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

Ankita Thomas, Bsc.

Nellie Campbell, Bsc.

Women Who Code Buddy program

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)
```

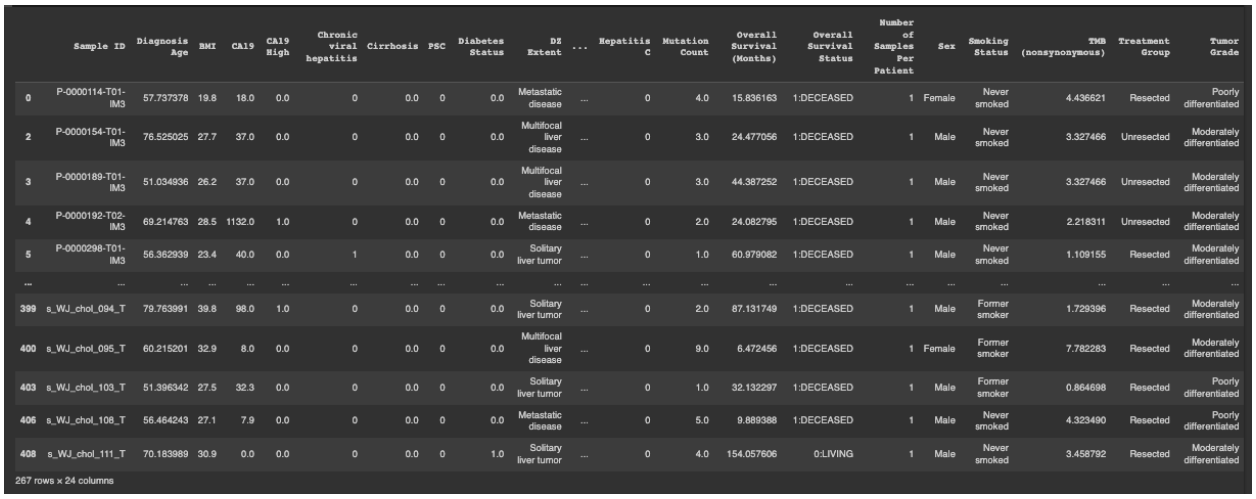
After these were removed, then rows with missing values were then removed. The final dataset contained 267 patients with 24 columns

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

# clean_data is a new dataframe

clean_data = data1.dropna()

clean_data
```



	Sample ID	Diagnosis Age	BMI	CA19	CA19 High	Chronic viral hepatitis	Cirrhosis	PSC	Diabetes Status	DZ Extent	...	Hepatitis C	Mutation Count	Overall Survival (Months)	Overall Survival Status	Number of Samples Per Patient	Sex	Smoking Status	TMB (nonsynonymous)	Treatment Group	Tumor Grade
0	P-0000114-T01-IM3	57.737378	19.8	18.0	0.0	0	0.0	0	0.0	Metastatic disease	...	0	4.0	15.836163	1:DECEASED	1	Female	Never smoked	4.436621	Resected	Poorly differentiated
2	P-0000154-T01-IM3	76.525025	27.7	37.0	0.0	0	0.0	0	0.0	Multifocal liver disease	...	0	3.0	24.477056	1:DECEASED	1	Male	Never smoked	3.327466	Unresected	Moderately differentiated
3	P-0000189-T01-IM3	51.034936	26.2	37.0	0.0	0	0.0	0	0.0	Multifocal liver disease	...	0	3.0	44.387252	1:DECEASED	1	Male	Never smoked	3.327466	Unresected	Moderately differentiated
4	P-0000192-T02-IM3	69.214763	28.5	1132.0	1.0	0	0.0	0	0.0	Metastatic disease	...	0	2.0	24.082795	1:DECEASED	1	Male	Never smoked	2.218311	Unresected	Moderately differentiated
5	P-0000298-T01-IM3	56.362939	23.4	40.0	0.0	1	0.0	0	0.0	Solitary liver tumor	...	0	1.0	60.979082	1:DECEASED	1	Male	Never smoked	1.109155	Resected	Moderately differentiated
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
399	s_WJ_chol_094_T	79.763991	39.8	98.0	1.0	0	0.0	0	0.0	Solitary liver tumor	...	0	2.0	87.131749	1:DECEASED	1	Male	Former smoker	1.729396	Resected	Moderately differentiated
400	s_WJ_chol_095_T	60.215201	32.9	8.0	0.0	0	0.0	0	0.0	Multifocal liver disease	...	0	9.0	6.472456	1:DECEASED	1	Female	Former smoker	7.782283	Resected	Moderately differentiated
403	s_WJ_chol_103_T	51.396342	27.5	32.3	0.0	0	0.0	0	0.0	Solitary liver tumor	...	0	1.0	32.132297	1:DECEASED	1	Male	Former smoker	0.864698	Resected	Poorly differentiated
406	s_WJ_chol_108_T	56.464243	27.1	7.9	0.0	0	0.0	0	0.0	Metastatic disease	...	0	5.0	9.889388	1:DECEASED	1	Male	Never smoked	4.323490	Resected	Poorly differentiated
408	s_WJ_chol_111_T	70.183989	30.9	0.0	0.0	0	0.0	0	1.0	Solitary liver tumor	...	0	4.0	154.057606	0:LIVING	1	Male	Never smoked	3.458792	Resected	Moderately differentiated

