**Machine Learning model to predict the Overall survival status in patients with Cholangiocarcinoma**

Ankita Thomas, Bsc.

Nellie Campbell, Bsc.

**Introduction**

To date, cholangiocarcinoma (CCA) is the second most common liver cancer after hepatocellular carcinoma (HCC), and patients with its diagnosis have increased over the years(1). Even though CCA is a rare cancer, its incidence, however, says otherwise. CCA can develop in one of three anatomical locations, intrahepatic (iCCA), perihilar(pCCA), and distal (dCCA)(2). Over the years, the incidence and mortality of CCA have increased enough to make it a world health problem(3), especially the iCCA(2). Nevertheless, CCA remains a rare type of cancer yet to be diagnosed at its early stages. This has led to patients having limited treatment options at the time of diagnosis, and surgery is the only possible treatment modality for a cure.

For this project we wanted to use a machine learning model to predict the overall survival status in patients with cholangiocarcinoma using the publicly available Cancer Genomics database The Cancer Genome Atlas Program (TCGA)

**Methods**

**Step 1 : Import Packages**

# import packages

import pandas as pd

import numpy as np

import plotly.express as px

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

from sklearn.feature\_selection import RFE

from sklearn.metrics import mean\_squared\_error

**Step 2: Data Access/Structuring**

The data for this project was obtained from the cBioPortal for Cancer Genomics database. This is a publicly available database that is accessible by anyone. This database comprises multiple cancer types, but for the purposes of this project, only the biliary tract cancer was of interest. In the database there were multiple sets of data for this specific cancer type, however each had different variables that were unique to that set of data. For this project, the specific dataset that was utilized was “Intrahepatic Cholangiocarcinoma (MSK, Hepatology 2021). This dataset consisted of only intrahepatic cholangiocarcinoma. This specific type of cholangiocarcinoma develops in the biliary cells within the bile ducts inside the liver. The raw downloaded data was in a csv format, and this was read by python for data analysis.

The final structural format for this dataset was a pandas data frame. To get the data set in a desired format, the .csv file was first converted into a pandas data frame.

data = pd.read\_csv('ihch\_msk\_2021\_clinical\_data.csv')

# visualize the first 5 rows of the data

data.head()

**Step 3 :Data Quality**

Quality steps were taken to assess the dataset assessing the impact the missing values had on addressing the needs of the research question. Initial modeling was conducted to determine the inclusion and exclusion criteria for the dataset and parse the dataset in the different categorical and continuous variables assessing the outcomes.

# check which columns have null values

missing\_values = data.isnull().sum()

#print(missing\_values)

with pd.option\_context('display.max\_rows', None,

'display.max\_columns', None,

'display.precision', 3,

):

print(missing\_values)

**# check again which columns have null values**

**missing\_values = data1.isnull().sum()**

**with pd.option\_context('display.max\_rows', None,**

**'display.max\_columns', None,**

**'display.precision', 3,**

**):**

**print(missing\_values)**

**# print(missing\_values)**

The columns that were not needed were the first ones to be excluded from the data frame. The rationale for excluding these columns ranged from more than 80% of the data missing to the data not being relevant to address this question.

**# drop by the columns that are not important for this analysis. Drop by column Name**

**data1 = data.drop(['Tumor Size','Systemic Chemotherapy', 'Steatosis','RFS Status','RFS Months','Positive Margin','Positive Lymoh Node','PNI','PD INF','OS Months from RX','Neoadjuvant Chemotherapy','LVI','Duct Type','Adjuvant Chemotherapy','Patient ID','Study ID','Cancer Type','Cancer Type Detailed','OncoTree Code'], axis=1)**

