****

**MICROFLUIDICS**

*Non-invasive procedure to predict Pancreatic Cancer in patients via liquid biomarkers*

**PROJECT REPORT**

**Prediction of Pancreatic Cancer using bio-markers**

**Under guidance of**

***Prof.*** [***Madhusudhanan Baskaran***](https://www.linkedin.com/in/drmadhub/overlay/about-this-profile/)

**By**

***Mohammed Aiman Khan***

***(18th June 2024 -18th July 2024)***

****

**CONTENTS**

[**ABSTRACT**](#_45hs213hpeep) **3**

[**LITERATURE REVIEW FOR NON-INVASIVE SOLUTIONS**](#_dzrifypri7cu) **4**

[Current Challenge](#_qs57ev9wocag) 4

[Medical Findings](#_5kmm75bt7ddo) 7

[**MARKET ANALYSIS   
(for Non-Invasive Biomarkers):**](#_kgkwst9aw46t) **8**

[Urinalysis Market Size & Trends](#_x0tlildoju78) 8

[Insights of Market Demand:](#_vzajlz3pr39d) 9

[Pancreatic cancer rates](#_vdajt7tjpehe) 10

[**ROLE OF DATA SCIENCE AND MACHINE LEARNING**](#_85qze0pth9az) **12**

[**FEASIBLE SOLUTION**](#_dzwfznepbm9i) **14**

[**TECHNICAL APPROACH:**](#_1oqagia8ln0z) **16**

[INSTALLING AND IMPORTING OF NECESSARY LIBRARIES FOR ACTION ON DATA](#_v5ga967lg50v) 16

[DATA PROCESSING:](#_xtxpnnvtxc63) 18

[Data Exploration:](#_1yaxqujmxeam) 18

[Data Separation as X and Y](#_za855dpzgq3) 21

[EDA - Data cleaning](#_x140s8i8pi3a) 22

[Data Populating:](#_48rhluzcqri6) 27

[ANALYSIS](#_it3o414nth8) 30

[Numerical Features Analysis](#_uqmvu2xvxt) 30

[Categorical Features Analysis](#_lxvmnf42hhg) 35

[CORRELATION](#_6v96ykwqpa0f) 39

[Correlation Matrix to find the relationship between features:](#_vf35oqsh600a) 39

[PAIR PLOTTING between biomarker features:](#_80ro1vs957a3) 42

[Model matrices Details:](#_gyc0uvgdq9hr) 43

[MULTICLASS MODEL and PREDICTIONS:](#_vsa5eldspb7j) 44

[MODEL BUILDING:](#_f2ezyfs2jux0) 44

[PREDICTING:](#_39i79cka7zhh) 47

[VALIDATING Predictions:](#_vg7tn6j85l6p) 48

[BINARY MODEL and PREDICTIONS:](#_sfkcc6cfyqgw) 49

[DATA CHANGES to Create a Binary Model:](#_t73n1v8y5tgo) 49

[MODEL BUILDING:](#_ndjolgms3897) 49

[PREDICTING:](#_hngrapgge2de) 52

[VALIDATING Predictions:](#_3ve7wzldf8os) 53

[**CONCLUSIONS AND FURTHER SCOPE**](#_kc8peapazoo6) **54**

[TECHNICAL CONCLUSION:](#_qy3txnduidxd) 54

[FURTHER SCOPE on Technical Front:](#_taa8f3xdvss8) 54

[PROJECT CONCLUSION:](#_7zisw5m2td6u) 54

[**REFERENCES**](#_eg55bgjpd140) **55**

# ABSTRACT

Pancreatic cancer is a formidable challenge due to its late diagnosis and poor prognosis. This project report investigates the landscape of pancreatic cancer diagnosis, focusing on non-invasive biomarkers for early detection. A thorough literature review explores pancreatic cancer biology and the role of biomarkers such as CA 19-9 and CEA, identified through genomic and proteomic analyses.

In addition, the report analyses the market for self-testing kits designed to detect pancreatic cancer early, utilising advancements in urine analysis. The evaluation assesses the accessibility and effectiveness of these diagnostic tools.

The study includes a detailed technical analysis of biomarkers identified through urine tests. A machine learning project was conducted on biomarkers from 590 individuals. A predictive model was trained and tested, highlighting key biomarkers crucial for pancreatic cancer diagnosis.

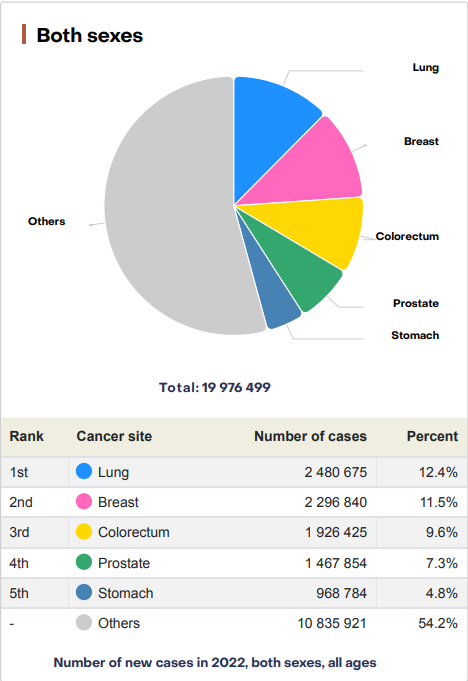
This report underscores the significance of early detection in improving patient outcomes and reducing mortality rates. It concludes by discussing future research directions to enhance early detection capabilities and save lives.

# 

# LITERATURE REVIEW FOR NON-INVASIVE SOLUTIONS

## Current Challenge

Cancer is one of the most common and deadly diseases worldwide, claiming millions of lives yearly. Despite significant advances in treatment, the overall survival rate remains low, primarily due to late-stage diagnosis.

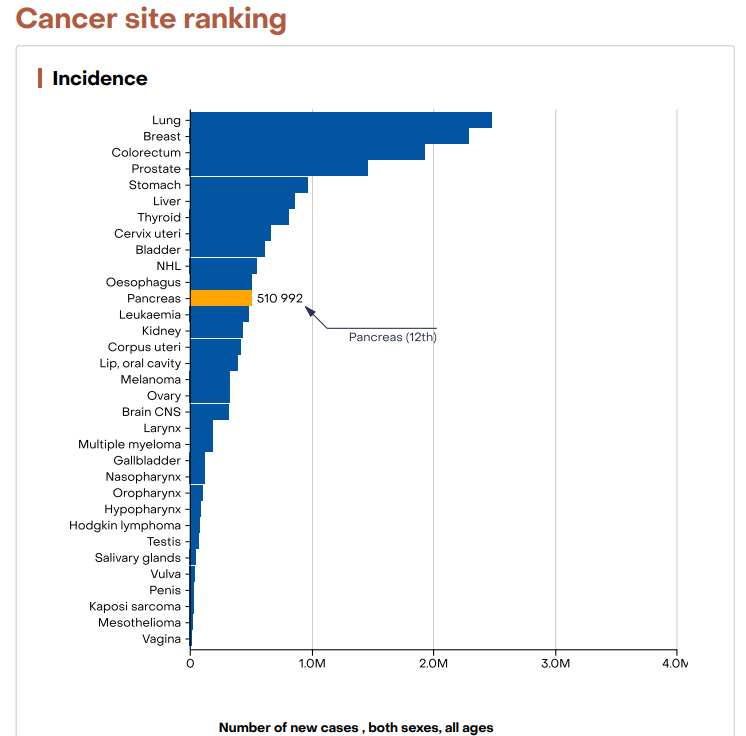


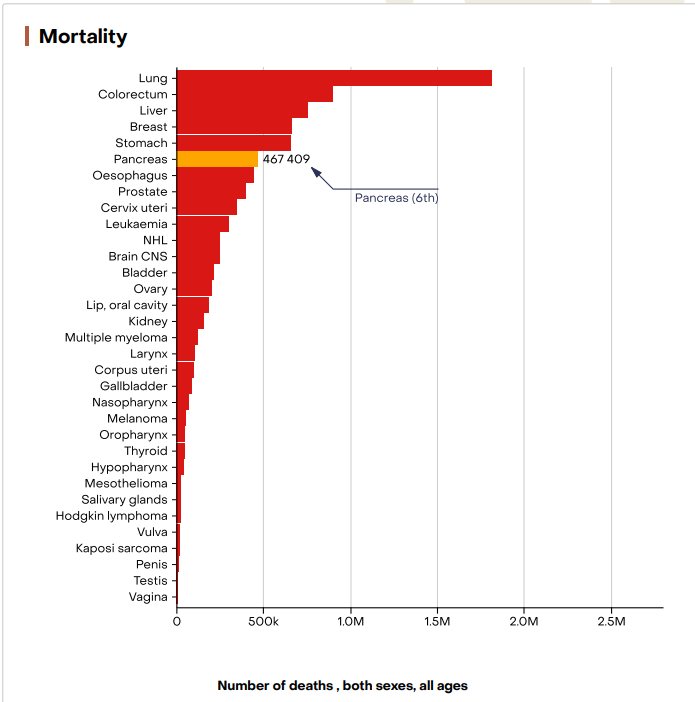
Pancreatic cancer is one of the most lethal gastrointestinal malignancies. Despite advances in cross-sectional imaging, chemotherapy, radiation therapy, and surgical techniques, the 5-year overall survival is only 12%.

Pancreatic cancer (PC) is the fourth leading cause of cancer-related mortality in the United States and, if current trends continue, will become the second most fatal cancer by 2030 . PC is transforming into a worldwide crisis with an estimated incidence of 355,317 cases by 2040 . Over 90% of PC develops from cells of the exocrine system, such as pancreatic ductal adenocarcinoma (PDAC) and pancreatic acinar cell carcinoma,PDAC is considered a silent killer due to a largely asymptomatic course, late clinical presentation, and rapid progression , with fewer than 20% of PDAC patients being diagnosed at early, resectable stages I and II .

