

Project Report

On

**URINE BIOMARKERS ANALYSIS FOR PANCREATIC CANCER
DIAGNOSIS**

Submitted

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

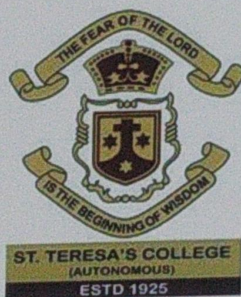
AZLAMIYA T A

(Register No. SM22AS007)

(2022-2024)

Under the Supervision of

Mrs. KAVYA KISHORE



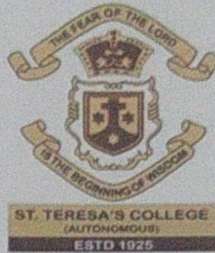
DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

APRIL 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

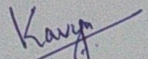


CERTIFICATE

This is to certify that the dissertation entitled, **URINE BIOMARKERS ANALYSIS FOR PANCREATIC CANCER DIAGNOSIS** is a bonafide record of the work done by **Ms. AZLAMIYA T A** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 29-04-2024

Place: Ernakulam

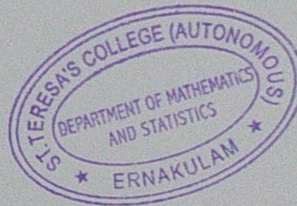

Mrs. Kavya Kishore

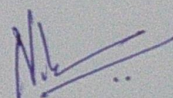
Assistant Professor,

Department of BCA (CT & ISM),

St. Teresa's College (Autonomous),

Ernakulam




Mrs. Nisha Oommen

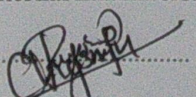
Assistant Professor & HOD,

Department of Mathematics and Statistics,

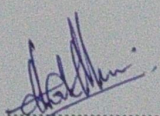
St. Teresa's College (Autonomous),

Ernakulam

External Examiners:

1: 

CHINU JOSEPH
29/4/2024.

2: 

LAKSHMI SURESH
29/04/2024

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **Mrs. Kavya Kishore**, Assistant Professor, Department of BCA (CT & ISM), St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Place: Ernakulam

AZLAMIYAT A

Date: 29-04-2024

SM22AS007

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout this study helped me to complete the work up to this stage

I am very grateful to my project guide Mrs. Kavya Kishore for the immense help during the period of work

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to the faculty, teaching, and non-teaching staff of the department and Colleagues.

I am also very thankful to HoD Mrs. Nisha Oommen for their valuable suggestions and critical examination of work during the process.

Place: Ernakulam

AZLAMİYAT A

Date: 29-04-2024

SM22AS007

ABSTRACT


Pancreatic cancer is one of the deadliest types of cancer with a high mortality rate it remains undetected until it reaches an advanced stage so that the early detection of this type of cancer is very crucial. In this study with the Kaggle Urinary Biomarkers for Pancreatic Cancer (2020) dataset for early diagnosis of pancreatic cancer, classification of samples into non-cancerous and cancerous cases based on the levels of urine biomarkers, plasma CA19-9, age and sex. Four machine learning models Support Vector Machine (SVM), Random Forest, K Nearest Neighbor (KNN), and Logistic Regression are built. From these models the best model is used for plotting features built upon the importance. The feature that is critical for the classification of classes is detected using model evaluation techniques. For model evaluation confusion matrix and Receiver operating curve (ROC) curve are used.

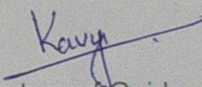



ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM

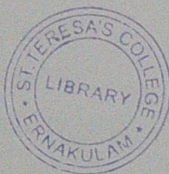
Certificate of Plagiarism Check for Dissertation

Author Name	AZLAMIYA T A
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. KAVYA KISHORE
Department	Post Graduate Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	URINE BIOMARKERS ANALYSIS FOR PANCREATIC CANCER DIAGNOSIS
Similarity	0% AI 12%
Paper ID	1669844
Submission Date	2024-04-20 13:01:31


Signature of Student


Signature of Guide


Checked By
College Librarian



* This report has been generated by DrillBit Anti-Plagiarism Software

Contents

Page no.