**# check again which columns have null values**

**missing\_values = data1.isnull().sum()**

**with pd.option\_context('display.max\_rows', None,**

**'display.max\_columns', None,**

**'display.precision', 3,**

**):**

**print(missing\_values)**

**# print(missing\_values)**

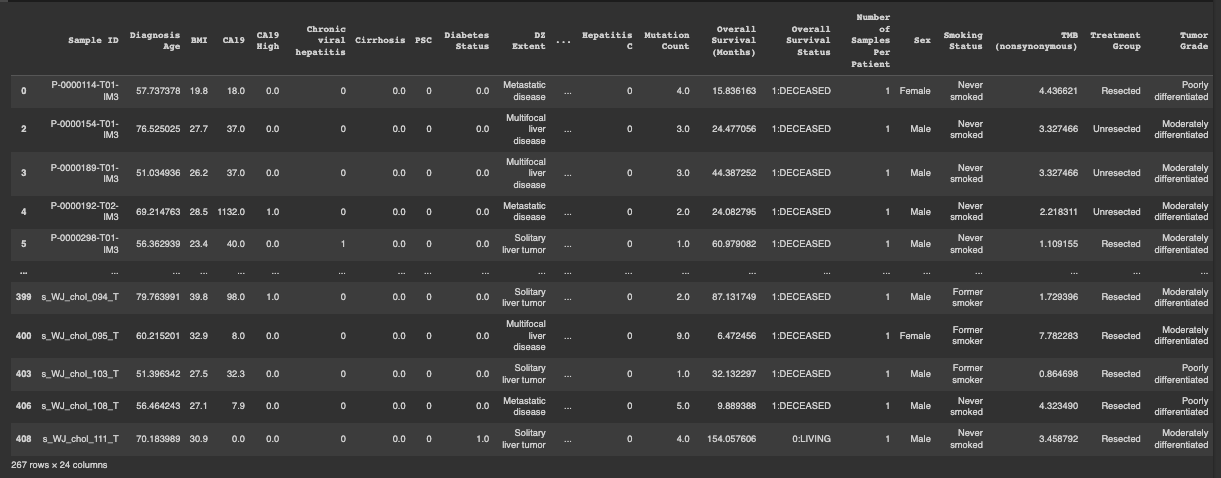
After these were removed, then rows with missing values were then removed. The final dataset contained 267 patients with 24 columns