267 rows × 24 columns

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

```
[ ] #Rename some the original columns to a format that i do the some statistical analysis
    clean_data.rename(columns={'Diagnosis Age':'Diagnosis_Age',"Overall Survival (Months)" :"Overall_Sur_Months"}, inplace=True)

    <ipython-input-10-df2e6d25e436>:2: SettingWithCopyWarning:
    A value is trying to be set on a copy of a slice from a DataFrame

    See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
      clean_data.rename(columns={'Diagnosis Age':'Diagnosis_Age',"Overall Survival (Months)" :"Overall_Sur_Months"}, inplace=True)
```

Prior to designing our ML model, we want to understand the statistics of the continous variables

```
[ ] clean_data.BMI.describe()

    count    267.000000
    mean      28.364794
    std        5.526311
    min       17.600000
    25%       24.250000
    50%       27.700000
    75%       31.400000
    max       52.100000
    Name: BMI, dtype: float64

[ ] clean_data.Diagnosis_Age.describe()

    count    267.000000
    mean      63.142065
    std       11.884687
    min       18.935494
    25%       55.781130
    50%       64.725745
    75%       71.280528
    max       88.508926
    Name: Diagnosis_Age, dtype: float64

[ ] clean_data.Overall_Sur_Months.describe()

    count    267.000000
    mean      32.447558
    std       30.556644
    min        0.821378
    25%       11.762129
    50%       23.491403
    75%       42.366663
    max      174.329208
    Name: Overall_Sur_Months, dtype: float64
```

Step 4.2: Categorical Variables

Prior to designing our ML model, we want to understand how these categorical variables actually impact the overall survival status and to do that we explored the dataset

```
[ ] #exploring the  overall survival status of the dataset by gender
    # New subset Dataframe data_sex_OS is created from data frame clean_data by fetching 2 columns only: Overall Survival Status and Sex
    # The new subset dataframe data_sex_OS is then divided into groups by Sex(M & F values ) and Overall Survival Status(0 & 1 Value) combina
    data_sex_OS = clean_data[["Overall Survival Status","Sex"]]
    data_sex_OS.groupby(["Sex","Overall Survival Status"]).size()

    Sex     Overall Survival Status
    Female  0:LIVING                  46
            1:DECEASED                94
    Male    0:LIVING                  32
            1:DECEASED                95
    dtype: int64

[ ] #exploring the  overall survival status of the dataset by treatment group
    data_treatmentgroup_OS = clean_data[["Overall Survival Status","Treatment Group"]]
    data_treatmentgroup_OS.groupby(["Treatment Group","Overall Survival Status"]).size()

    Treatment Group  Overall Survival Status
    Resected         0:LIVING                  42
                     1:DECEASED                71
    Unresected       0:LIVING                  36
                     1:DECEASED               118
    dtype: int64

[ ] #exploring the  overall survival status of the dataset by DZ extent
    data_DZ_extent_OS = clean_data[["Overall Survival Status","DZ Extent"]]
    data_DZ_extent_OS.groupby(["DZ Extent","Overall Survival Status"]).size()

    DZ Extent                Overall Survival Status
    Metastatic disease       0:LIVING                  27
                             1:DECEASED               112
    Multifocal liver disease 0:LIVING                  19
                             1:DECEASED                32
    Solitary liver tumor     0:LIVING                  32
                             1:DECEASED                45
    dtype: int64

[ ] #exploring the  overall survival status of the dataset by Chronic viral hepatitis
    data_Chronic_viral_hepatitis_OS = clean_data[["Overall Survival Status","Chronic viral hepatitis"]]
    data_Chronic_viral_hepatitis_OS.groupby(["Chronic viral hepatitis","Overall Survival Status"]).size()

    Chronic viral hepatitis  Overall Survival Status
    0                        0:LIVING                  76
                             1:DECEASED               167
    1                        0:LIVING                   2
                             1:DECEASED                22
    dtype: int64

[ ] #exploring the  overall survival status of the dataset by Cirrhosis
    data_Cirrhosis_OS = clean_data[["Overall Survival Status","Cirrhosis"]]
    data_Cirrhosis_OS.groupby(["Cirrhosis","Overall Survival Status"]).size()

    Cirrhosis  Overall Survival Status
    0.0        0:LIVING                  75
               1:DECEASED               174
    1.0        0:LIVING                   3
               1:DECEASED                15
    dtype: int64

[ ] #exploring the  overall survival status of the dataset by PSC
    data_PSC_OS = clean_data[["Overall Survival Status","PSC"]]
    data_PSC_OS.groupby(["PSC","Overall Survival Status"]).size()

    PSC  Overall Survival Status
    0    0:LIVING                  77
         1:DECEASED               186
    1    0:LIVING                   1
         1:DECEASED                 3
    dtype: int64

[ ] #exploring the  overall survival status of the dataset by diabetes status
    data_Diabetes_OS = clean_data[["Overall Survival Status","Diabetes Status"]]
    data_Diabetes_OS.groupby(["Diabetes Status","Overall Survival Status"]).size()

    Diabetes Status  Overall Survival Status
    0.0              0:LIVING                  62
                     1:DECEASED               148
    1.0              0:LIVING                  16
```