Importantly, if PDAC is detected when still a localised disease, the 5-year survival is around 32% , but it can approach 70% following resection of incidentally diagnosed stage I tumours .

It is estimated that it takes at least 10 years between initiating mutation and the birth of parental founder ; combined with an increased PDAC incidence with age , resectable PDAC has likely already been growing for a number of years. We therefore further on use the term “earlier” rather than “early” stage. while the remainder of pancreatic tumours originate from endocrine cells, called neuroendocrine tumours, or pancreatic cystic lesions (PCLs), including intraductal papillary mucinous neoplasm (IPMN) and mucinous cystic neoplasms (MCNs). Some PCLs carry a risk of malignant transformation and an opportunity for curative resection prior to malignant transformation if detected early . Early-stage PC is often indolent, presenting with vague abdominal symptoms, jaundice, weight loss, or no symptoms at all. As a result, only 10–15% of patients have a resectable disease at diagnosis while approximately 50% present with a metastatic disease.





Although a biopsy is not needed prior to surgical resection when clinical suspicion is high, standard staging workup for PC involves a multiphasic cross-sectional imaging of the chest, abdomen, and pelvis with thin cuts through the pancreas. Furthermore, endoscopic interventions including esophagogastroduodenoscopy (EGD) with an endoscopic ultrasound (EUS) and fine-needle biopsy (FNB) or endoscopic retrograde cholangiopancreatography (ERCP) may be needed prior to the initiation of neoadjuvant chemotherapy when a tissue diagnosis is required or when the patient needs relief from biliary obstruction .

To date, there exists no reliable, non-invasive screening test, either molecular or imaging-based, that will allow accurate PDAC detection at an earlier stage in asymptomatic patients. The conventional methods such as computerised tomography, magnetic resonance imaging or endoscopic ultrasonography are expensive and have low sensitivity (SN) and specificity (SP) for detection of small premalignant lesions .

Furthermore,a lot of research is in-progress on non-invasive methods for diagnosing.

## Medical Findings

National Library of Medicine publication “Identification of a Three-Biomarker Panel in Urine for Early Detection of Pancreatic Adenocarcinoma”, published the below details of three non-invasive protein biomarkers which can detect pancreatic cancer

LYVE-1, REG1A, and TFF1 were selected as candidate biomarkers. When comparing PDAC (n = 192) with healthy (n = 87) urine specimens, the resulting areas under the receiver-operating characteristic curves (AUC) of the panel were 0.89 [95% confidence interval (CI), 0.84-0.94] in the training (70% of the data) and 0.92 (95% CI, 0.86-0.98) in the validation (30% of the data) datasets. When comparing PDAC stage I-II (n = 71) with healthy urine specimens, the panel achieved AUCs of 0.90 (95% CI, 0.84-0.96) and 0.93 (95% CI, 0.84-1.00) in the training and validation datasets, respectively. In PDAC stage I-II and healthy samples with matching plasma CA19.9, the panel achieved a higher AUC of 0.97 (95% CI, 0.94-0.99) than CA19.9 (AUC = 0.88; 95% CI, 0.81-0.95, P = 0.005). Adding plasma CA19.9 to the panel increased the AUC from 0.97 (95% CI, 0.94-0.99) to 0.99 (95% CI, 0.97-1.00, P = 0.04), but did not improve the comparison of stage I-IIA PDAC (n = 17) with healthy urine.

This proves that using non-invasive biomarkers pancreatic cancer can be predicted.

Another journal “A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study” which aimed to establish the accuracy of an improved panel, including REG1B instead of REG1A, and an algorithm for data interpretation, the PancRISK score, in additional retrospectively collected urine specimens. We also assessed the complementarity of this panel with CA19-9 and explored the daily variation and stability of the biomarkers and their performance in common urinary tract cancers.

And concluded that with appropriate substituting REG1A with REG1B. At a pre-selected cutoff of >80% SN and SP for the affiliated PancRISK score, we demonstrate a clinically applicable risk stratification tool with a binary output for risk of developing PDAC ('elevated' or 'normal').

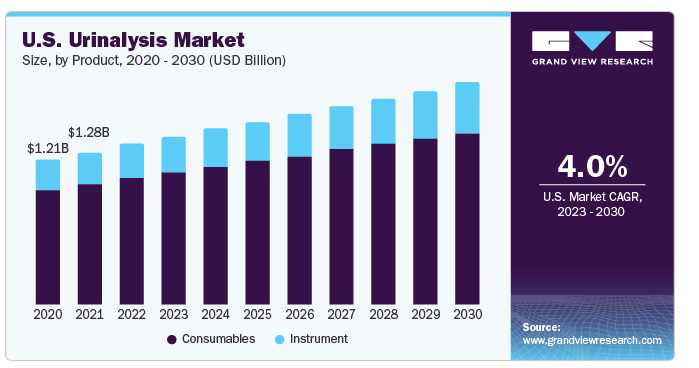
There are a lot of studies being done on the medical front.

Computer science is also playing a huge role to work with data and get accuracy of prediction as near as possible to the truth.

# MARKET ANALYSIS (for Non-Invasive Biomarkers):

## Urinalysis Market Size & Trends

The global [urinalysis market size](https://www.grandviewresearch.com/horizon/outlook/urinalysis-market-size/global) was valued at USD 2.14 billion in 2022 and is anticipated to expand at a compound annual growth rate (CAGR) of 4.7% from 2023 to 2030. The market is witnessing growth due to factors including the rising incidence of diseases such as Urinary Tract Infections (UTIs), diabetes, and kidney diseases. For instance, according to the National Diabetes Statistics Report, 2020, around 34.1 million adults aged 18 years and above had diabetes in the U.S. About one-third of diabetic people have kidney-related comorbidities due to high blood sugar and blood pressure. Urinalysis is an informative and noninvasive diagnostic tool accessible to clinicians to diagnose kidney diseases. Hence, the rising prevalence of kidney ailments is anticipated to increase demand for products used in urinalysis, thereby driving market growth.



## 

## 

## 

## 

## Insights of Market Demand:

The consumables segment held the dominant market share of 78.61% in terms of revenue in 2022. This dominance can be attributed to the increasing demand and frequent purchases of reagents and dipsticks by clinical and hospital laboratories. Dipsticks are the most common type of consumable in urinalysis, accounting for the largest share of the market. Reagents measure the levels of specific substances in the urine. The development of new and innovative consumables, such as rapid test kits and point-of-care testing devices, makes urinalysis more accessible and convenient, which drives demand.

Many major providers, such as Siemens Healthcare GmBH, Abbott, and Roche, provide high-quality consumables for urinalysis testing, which is expected to boost segment growth. For instance, Siemens Healthineers GmBH offers Multistix 10 SG Reagent Strips to diagnose UTI, diabetes, and kidney diseases. In September 2022, Sysmex Corp. announced the launch of its UF-1500 Fully Automated Urine Particle Analyzer that helps in urine sediment testing. The National Kidney Foundation raised over USD 600 million in 2022 to fund research, offer patient care, and advocate for people with kidney disease.

Similar kind of growth story can be looked at for the diagnosis of pancreatic cancer.

When we look at another paper which states the statistics of pancreatic cancer at each major country is listed. Which proves the market growth all over major countries.

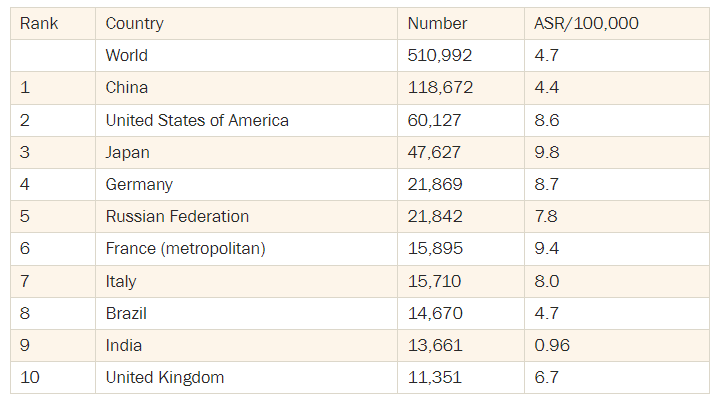


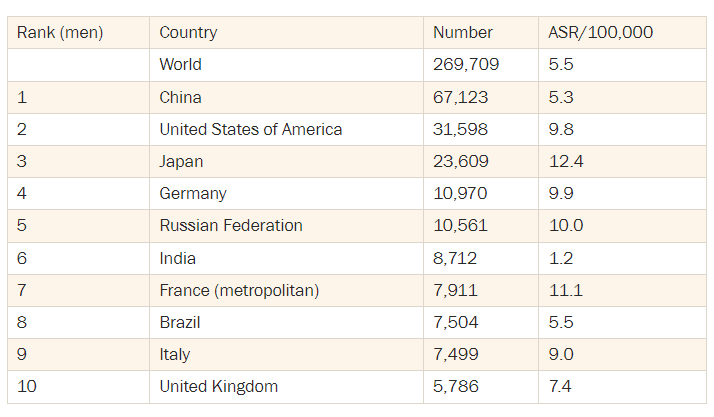
## 

## 

## Pancreatic cancer rates

The following 3 tables show total global pancreatic cancer incidence and rates in 2022, followed by the figures for men and women. Uruguay had the highest overall rate of pancreatic cancer in 2022,.