1. Introduction	8
1.1 Pancreas	8
1.2 Pancreatic Cancer	8
1.2.1 Pancreatic Neuroendocrine Tumor	9
1.2.2 Pancreatic Ductile Adenocarcinoma	9
1.3 Urine Biomarkers	9
1.4 Objectives of the study	10
2. Literature Review	11
3. Materials and Methodology	16
3.1 Dataset	16
3.2 Machine Learning	16
3.2.1 Supervised Learning	17
3.2.2 Unsupervised Learning	17
3.2.3 Semi-Supervised Learning	18
3.2.4 Reinforcement Learning	18
3.3 Exploratory Data Analysis (EDA)	19
3.4 Classification Models	19
3.4.1 Support Vector Machine (SVM)	20
3.4.2 Random Forest Classifier	20
3.4.3 Logistic Regression	21
3.4.4 K Nearest Neighbor (KNN)	21
3.5 Model Evaluation	22
3.5.1 Confusion Matrix	22
3.5.2 ROC-AUC Curve	23
3.6 Feature Importance	24

3.7 Methodology of the study	25
4. Data Description And Exploratory Data Analysis	26
4.1 Attributes	26
4.2 Sample Dataset	27
4.3 Data Pre-processing	28
4.4 Exploratory Data Analysis (EDA)	29
5. Results And Discussions	33
5.1 Model Building	33
5.1.1 Support Vector Machine (SVM)	33
5.1.2 Random Forest	34
5.1.3 Logistic Regression	35
5.1.4 K Nearest Neighbor (KNN)	36
5.2 Comparison of Models	37
5.3 Feature Importance Plot	37
6. Conclusion	39
7. References	40

Chapter 1

Introduction

1.1 Pancreas

The pancreas is a major organ in the digestive system that lies in the human abdomen behind the stomach. The pancreatic gland acts as a digestive gland in the digestive system. The pancreas serves as both an endocrine gland and an exocrine gland. Functioning as an endocrine gland it secretes insulin which helps in the movement of glucose to muscles and other tissues for use as energy from the blood if the blood sugar level is high. When the blood sugar level is low it secretes, glucagon helps break down stored sugar into glucose by the liver to maintain the balance in sugar level. Thus, the pancreas as an endocrine part regulates the blood sugar level. Now the pancreas as an exocrine gland secretes pancreatic juice that contains bicarbonate into the duodenum through the pancreatic duct. This juice neutralizes the acids entering the duodenum from the stomach and also the digestive enzymes that break down carbohydrates, proteins, and fats in the food entering the duodenum from the stomach.

1.2 Pancreatic Cancer

Pancreatic cancer arises due to the uncontrolled multiplication of cells that develop into tumor cells forming a mass. These tumor cells can spread into other parts of the body. It can be mostly seen as adenocarcinoma in the head of the pancreas. This type of cancer presents at a later stage because the symptoms do not occur in the early stages and have only limited treatment options. Individuals having age younger than 40 are less prone to and individuals around 70 years of age are highly prone to pancreatic cancer. The risk factors include smoking, obesity, older age, diabetes, chronic pancreatitis, and some genetic conditions. There will be no indication of the cancer or any symptoms portrayed till it spreads to other parts of the body and reaches an advanced stage. This shows the

forms of symptoms such as yellow skin, unclear weight loss, dark urine, abdominal or back pain, light-colored stool, and loss of appetite at the higher stage. Using techniques such as computed tomography, ultrasound, examination of tissue samples, and blood tests pancreatic cancer is diagnosed. There are two types of Pancreatic Cancer Pancreatic Neuroendocrine Tumor and Pancreatic Ductile Adenocarcinoma.

1.2.1 Pancreatic Neuroendocrine Tumor

Pancreatic Neuroendocrine Tumor also called Pancreatic Endocrine Cancer arises in the hormone-producing cells and is less aggressive than the other type. This has different clinical characteristics generated in the hormone-producing tissue of the pancreas also abbreviated as PanNETs which is a small minority. These tumors are produced from the neuroendocrine cells in the form of a group of malignant or benign tumors. There are functioning type PanNET tumors which are responsible for secreting glucagon, gastrin, and insulin, and non-functioning types do not secrete any hormones. This type is detected only after it gets spread to other parts of the body. This is also sometimes called islet cell cancer and has the symptoms of jaundice, pain in the upper abdomen, and unexplained weight loss.