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:

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.

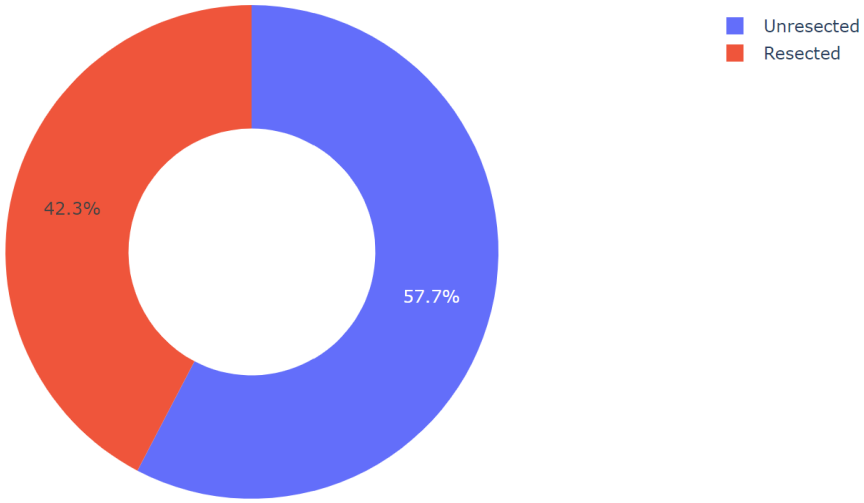


Figure 1 : Chart showing the treatment group of patients

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.