**#remove all the rows that missing values from the dataset**

**# clean\_data is a new dataframe**

**clean\_data = data1.dropna()**

**clean\_data**

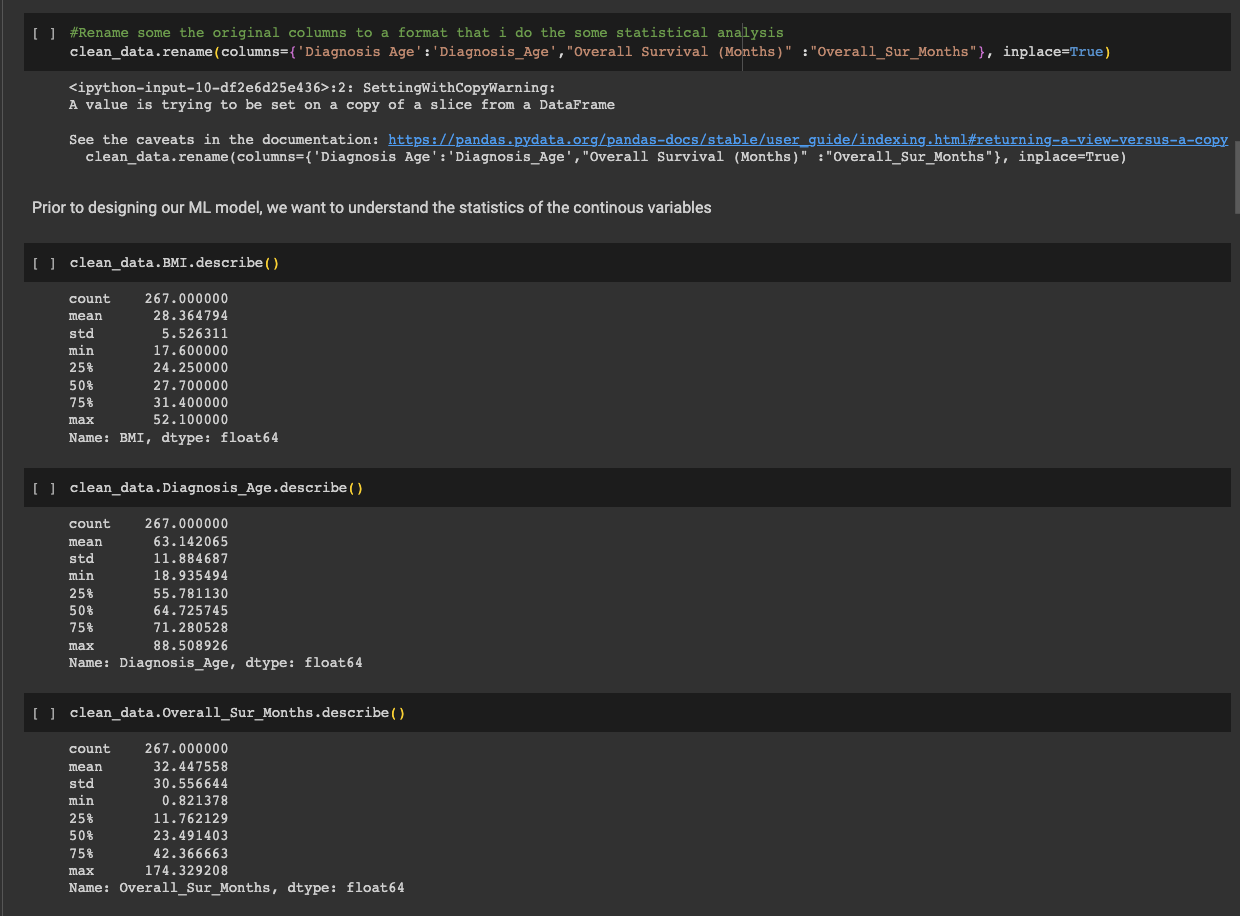
****

**Step 4 : Exploratory Data Analysis**

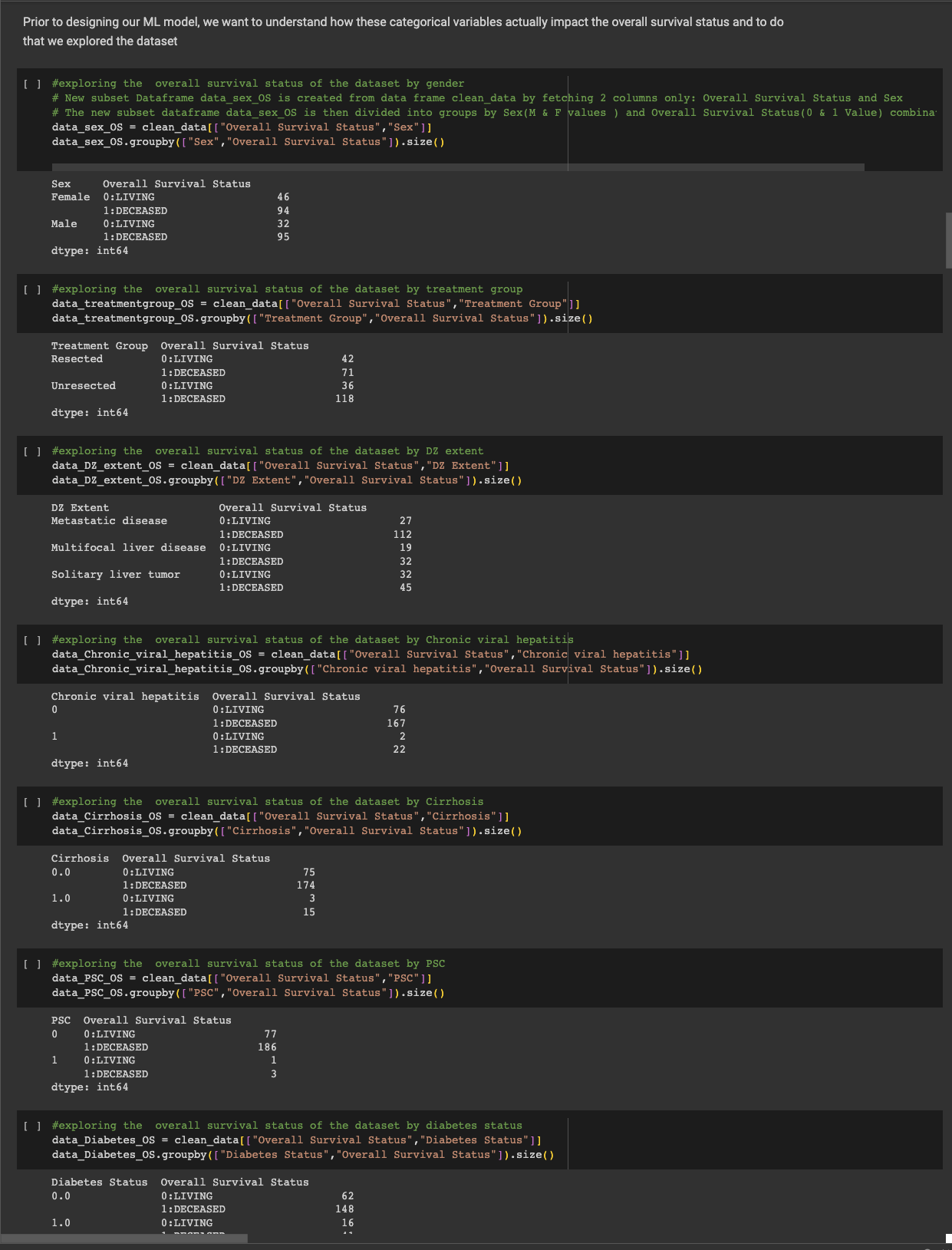
Inorder to choose the right analysis approach we conducted an exploratory data analysis for both the numerical and categorical variables.

