64
dtypes: float64(1), int64(3), object(26)
memory usage: 20.6+ MB
```

EDA(Exploratory Data Analysis)





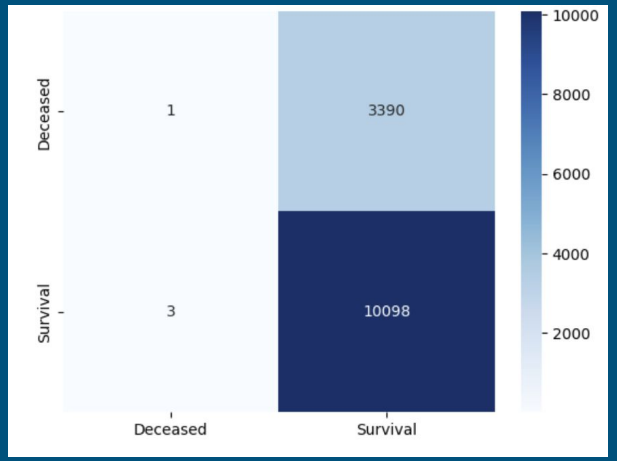
Modeling

I. Before modelling the data the following steps are undertaken to preprocess it:
   A. One hot encoding
   B. Normalization(A form of Standardization)

II. The following models have been used for the project:
   1. KNN
   2. Decision Tree Classifier
   3. Gaussian Naive Bayes
   4. Multinomial Naive Bayes
   5. Random Forest Classifier
   6. Balanced Boosting Classifier
   7. Logistic Regression

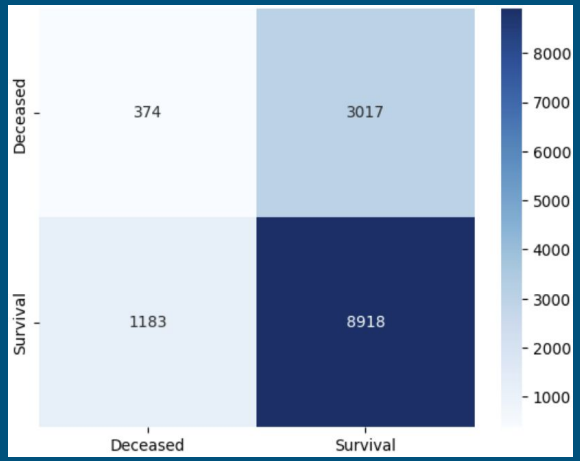
Modeling (Best Performing Models)

	precision	recall	f1-score	support
0	0.25	0.00	0.00	3391
1	0.75	1.00	0.86	10101
accuracy			0.75	13492
macro avg	0.50	0.50	0.43	13492
weighted avg	0.62	0.75	0.64	13492



Random Forest Classifier

	precision	recall	f1-score	support
0	0.24	0.11	0.15	3391
1	0.75	0.88	0.81	10101
accuracy			0.69	13492
macro avg	0.49	0.50	0.48	13492
weighted avg	0.62	0.69	0.64	13492



Balanced Bagging Classifier

Conclusions and Future Scope

- Best Model:
  - The Balanced Bagging Classifier achieved the highest accuracy, precision, recall, and F1-score among all models tested.
- Prediction Insights:
  - Predicting survival status remains challenging, with room for improvement in model performance. Most features-except patient ID, gender, and race-significantly influence survival outcomes.
- Future Directions:
  - Incorporate larger and higher-quality datasets to enhance analysis.
  - Explore Bayesian modeling approaches for deeper probabilistic insights, complementing the current frequentist perspective.

References and Links

- Dataset Link: https://www.kaggle.com/datasets/ankushpanday1/colorectal-cancer-risk-and-survival-data/data
- Code Repo link: https://github.com/rockaroll/rockaroll.github.io.git
- Code Deployment link: https://rockaroll.github.io/