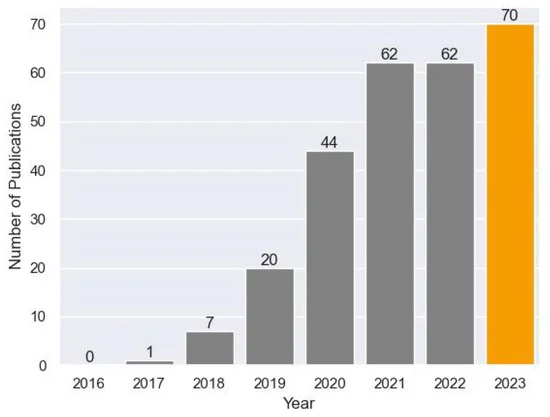




# ROLE OF DATA SCIENCE AND MACHINE LEARNING

Data Science along with ML is a very useful tool , data is collected and acted upon in data science which is then fed to the ML model which makes close to accurate predictions.

Artificial intelligence (AI) includes a large family of techniques to generate simplified representations from complex data that can be used for decision-making or classification . AI is constantly improving the health care sector. With the growing usage of electronic health records, the advances in computer power and continuous monitoring systems and the availability of large data, AI technologies have become the ideal medium for improving health care . Even though it remains expensive and unfeasible to screen the general population for pancreatic cancer with the present technology, the ability to identify susceptible populations with a higher possibility of harbouring such lesions may lead to earlier interception and enhanced survival rates. The use of AI and machine learning as risk stratification tools have the potential to change the diagnosis and the diagnostic landscape. Previous studies reported that AI algorithms have benefited physicians through clinical diagnostic prediction and imaging-based testing.

****

The ability of AI to analyse biomarker data is outstanding. With the use of relevant algorithms, AI can analyse huge databases to detect patterns and correlations within biomarker data. It can help detect relationships between cancer characteristics and biomarkers that may not be detectable by human analysis. AI analysis of patient data including genomics, transcriptomics, proteomics, and clinical information can assist in detecting biomarkers associated with pancreatic cancer that can help in setting treatment plans and prognoses .

Combining multiple biomarkers and analysing them by AI can aid in the development of highly accurate models and algorithms that are capable of detecting pancreatic cancer in its early stages, staging, prognosis prediction, and therapeutic response prediction

ML includes many different computer algorithms with each algorithm having particular advantages and disadvantages. The major types of algorithms are broadly classified as supervised and unsupervised learning. Supervised learning is a form of ML that requires entering both the input and output data into the model dataset, the main goal of which is to learn new functions or predictions on new unseen data. There are two main subsets of supervised learning: classification and regression. In classification, the algorithm is meant to assign the new input data to a specific category, e.g. diagnosing CT images as liver cancer or gallbladder cancer. Meanwhile, in regression, the algorithm derives output based on previous examples of input data, allowing for a backward analysis . Unsupervised learning, on the other hand, is a form of ML where unlabeled input data is used in algorithms to produce patterns and relationships that provide new answers and questions that might have not been previously apparent to the investigators. There are two main subsets of unsupervised learning: clustering and dimensionality reduction. In clustering, algorithms specify and identify different patterns and establish different identification subgroups based on the heterogeneous input data given. Meanwhile, in dimensionality reduction, algorithms aim to decrease the number of subgroups identified with the input data given to simplify analysis and thereby draw fewer differences between different data .

Using ML algorithms, we can apply different datasets that are relevant to pancreatic cancer, including imaging, biomarkers, histology slides, and more, introducing them into a desired algorithm. This could either be under supervised learning, like support vector machine, linear regression, and Naive Bayes, or under an unsupervised algorithm that generates output data like cluster analysis or automated ML. Thereupon, the algorithm runs through a great amount of information that could help researchers with the tedious manual labour of having to go through medical imaging or histology slides one at a time. Automating data analysis can help in diagnosing pancreatic cancer early on, especially with the help of large databases. ML could also potentially play a part in enhancing prognosis and treatment management in a more individualised approach.

# FEASIBLE SOLUTION

Data Science Application to Urine Biomarker Analysis

This will provide a brief overview of the technical project, including the purpose of analysing urine biomarker data and the potential benefits of applying data science techniques. It will also highlight the specific goals of the analysis.

**Data Acquisition and Preparation**

* **Data Collection:** Describe the process of obtaining the urine biomarker dataset. This may involve specifying the source of the data (e.g., hospital, research institution), sample size, and relevant patient demographics.
* **Data Cleaning and Preprocessing:** Outline the steps involved in cleaning and preparing the data for analysis. This may include handling missing values, outliers, inconsistencies, and data normalisation.
* **Feature Engineering:** Discuss the creation of new features or transformations of existing features to enhance the predictive power of the model.

**Data Exploration and Visualization**

* **Exploratory Data Analysis (EDA):** Summarise the key characteristics of the dataset using descriptive statistics and visualisations. This may include histograms, box plots, correlation matrices, and other relevant plots.
* **Data Visualization:** Create informative visualisations to understand the relationships between variables, identify patterns, and uncover potential insights.

**Model Development and Selection**

* **Model Selection:** Discuss the choice of appropriate machine learning algorithms based on the problem type (e.g., classification, regression, clustering). Consider algorithms such as logistic regression, decision trees, random forests, support vector machines, or neural networks.
* **Model Training and Evaluation:** Describe the process of training and evaluating the selected models using appropriate metrics (e.g., accuracy, precision, recall, F1-score, AUC-ROC).
* **Model Optimization:** Explore techniques to improve model performance, such as hyperparameter tuning, feature selection, and ensemble methods.

**Predictive Modeling and Insights**

* **Model Deployment:** Discuss the deployment of the best-performing model into a production environment for real-time or batch predictions.
* **Model Interpretation:** Explain how to interpret the model's predictions and extract meaningful insights from the results.
* **Clinical Validation:** Outline the plan for validating the model's performance in a clinical setting to assess its practical utility.

**Technology Stack**

* **Data Science Tools and Libraries:** Specify the software and programming languages to be used for data analysis and modeling (e.g., Python, R, SQL, Pandas, NumPy, Scikit-learn, TensorFlow, PyTorch).
* **Cloud Platforms:** If applicable, mention the cloud platforms to be utilized for data storage, processing, and model deployment (e.g., AWS, GCP, Azure).

**Ethical Considerations**

* **Data Privacy and Security:** Address the importance of protecting patient data and ensuring compliance with relevant regulations (e.g., HIPAA, GDPR).
* **Bias and Fairness:** Discuss the potential for bias in the data and model, and strategies to mitigate it.

***To summarise the key findings and potential impact of the project is to highlight the value of data science in understanding urine biomarker data and its potential applications in clinical practice.***

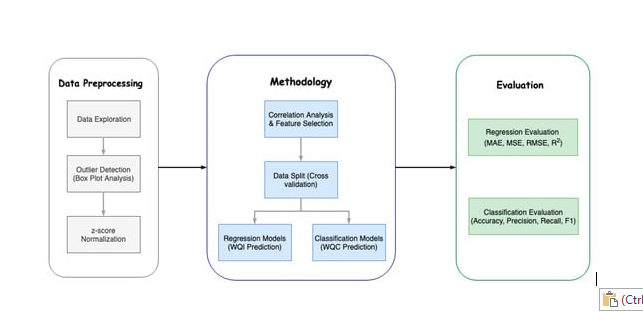
# 

# 

# 

# TECHNICAL APPROACH:

To build a model using data science and ML using google co labs as a platform to perform the analysis.



## INSTALLING AND IMPORTING OF NECESSARY LIBRARIES FOR ACTION ON DATA

***Installed pycaret for automated ML analysis***



PyCaret is a Python open source machine learning library designed to make performing standard tasks in a machine learning project easy.

It is a Python version of the Caret machine learning package in R, popular because it allows models to be evaluated, compared, and tuned on a given dataset with just a few lines of code.

The PyCaret library provides these features, allowing the machine learning practitioner in Python to spot check a suite of standard machine learning algorithms on a classification or regression dataset with a single function call.

***Installed scikit-learn for automated ML analysis***



Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

***Import other libraries to help with data presentation and analysis***



**Pandas:** Pandas is chiefly used for machine learning in the form of DataFrames. Pandas allows for importing and exporting tabular data in various formats, such as CSV or JSON files. (In this project we use .csv format data)

**Seaborn:**Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures. Here is some of the functionality that seaborn offers: A dataset-oriented API for examining relationships between multiple variables.

**Missingno:**Missingno is an excellent and simple to use Python library that provides a series of visualisations to understand the presence and distribution of missing data within a pandas dataframe. This can be in the form of either a barplot, matrix plot, heatmap, or a dendrogram.

**Numpy:** Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Moreover Numpy forms the foundation of the Machine Learning stack.

**Matplotlib.pyplot:**It is a very comprehensive library and designed in such a way that most of the functions for plotting in MATLAB can be used in Python. It consists of several plots like the Line Plot, Bar Plot, Scatter Plot, Histogram e.t.c through which we can visualise various types of data.

**Plotly.express:**Plotly express is a high-level data visualisation package that allows you to create interactive plots with very little code. It is built on top of Plotly Graph Objects, which provides a lower-level interface for developing custom visualisations.