The numerical variables explored were BMI, Diagnosis Age and Overall survival in months.

Step 4.1 : Numerical Variables

****

Step 4.2: Categorical Variables

****

**Step 5: Machine learning model used: Logistic Regression**

**Rationale:** The overall survival outcome of these patients is either 1:Deceases or 0: Living making it a binary classification and considering that this is a relatively small dataset, logistic regression can perform well as it requires fewer data points compared to more complex models.

However the use of more complex models will be useful for comparison.

# for the independent categorical variables, there will be convert to numeric using one-hot encoding

# Convert categorical variable 'gender' to numeric using one-hot encoding

clean\_data2 = pd.get\_dummies(clean\_data, columns=['DZ Extent',

'ECOG BIN',

'Sex','Smoking Status',

'Treatment Group',

'Tumor Grade',

'HAIC','Hepatitis B',

'Hepatitis C',], drop\_first=True)

clean\_data2

#information about the data types for each column

clean\_data2.info()

# Split the data into features (X) and target variable (y)

X = clean\_data2[['Diagnosis\_Age', 'BMI', 'CA19','CA19 High','Chronic viral hepatitis','Cirrhosis','PSC','Diabetes Status','Fraction Genome Altered','Mutation Count','Overall\_Sur\_Months','Number of Samples Per Patient','TMB (nonsynonymous)','DZ Extent\_Multifocal liver disease','DZ Extent\_Solitary liver tumor','ECOG BIN\_1','ECOG BIN\_3-Feb','Sex\_Male','Smoking Status\_Former smoker','Smoking Status\_Never smoked','Treatment Group\_Unresected','Tumor Grade\_Poorly differentiated','Tumor Grade\_Well differentiated','HAIC\_1','Hepatitis B\_1','Hepatitis C\_1']]

y = clean\_data2['Overall Survival Status']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Initialize the Logistic Regression model

logreg = LogisticRegression()

# Initialize RFE with the linear regression model

rfe = RFE(estimator=logreg, n\_features\_to\_select=1)

# Fit RFE on the training data

rfe.fit(X\_train, y\_train)