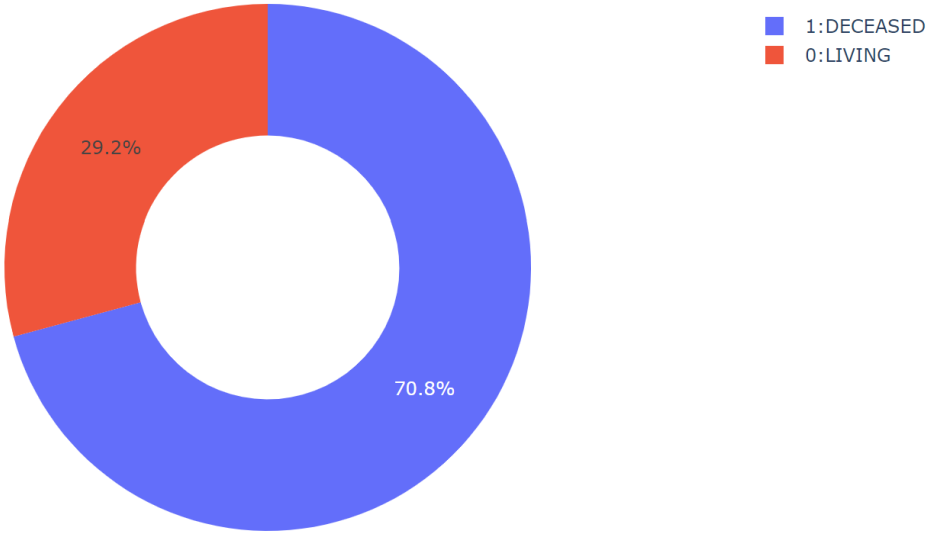


Figure 2: Chart showing the overall survival status of patients

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

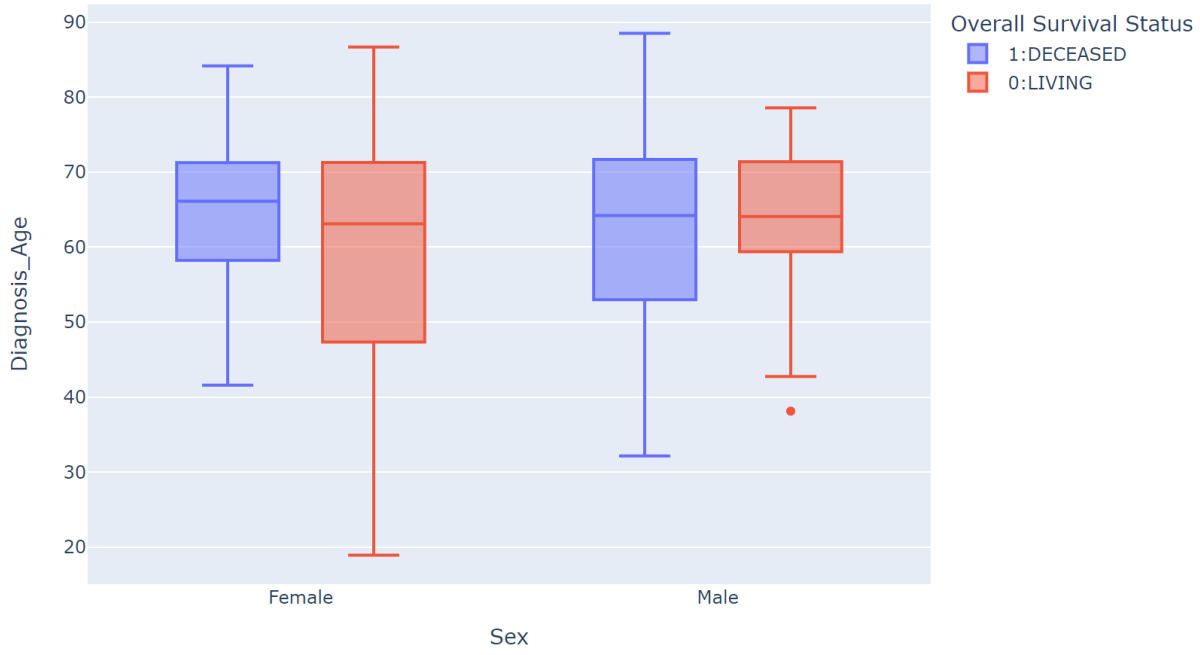


Figure 3: Boxplot showing the age distribution for the overall survival status partitioned by sex

Logistic regression model

The overall Accuracy of the logistic regression model was 0.70. When the dataset is balanced, which means that the target variable's classes (categories) are distributed fairly evenly, accuracy is frequently employed as the key performance indicator. The most important evaluation metric in this logistic regression model is shown in table 1.

Table 1 : Classification Report of the Logistic regression model

	precision	**recall**	**f1-score**	**support**
0	**0.00**	**0.00**	**0.00**	**24**
1	**0.70**	**1.00**	**0.83**	**57**

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.

References

1. Rizvi S, Khan SA, Hallemeier CL, Kelley RK, Gores GJ. Cholangiocarcinoma evolving concepts and therapeutic strategies. Nat Rev Clin Oncol 2018;15:95-111.
2. Zori AG, Yang D, Draganov PV, Cabrera R. Advances in the management of cholangiocarcinoma. World J Hepatol 2021;13:1003-1018.
3. Data was obtained from The Cancer Genome atlas (TCGA) https://www.cbioportal.org/study/clinicalData?id=ihch_msk_2021