## DATA PROCESSING:

### Data Exploration:

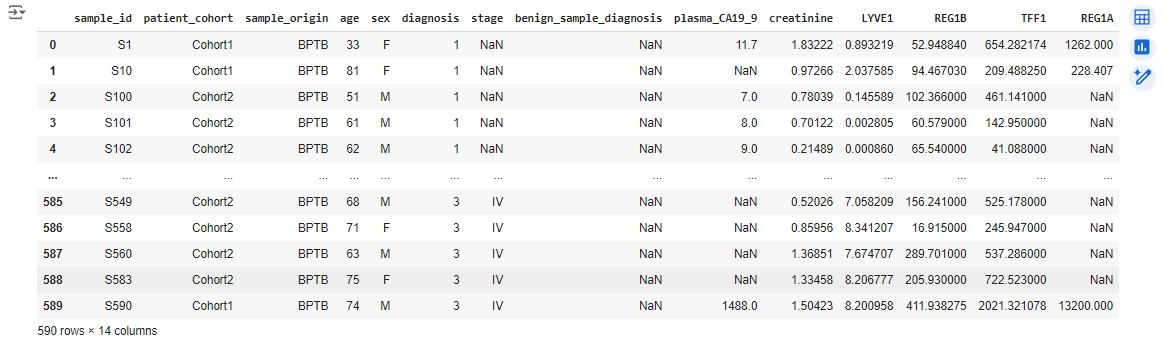
* Explored various data sets containing data for
  + Breast cancer
  + Lung cancer
  + Pancreatic cancer
  + Etc..

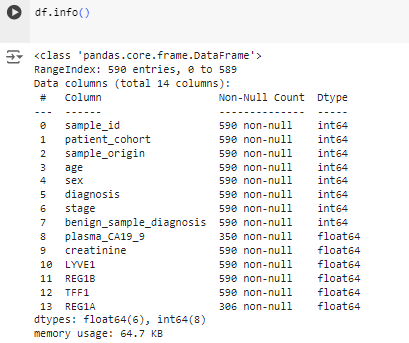
Picked data pertaining to Pancreatic cancer as it had biomarker data collected by non-invasive approach

* Conversion of data into CSV to work on colab
* ***Import the CSV file into a co lab notebook.***
* ***Create a DataFrame -* using pandas library**

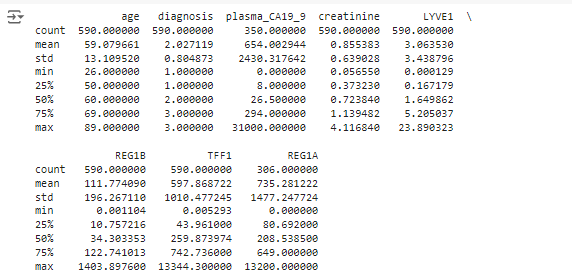


* ***View the DataFrame Created***



****

### 



Can see quite a outliars like in plasma\_ca19\_9 we see max value as 31000 with 75% is around 294 \**Data set may need CLEANING for outlier removals \**

### 

### 

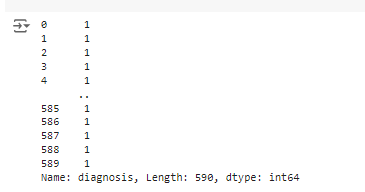
### 

### Data Separation as X and Y

It’s common to consider features of a dataset as the input to a model, and labels of the same dataset as the model’s output. This approach has, however, two important problems that limit its capacity for generalisation.

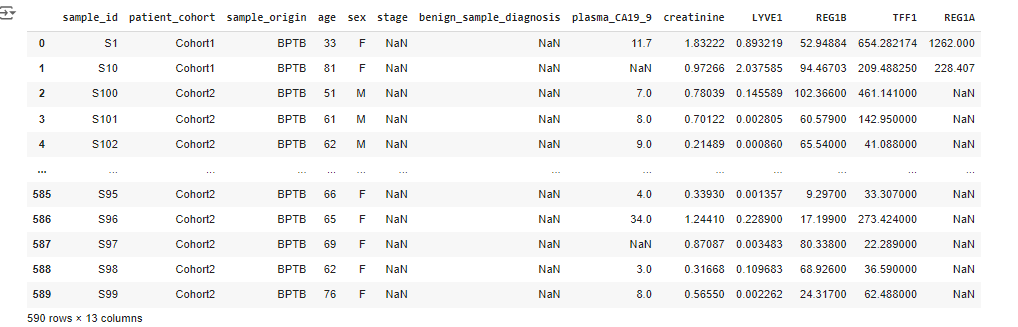
A label in this case is diagnosis to which we assign arbitrarily high importance as our final outcome is to predict the diagnosis based on the bio-markers. Hence we create the label as Y.

****

****

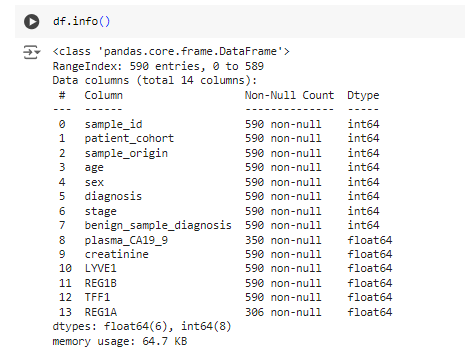
features are characteristics of observations, In our dataset we initially assume all the columns except diagnosis are features. Hence we create a feature set X.

****

****

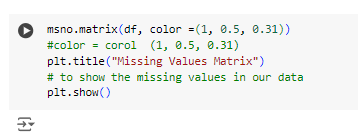
### EDA - Data cleaning

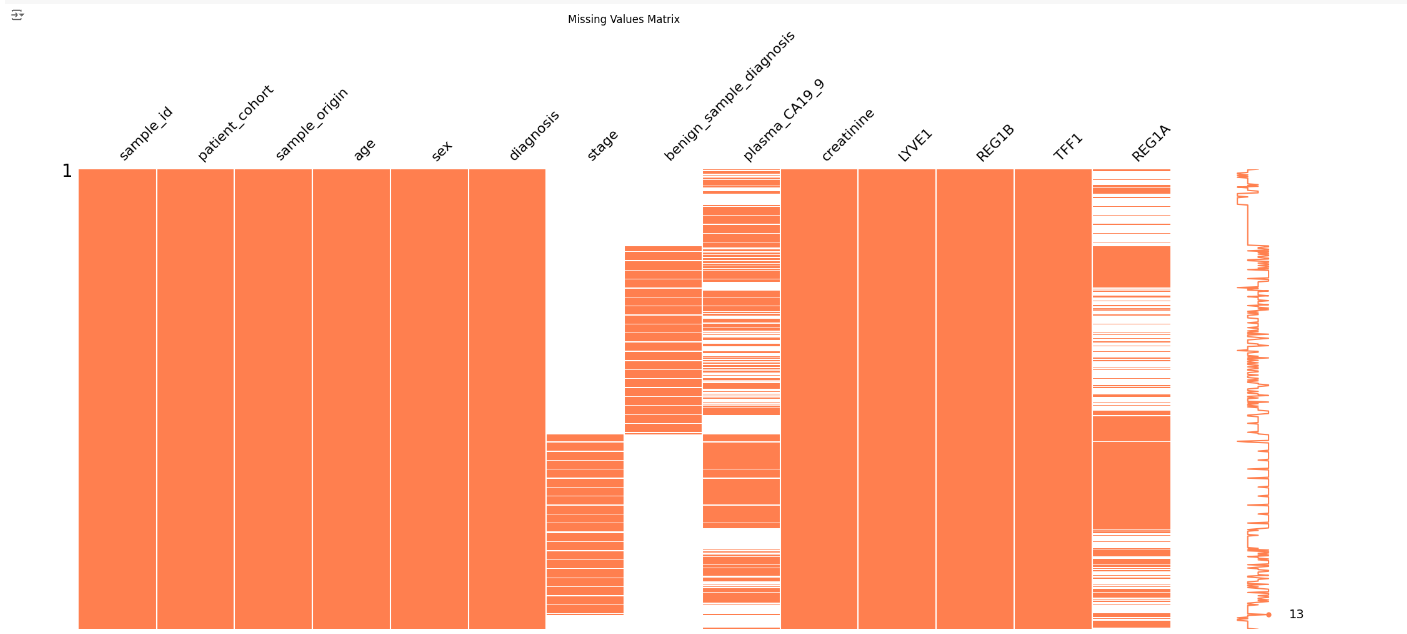
* Identify Empty Numeric data values in the DataFrame

****

* ***We observe the numeric column plasma\_CA19\_9 and REG1A have few Null datavalues***

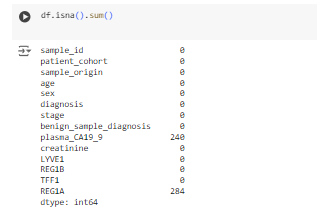
Analysed the data to find empty data in each column ( few columns having missing values or NaN values) for both numeric and non-numeric data





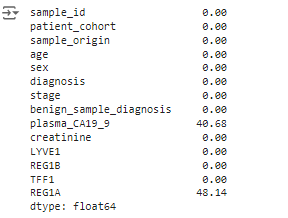
***Data shows the columns Stage, benaing\_sample\_diagonise , plasma\_CA19\_9, REGIA contains empty data values we observe stage and benaing\_sample\_diagonise have null data exclusively.***

* Get count of Empty Numeric data values in the DataFrame



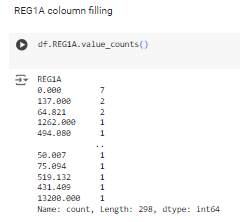
* Calculate the ratio value of the missing data values in the 2 features.