# Get the ranking of features (1: most important, 2: second most important, etc.)

feature\_ranking = rfe.ranking\_

feature\_ranking

# Select the most important feature(s) based on the ranking

selected\_features = X\_train.columns[rfe.support\_]

# Train the model on the selected features

logreg.fit(X\_train[selected\_features], y\_train)

# Make predictions on the test data

y\_pred = logreg.predict(X\_test[selected\_features])

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

classification\_rep = classification\_report(y\_test, y\_pred)

print("Accuracy:", accuracy)

print("Confusion Matrix:")

print(conf\_matrix)

print("Classification Report:")

print(classification\_rep)

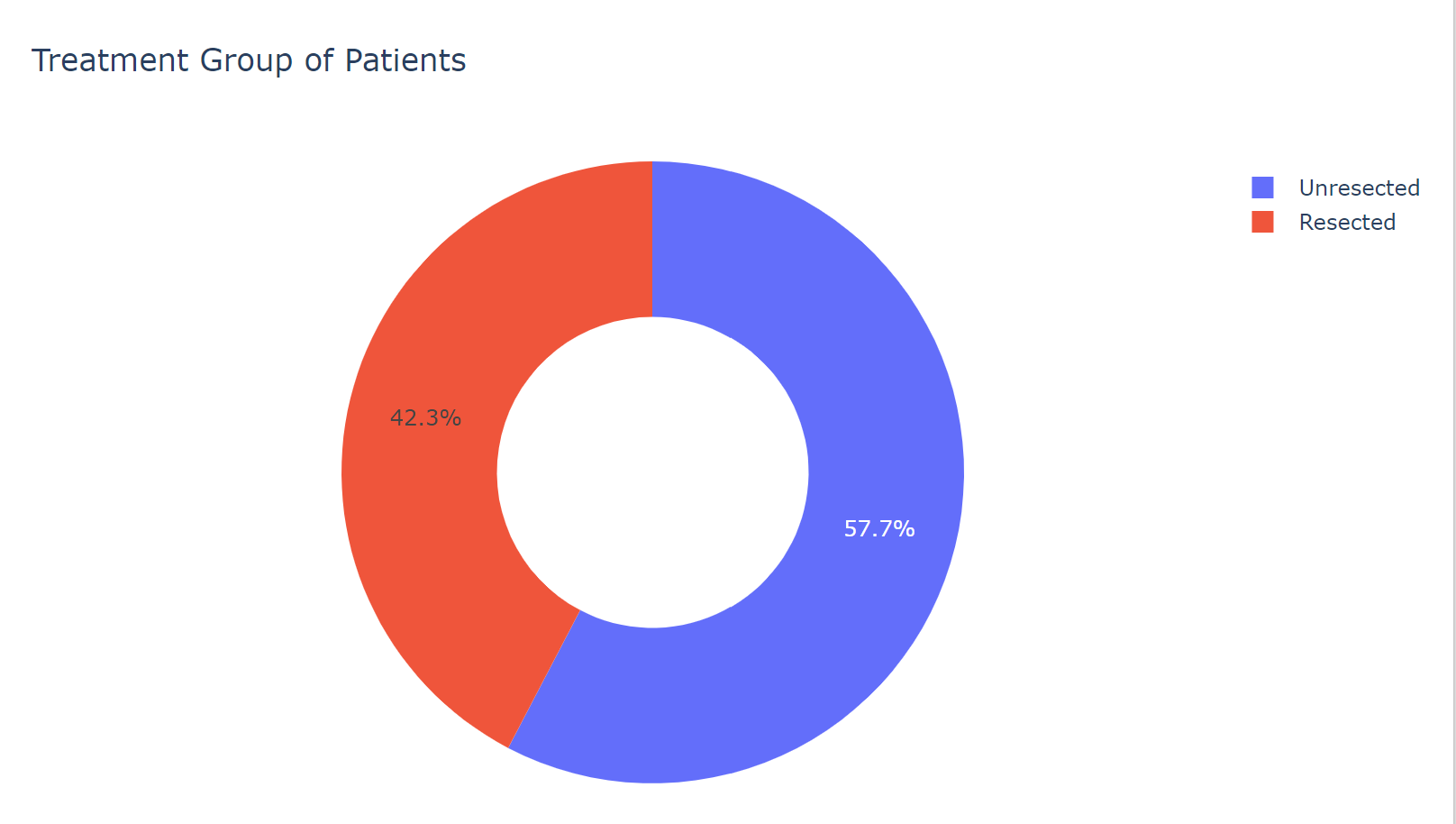
**Result:**

Based on Exploration of the data some interesting Visualizations were attained and divided in terms of

**Treatment Group of Patients:**

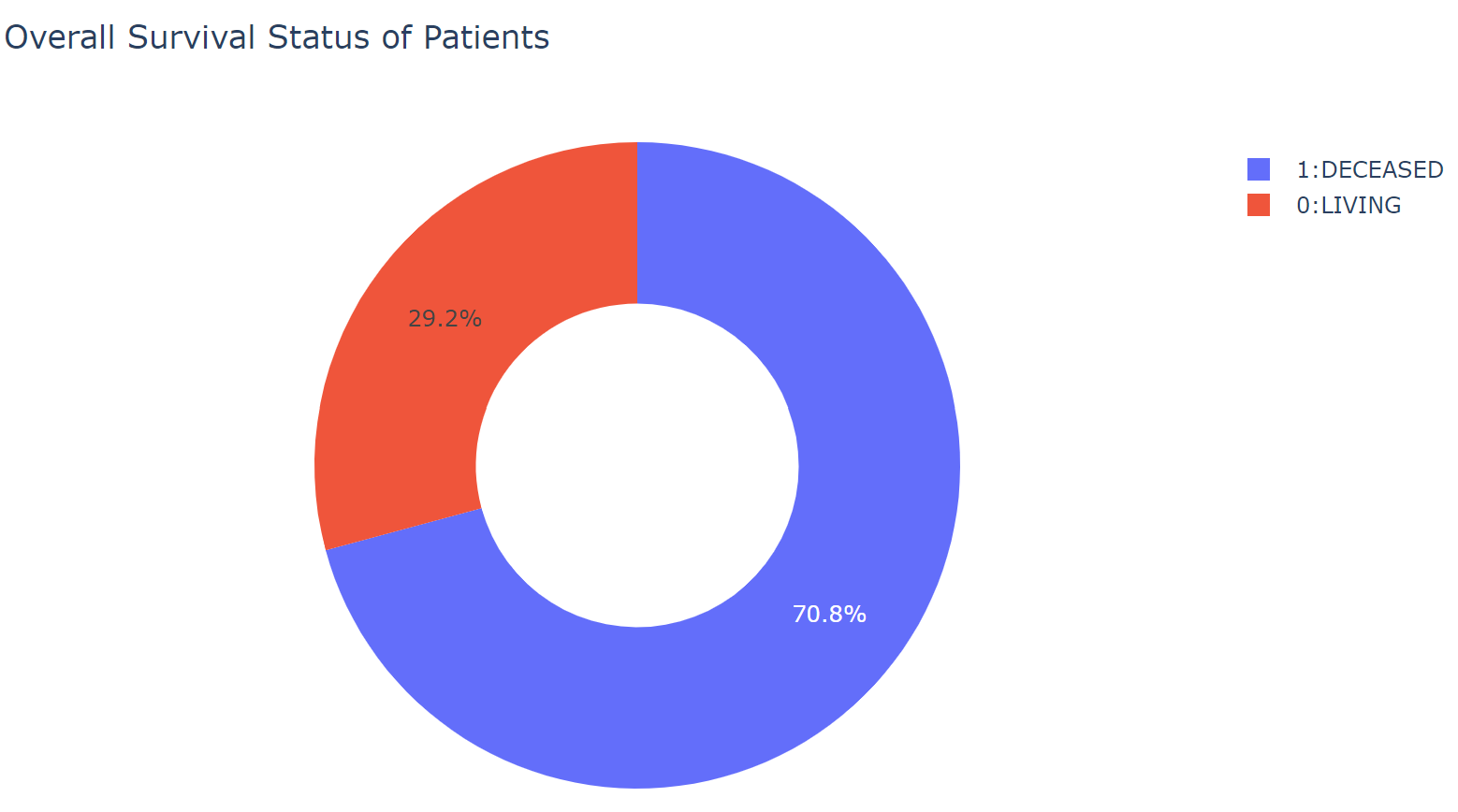
**Unresected group of patients**: This refers to a group of patients who have not undergone surgical resection. It means that the tumor or affected tissue has not been removed through surgery.