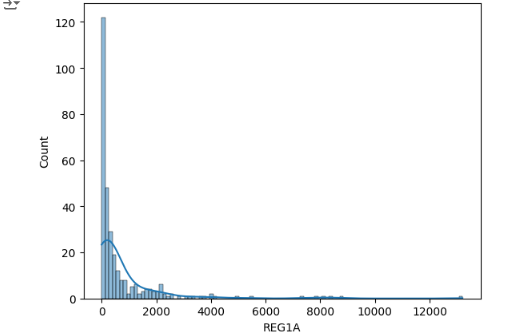
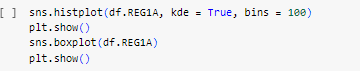


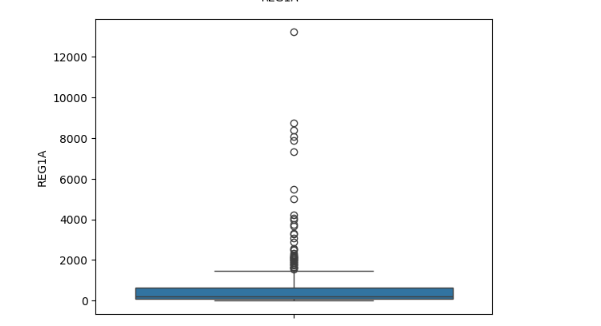


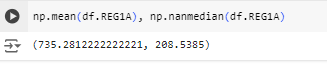
Since we have empty values for important biomarkers we use median to fill up the values of REG1A and plasma\_CA19\_9 columns.

**Study on REG1A feature:**



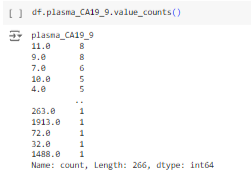


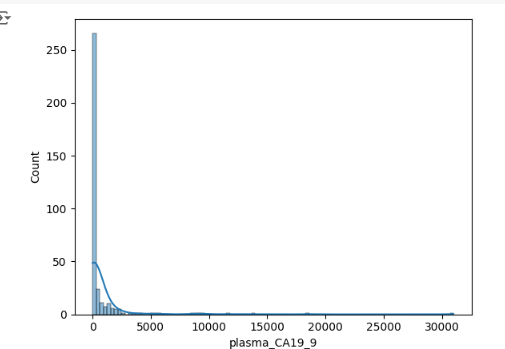
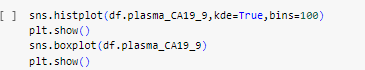


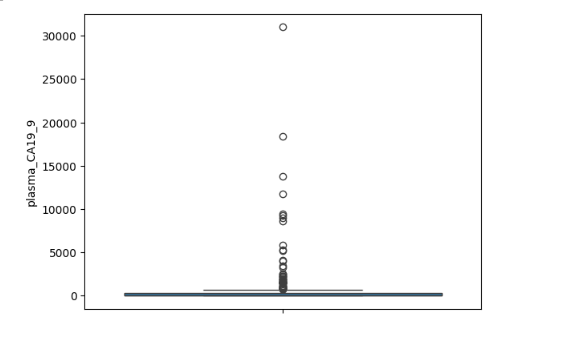
******

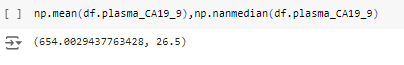
***here we can see there are outliers in the dataframe for REG1A column and median is a suitable option to fill in the data values when outliers are present in data***

**Study on PLASMA\_CA19\_9 feature:**









***here we can see there are outliers in the dataframe for REG1A column and median is a suitable option to fill in the data values when outliers are present in data***

Here in both the case’s we use median to fill up empty values of a column . We use Median

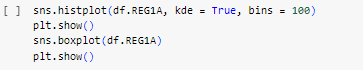
* As Median is a position-based measure.
* Median-Imputation is a robust technique to handle missing data with outliers.

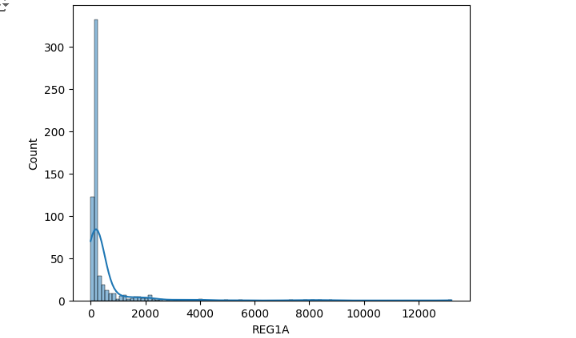
***So, I would like to conclude that Mean-Imputation is generally not a recommended practice when your data has extreme values, due to its misleading nature.Hence selected Median as fill up option.***

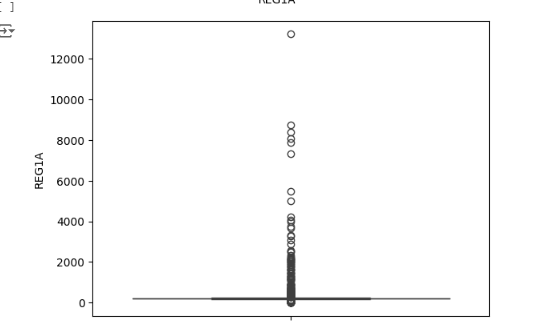
### Data Populating:

**Populating REG1A and PLASMA\_CA19\_9 feature:**



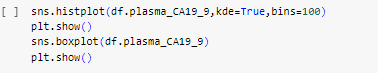


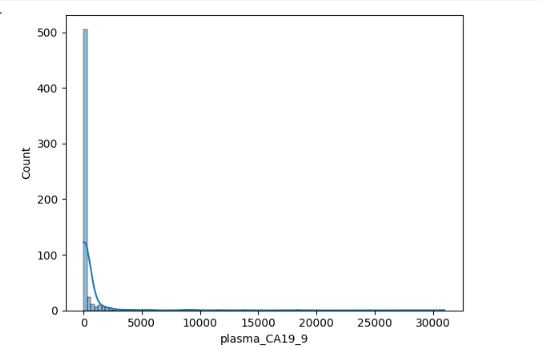


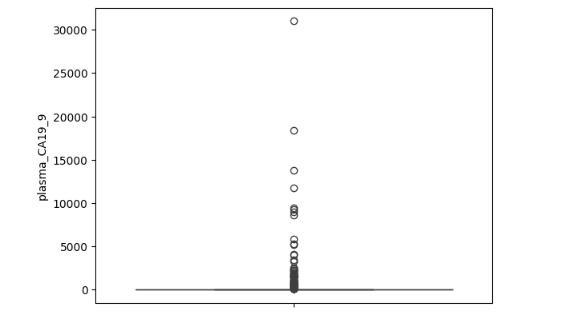
******

***With majority of values falling between 0 - 2100.***



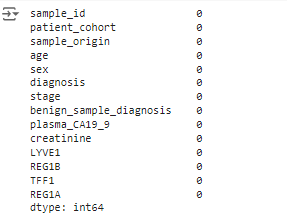






***Which shows majority of the values falling in 0-3000 rang***





Empty data fields are now populated.

### 

### 

### 

### 

### 

### 

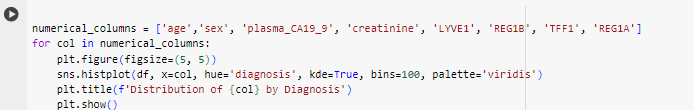
### 

### 

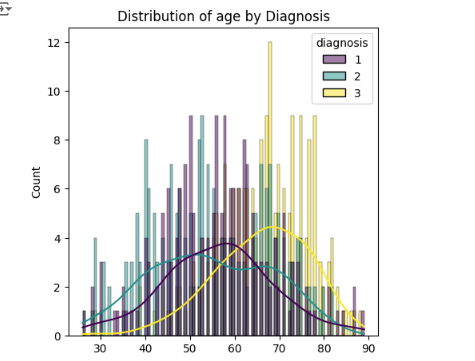
### 

## ANALYSIS

### Numerical Features Analysis



***AGE wrt DIAGNOSIS:***

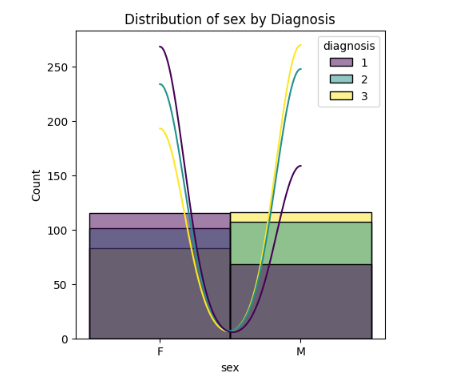
******

***We observe age feature has a strong relationship with diagnosis,***

***Diagnosis who are classified as cancer as 3, below is the analysis based on the above graph interpretation.***

* ***Diagnosis begins usually around 63 years***
* ***Diagnosis is usually peak for patient in early 70’s***
* ***And then drops gradually drops latter***

***SEX wrt DIAGNOSIS:***

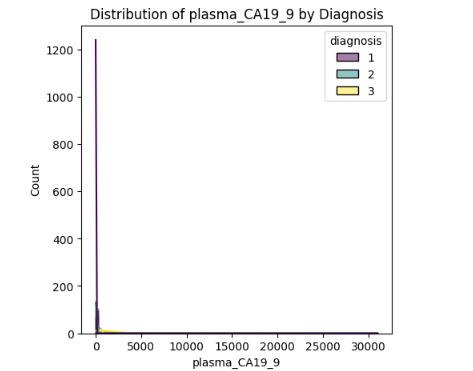
******

***We observe sex feature has a strong relationship with diagnosis,***

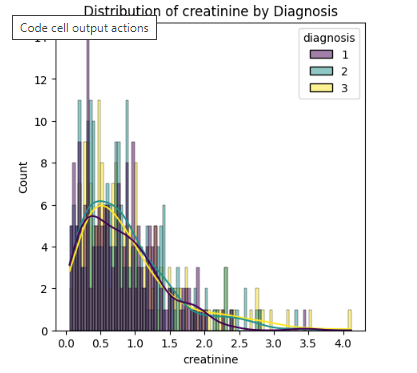
***Diagnosis who are classified as cancer as 3, below is the analysis based on the above graph interpretation.***

* ***Men are at higher risk of pancreatic cancer when compared with women***

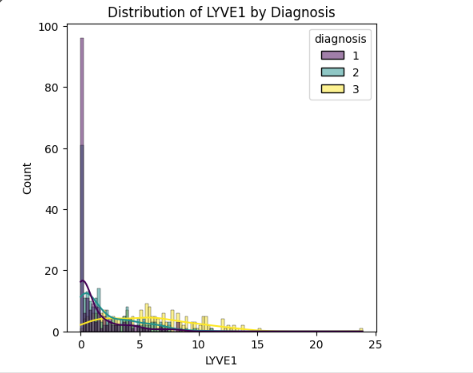
***PLASMA\_CA19\_9 wrt DIAGNOSIS:***



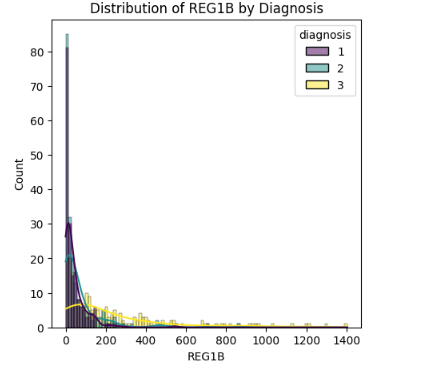
***CRETININE wrt DIAGNOSIS:***

******

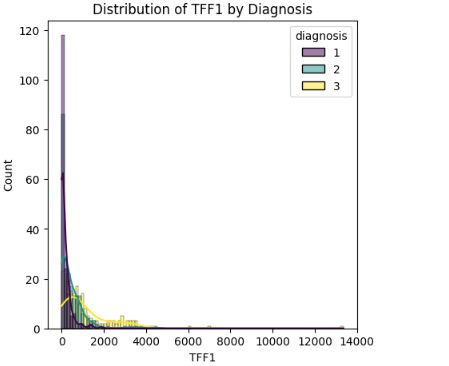
***LYVE1 wrt DIAGNOSIS:***



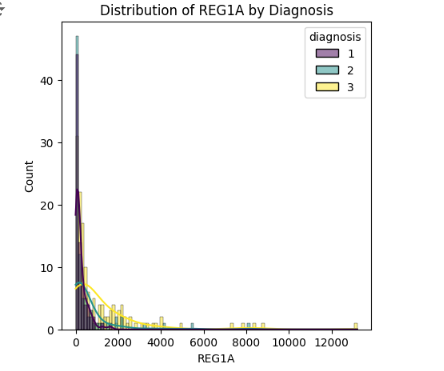
***REG1B wrt DIAGNOSIS:***

******

***TFF1 wrt DIAGNOSIS:***



***REG1A wrt DIAGNOSIS:***



**HISTOGRAM PLOT ANALYSIS:**

***Diagnosis Count by Age: most diagnoses are in the 60-70 age group, with fewer diagnoses in younger and older age groups***

***Diagnosis Count by Sex: By comparing the heights of the bars, you can see how the count of diagnoses differs between males and females. For instance, a taller bar for "Male" would indicate that there were more diagnoses in the male group compared to females in this dataset.***

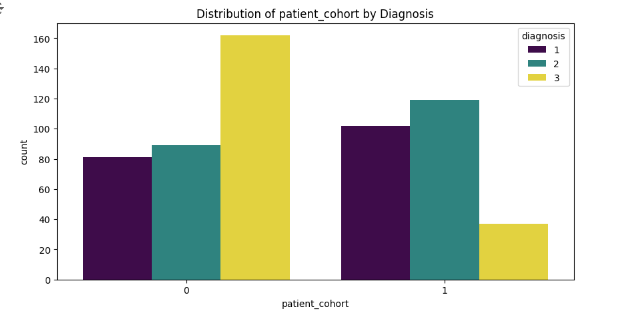
***Diagnosis Count vs creatinine : The graph suggests a positive relation between creatinine levels and the number of diagnoses a patient has.***

***Diagnosis Count by other attributes/features: There is no clear linear correlation with the count of diagnoses. The data points are scattered throughout the chart, and there's no distinct upward or downward trend. This suggests that there might not be a strong relationship between the other feature values and the number of diagnoses.***

### Categorical Features Analysis

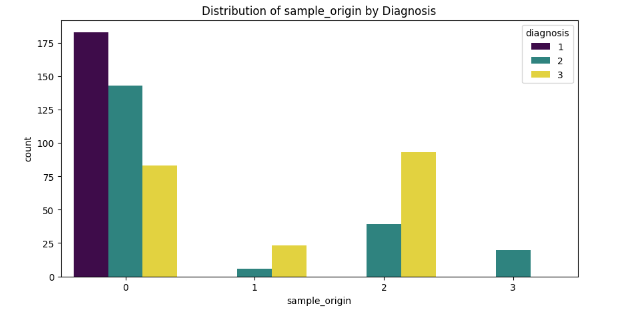
****

***PATIENT COHORT wrt DIAGNOSIS:***

****

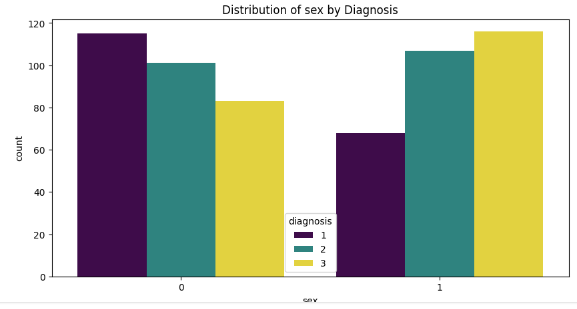
Patient Cohort of category 0 are more inclined to diagnosis of pancreatic cancer than in category 1

***SAMPLE ORIGIN wrt DIAGNOSIS:***

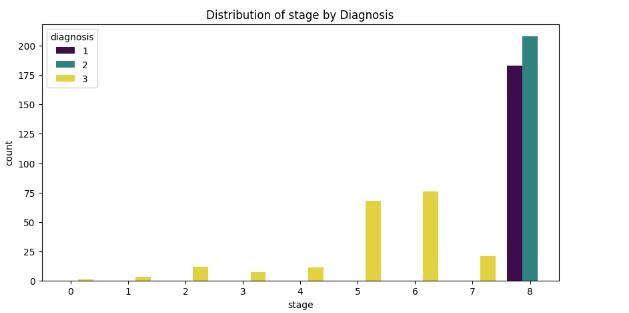


Sample origin of Location 2 and than 0 are more inclined to diagnosis of pancreatic cancer than in category 1 and 3.

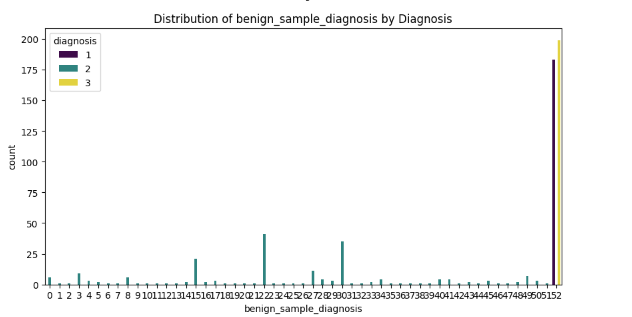
***SAMPLE ORIGIN wrt DIAGNOSIS:***



***STAGE wrt DIAGNOSIS:***



***BENIGN\_SAMPLE wrt DIAGNOSIS:***



**HISTOGRAM PLOT ANALYSIS:**

***Diagnosis Count by Age: most diagnoses are in the 60-70 age group, with fewer diagnoses in younger and older age groups***

***Diagnosis Count by Sex: By comparing the heights of the bars, you can see how the count of diagnoses differs between males and females. For instance, a taller bar for "Male" would indicate that there were more diagnoses in the male group compared to females in this dataset.***

***Diagnosis by patient cohort: The graph suggests a positive relation patient cohort and number of diagnoses a patient has.***

***Diagnosis by Sample Origin : This also suggest strong relationship***

## CORRELATION

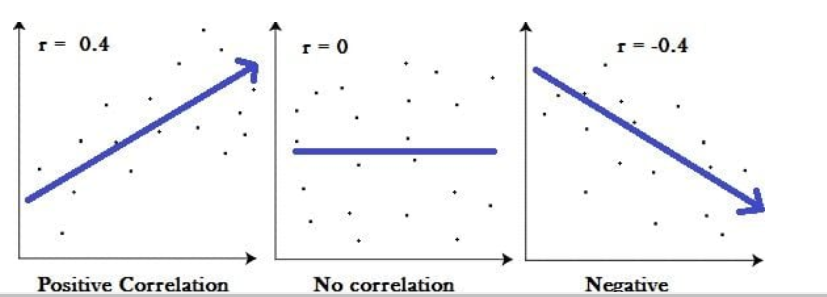
### Correlation Matrix to find the relationship between features:

A correlation matrix involves a rows and columns table that shows the variables. Every cell in a matrix contains the correlation coefficient.

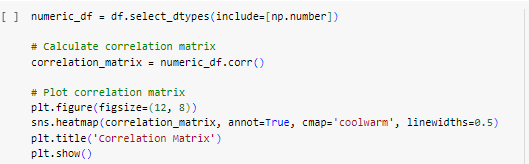
A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where

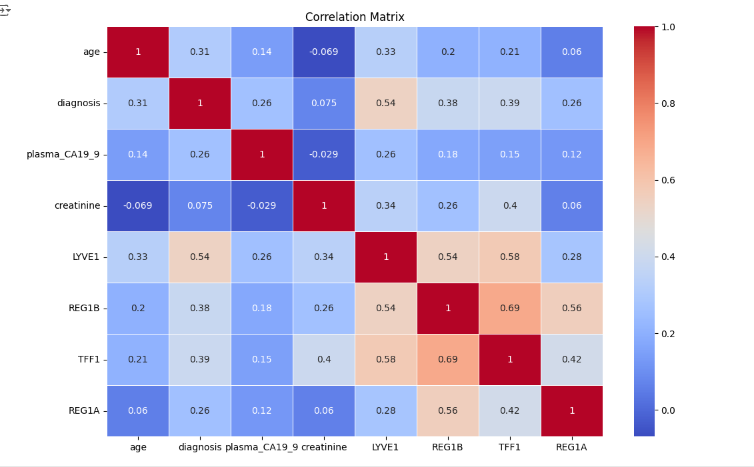
* 1 is considered a strong relationship between variables
* 0 a neutral relationship and
* -1 a not strong relationship or negative relationship.