**Resected group of patients:** This refers to a group of patients who have undergone surgical resection. It means that the tumor or affected tissue has been surgically removed from the body.

****

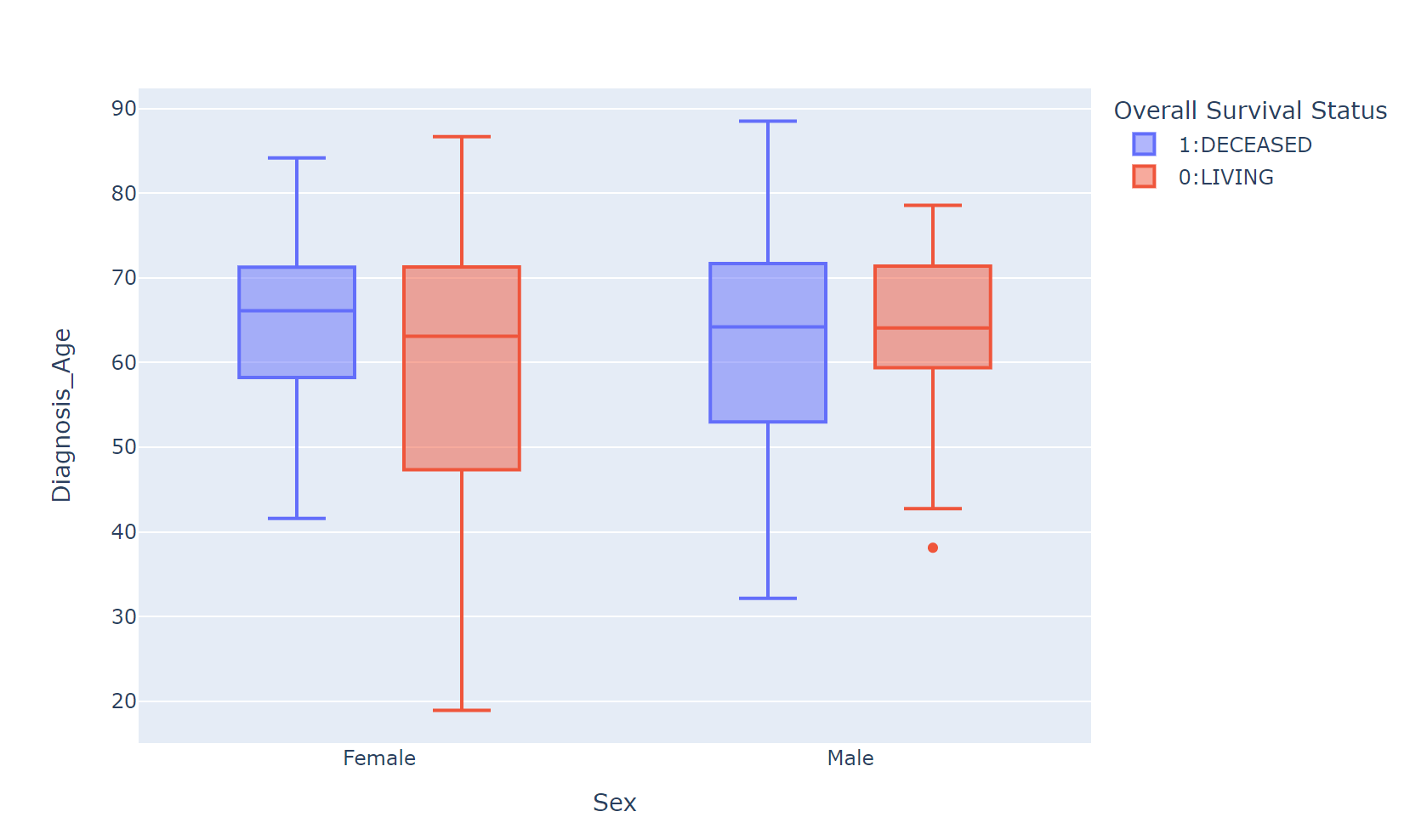
**Overall Survival Status of Patients:**