It’s most commonly used in building regression models.



**Correlation between NUMERIC FEATURES:**





Here are some of the strongest correlations in the matrix:

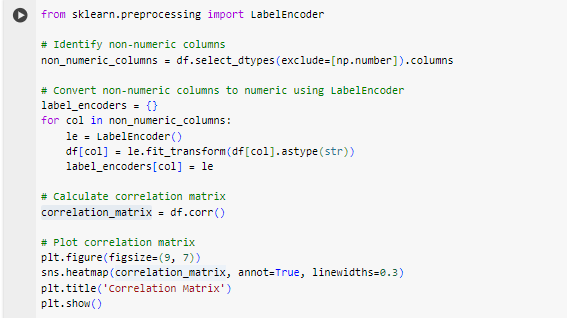
* LYVEL and REGIB (0.54)
* LYVEL and TFFI (0.58)
* TFFI and REGIA (0.42)
* Diagnosis and LYVEL (0.33)

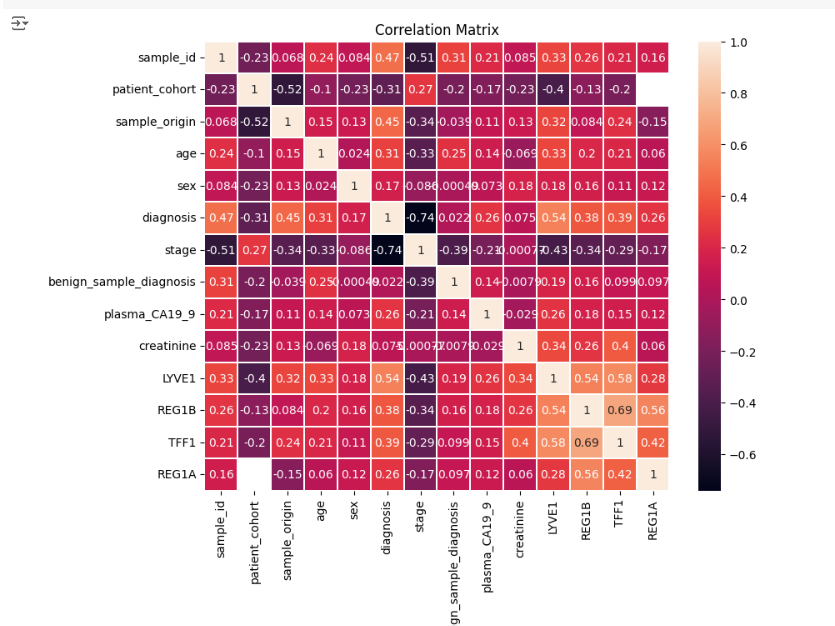
A favourable relationship is seen between REG1B and TFF1 bio-marker, but not strong enough to analyse the redundancy.

As next step, we will try to include non-numeric features as well to get a view of relationships between all the feature.

**Correlation among ALL the FEATURES:**

As 1st step we will convert each non-numeric attribute to an enumeration numeric value

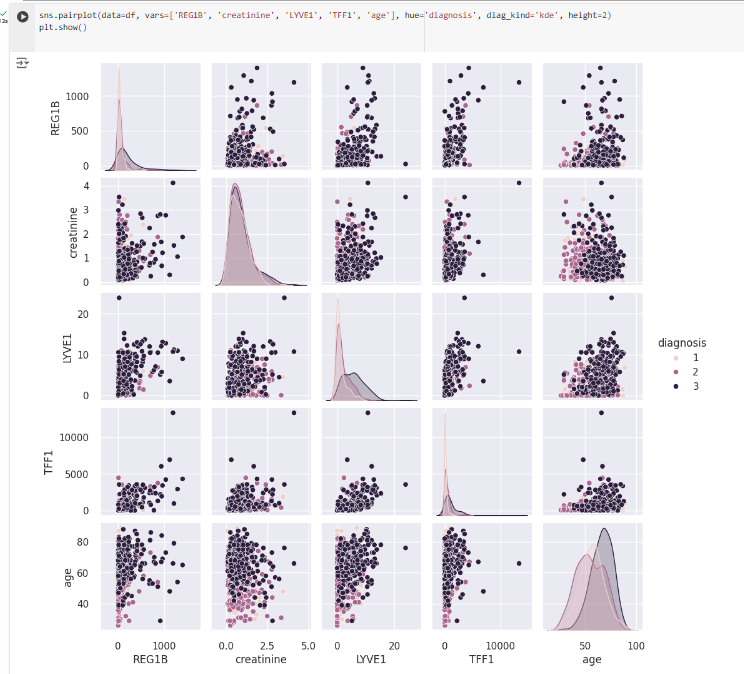


****

Here are some of the strongest correlations in the matrix:

* LYVEL and REGIB (0.54)
* LYVEL and TFFI (0.58)
* TFFI and REGIA (0.42)
* Diagnosis and LYVEL (0.33)

### PAIR PLOTTING between biomarker features:



***Do not see significant pairing and correlation.***

***Hence redundancy is not a case here and we consider all the features for model building.***

## Model matrices Details:

**1. Accuracy:**

* **Definition:** The proportion of correct predictions made by the model over the total number of predictions. It's calculated as the number of true positives (TP) and true negatives (TN) divided by the total number of instances.
* **Formula:** Accuracy = (TP + TN) / Total
* **Interpretation:** A high accuracy indicates the model is generally performing well at classifying instances correctly. However, it can be misleading if the class distribution is imbalanced (e.g., many more negative than positive examples).

**2. Area Under the ROC Curve (AUC):**

* **Definition:** The AUC is a probability metric that measures the model's ability to distinguish between positive and negative classes. It represents the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds.
* **Interpretation:** An AUC of 1 indicates perfect discrimination between classes, while an AUC of 0.5 is equivalent to random guessing. Higher AUC values generally indicate better performance.

**3. Recall (Sensitivity, True Positive Rate):**

* **Definition:** The proportion of actual positive instances that were correctly identified by the model. It's calculated as the number of true positives (TP) divided by the total number of positive instances (TP + FN).
* **Formula:** Recall = TP / (TP + FN)
* **Interpretation:** A high recall indicates the model is good at identifying most of the positive cases. It's important for tasks where missing positive cases is costly (e.g., medical diagnosis).

**4. Precision (Positive Predictive Value):**

* **Definition:** The proportion of predicted positive instances that were actually correct. It's calculated as the number of true positives (TP) divided by the total number of instances predicted positive (TP + FP).
* **Formula:** Precision = TP / (TP + FP)
* **Interpretation:** A high precision indicates the model is good at avoiding false positives. It's important for tasks where wrongly classifying negative instances is costly (e.g., spam filtering).

**5. F1-Score:**

* **Definition:** The F1-score is a harmonic mean of precision and recall, combining both metrics into a single score. It's a good way to balance the importance of precision and recall when both are important for your task.
* **Formula:** F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)
* **Interpretation:** F1-score considers both false positives and false negatives, providing a balanced view of the model's performance. A high F1-score indicates the model is good at both identifying true positives and avoiding false positives.

**6. Kappa Statistic (Cohen's Kappa):**

* **Definition:** The kappa statistic is a metric that considers the agreement between the model's predictions and the actual labels, accounting for chance agreement. It's useful for evaluating classification performance when the class distribution is imbalanced.
* **Formula:** Kappa = (Observed Agreement - Expected Agreement) / (1 - Expected Agreement)
* **Interpretation:** Kappa values range from -1 (perfect disagreement) to 1 (perfect agreement). However, interpreting the specific value depends on the context and number of classes in your problem.

**7. Matthews Correlation Coefficient (MCC):**

* **Definition:** The MCC is another balanced metric that takes into account true positives, true negatives, false positives, and false negatives. It's particularly useful for imbalanced class problems.
* **Formula:** MCC = (TP \* TN - FP \* FN) / sqrt((TP + FP) \* (TP + FN) \* (TN + FP) \* (TN + FN))
* **Interpretation:** MCC values range from -1 (perfect disagreement) to +1 (perfect agreement). Similar to kappa, the specific value interpretation depends on the context of your problem.