**70.8% patients who were unresected passed away** and the remaining **29.2% resected patients survived**.The information does not provide details about the specific conditions being treated, the time frame over which these outcomes were observed, or any other factors that might have influenced the results.For understanding of the effectiveness of surgical resection and the prognosis for patients with specific medical conditions,factors such as the stage of the disease, the health status of the patients, the presence of any other conditions, and the overall quality of medical care provided are all to be considered.

****

**Overall Survival Status Based on Sex & Diagnosis Age**

The Box plot explains among females those who are deceased lived up to 84 years, The third quartile is the value that separates the lower 75% of the data from the upper 25% separates at age of 71.27 meaning 25 percentile of the deceased female population was 71 years. Median age is depicted as 66 and the first quartile that separates the lower 25% of the data from the upper 75%. separates at 58.24 and the minimum age is depicted at 18

****

**Discussion:**

The machine learning model developed in this project aimed to predict the overall survival status in patients with cholangiocarcinoma (CCA). Several key findings and insights emerged from the analysis, which should be discussed:

1. **Data Quality and Preprocessing:** The initial dataset obtained from cBioPortal for Cancer Genomics database was subjected to rigorous data cleaning and preprocessing. Missing values were handled by removing rows with missing data, and irrelevant columns were dropped. This ensured that the dataset used for modeling was of high quality and relevance to the research question.

2. **Exploratory Data Analysis (EDA)**: EDA was conducted to gain a deeper understanding of the data. The analysis explored both numerical and categorical variables. Notable observations include the distribution of overall survival status among patients who underwent resection versus those who did not, as well as insights into age and gender-based survival trends.

3. **Machine Learning Model Selection**: Logistic Regression was chosen as the machine learning model for predicting overall survival status. This choice was made based on the binary nature of the target variable (survived or deceased) and the relatively small dataset size. Logistic Regression is known for its simplicity and interpretability, making it suitable for this context. It's worth noting that more complex models could be explored in future studies for comparison.

4. **Feature Selection:** Recursive Feature Elimination (RFE) was employed to select the most important features for the model. This process helps in identifying the key predictors that have the most influence on the outcome. These selected features were then used to train the logistic regression model.

5. **Model Evaluation:** The model's performance was evaluated using standard classification metrics, including accuracy, confusion matrix, and classification report. These metrics provide insights into the model's ability to correctly classify patients' survival status.

**Conclusion:**

In conclusion, this project successfully developed a machine learning model using Logistic Regression to predict the overall survival status in patients with cholangiocarcinoma. The model demonstrated promising results in terms of accuracy and provided insights into the factors that influence survival outcomes.

Key findings from the analysis include differences in survival rates between resected and unresected patients, as well as age and gender-based survival trends among the deceased population. However, it's important to note that the findings should be interpreted with caution, and further research is needed to validate and refine the model's predictions.

Future directions for this research could involve the exploration of more advanced machine learning models, incorporating additional clinical and genetic features, and conducting a more comprehensive analysis of patient outcomes. Additionally, external validation of the model on independent datasets would strengthen its reliability for clinical use.

Overall, this project represents a valuable step toward developing predictive models for cholangiocarcinoma patients, potentially aiding in treatment decision-making and improving patient care in the future.

**Reference**