## MULTICLASS MODEL and PREDICTIONS:

**Diagnosis Target:**

0 → Healthy

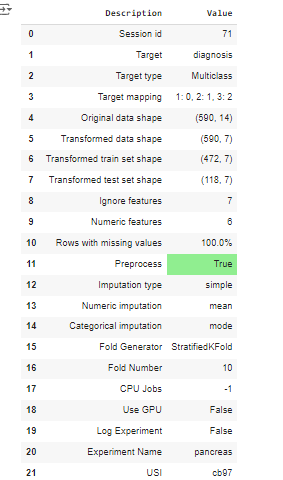
1→ Non Healthy and Non Cancerous

2 → Cancerous

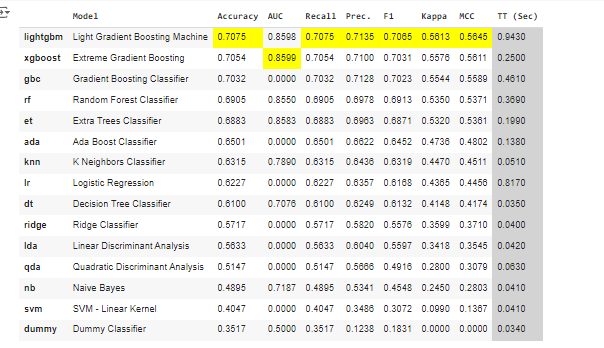
### MODEL BUILDING:

Setting up model to consider only bio-markers





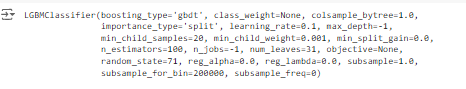


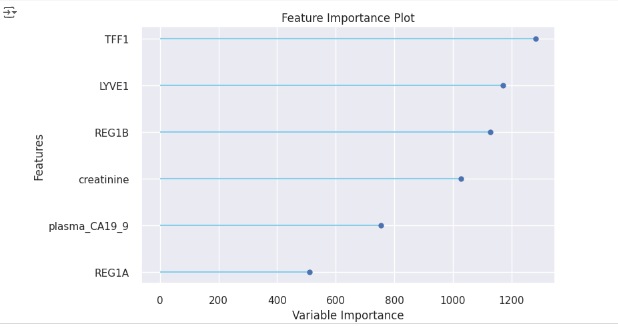


**Here we are using Accuracy as the best matrix.**

Best Model Details:







***We observe the bio-markers are important features***

* ***TFF1***
* ***LYVE1***
* ***REG1B***

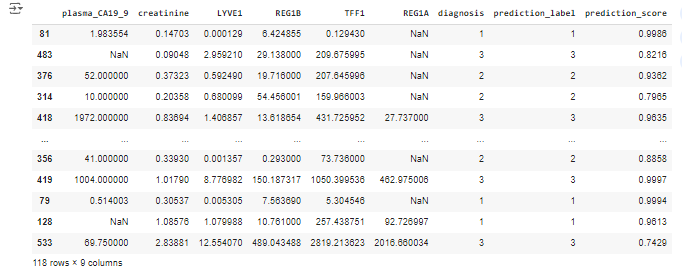
### PREDICTING:

Identifying Best Predicting model based on the best model built:



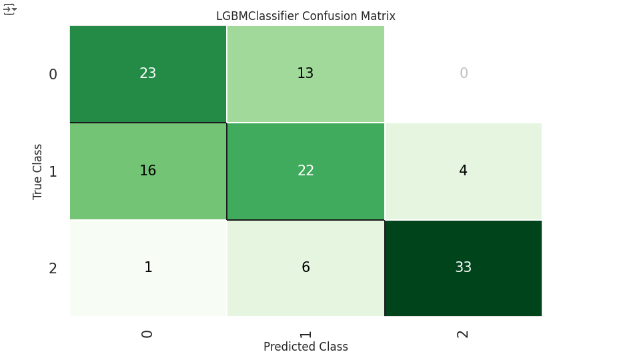






### VALIDATING Predictions:

****

****

There seems to be a good performance based on the high values on the diagonal (23, 22, and 33). These represent correctly classified instances.

There are still some misclassifications indicated by the off-diagonal values.

***GENERAL ASSESSMENT:***

* ***Yes, it's a good starting point:*** *The high values on the diagonal indicate a good overall performance in correctly classifying most instances.*
* ***There's room for improvement:*** *The off-diagonal values suggest some misclassifications are still occurring.*
* ***Further analysis can be looked.***

## BINARY MODEL and PREDICTIONS:

**Diagnosis Target:**

0 → Healthy

1 → Cancerous

### DATA CHANGES to Create a Binary Model:

Remove the non-healthy and non-cancerous



We are removing the data of diagnosis as 2 which are not healthy and not cancerous ..thus comparing only between healthy and cancerous data

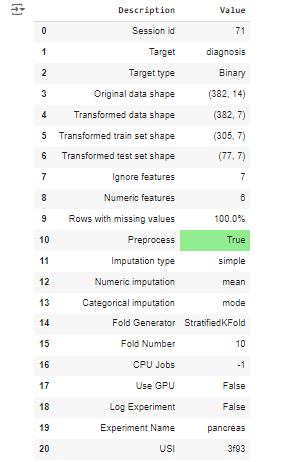




### MODEL BUILDING:

Setting up model to consider only bio-markers

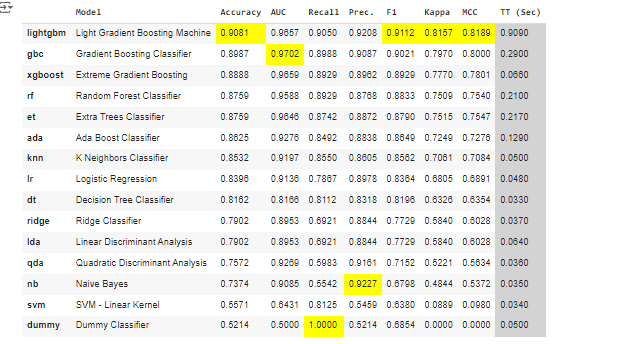




**Here we are using Accuracy as the best matrix.**

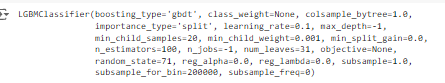




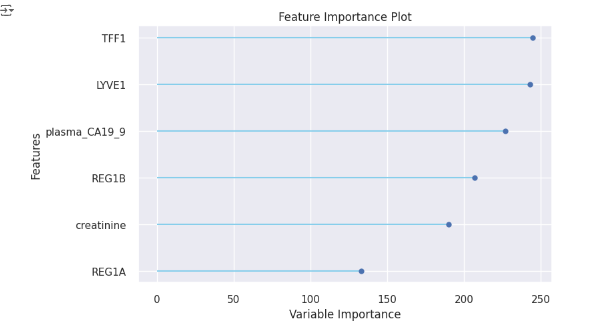


Best Model Details:









***We observe the bio-markers are important features***

* ***TFF1***
* ***LYVE1***
* ***plasma\_CA19\_9***

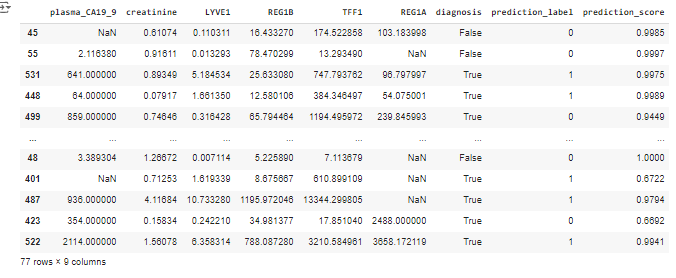
### PREDICTING:

Identifying Best Predicting model based on the best model built:



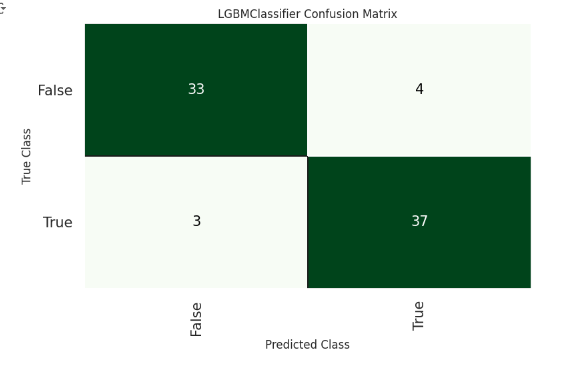






### VALIDATING Predictions:

****

****

There seems to be a good performance based on the high values on the diagonal (23, 22, and 33). These represent correctly classified instances.

There are still some misclassifications indicated by the off-diagonal values.

***GENERAL ASSESSMENT:***

* ***Based on the confusion matrix, the model shows promising signs of good performance.***

# CONCLUSIONS AND FURTHER SCOPE

## TECHNICAL CONCLUSION:

Our two models in the last section achieve a respectable overall classification score but still have some false negatives.

Prediction seems good for the Binary Model but the other matrices are weaker than the multiclass model.

## FURTHER SCOPE on Technical Front:

The best way forward could be to develop the multiclass approach further with feature engineering and a more sophisticated strategy to replace the missing values.

A bigger data size will also be a good starting point for further investigation.

## PROJECT CONCLUSION:

The comprehensive technical review and subsequent data science study have demonstrated promising potential for the early diagnosis of pancreatic cancer through the analysis of urine biomarker data. The findings underscore the need for further in-depth research and analysis to solidify these initial observations.

Expanding upon this research will be instrumental in the development of innovative pancreatic cancer diagnostic kits and strips. Moreover, conducting thorough market analysis will provide valuable insights into market size and potential opportunities.

Ultimately, the successful realisation of this project holds the promise of significantly improving early detection rates for pancreatic cancer, ultimately saving countless lives. By enabling timely intervention, we can make a substantial impact on the lives of patients and their families.

This research represents a critical step forward in the fight against pancreatic cancer.

**Above all, This will be a great help for early diagnosis and save human life ..the precious ONE!**

# REFERENCES

<https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf>

<https://gco.iarc.who.int/media/globocan/factsheets/cancers/13-pancreas-fact-sheet.pdf>

<https://raw.githubusercontent.com/tararelan/Kaggle-Pancreatic-Cancer/main/Debernardi%20et%20al%202020%20data.csv>

<https://pubmed.ncbi.nlm.nih.gov/33301466/>

<https://pubmed.ncbi.nlm.nih.gov/26240291/>

<https://www.grandviewresearch.com/industry-analysis/urinalysis-market>

<https://www.wcrf.org/cancer-trends/pancreatic-cancer-statistics/>

<https://www.gminsights.com/industry-analysis/self-testing-market>