1.2.2 Pancreatic Ductile Adenocarcinoma

Pancreatic Ductile Adenocarcinoma arises in the duct that lines the pancreas which makes digestive enzymes. This is the deadliest type of cancer which remains undetected until it reaches the advanced stage. There will be symptoms such as yellow skin, back or abdominal pain, unexpected weight loss, dark urine, and loss of appetite which can be seen only after it reaches a severe condition. Most of the pancreatic cancer cases are the pancreatic ductile adenocarcinoma.

Thus, the need for early detection of pancreatic cancer is crucial since it remains undetected till it reaches the advanced stage in both types of cancer cases.

1.3 Urine Biomarkers

Biomarkers also known as molecular markers and signature molecules are characters that are measured in the blood, body fluids, and tissues. They may occur when there is something gone wrong such that they can be symptoms of disorders, diseases, or regular

body functions. Due to this reason, biomarkers are significant in providing crucial information regarding whether the health of a system in the body is normal or abnormal.

In the case of cancer, biomarkers include DNA, proteins, and other substances that are helpful in the early detection of malignancies and also abnormalities in the body. It also indicates how far cancer is from the body and how well a treatment works by predicting the impact of cancer. Because of these specialties, investigations are going on for the replacement of image-based tests for cancer with biomarker tests. Doctors sometimes test for biomarkers in the tumor cells and will assign a risk of recurrence score since biomarkers detect the chance of reoccurrence of cancer.

The urine biomarkers that play a crucial role in the early detection of pancreatic cancer include Urine Creatinine, LYVE1, REG1B, REG1A, Plasma_CA19_9, and TFF1. Urine Creatinine level reflects renal functioning, LYVE1 functions in malignant tumor development, REG1B and REG1A are linked to pancreatic regenerating cells, and TFF1 is a potential predictive biomarker for PDAC.

1.4 Objectives of the study

1. To conduct Exploratory Data Analysis (EDA) on the dataset containing urinary biomarkers and plasma CA19-9 levels to gain insights into the data.
2. To develop a classification model using a Support Vector Machine (SVM), Random Forest, Logistic Regression, and K Nearest Neighbors (KNN) to distinguish between Non-Cancerous and Cancerous cases based on the urinary biomarkers and plasma CA19-9 levels.
3. To compare Support Vector Machine (SVM), Random Forest, Logistic Regression, and K Nearest Neighbors (KNN) models ability to classify Non-Cancerous and Cancerous cases based on the urinary biomarkers and plasma CA19-9 levels.
4. To determine the most important urinary biomarkers and plasma CA19-9 levels in classifying Non-Cancerous and Cancerous cases, derived from the best model's feature ranking analysis.

Chapter 2

Literature Review

Analysis from the previous research paper related to this paper uses various clinical datasets and makes predictions using advanced statistical techniques, AI technologies, machine learning algorithms, data mining techniques, etc.

- Floyd (2007) examined ‘Data Mining Techniques for Prognosis in Pancreatic Cancer and the anticipated survival times of pancreatic cancer patients’. Various machine learning algorithms, such as artificial neural networks (ANN), Bayesian networks, and support vector machines, were employed for analyzing the clinical data to enhance the reliability of survival models and to anticipate patient outcomes with high accuracy. His results showed that, with a statistically significant difference ($p < 0.05$), the data mining approach facilitated a more precise decision based on the survival of the pancreatic cancer patients when compared with the logistic regression model.
- Ge and Wong (2008) analyzed ‘Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles’. For the classification of premalignant pancreatic cancer, mass-spectrometry data and decision tree ensembles were used. Feature selection techniques such as the student t-test, Wilcoxon rank sum test, and genetic algorithms are used to identify the effectiveness of selecting key features. The various decision-tree-based classifier ensembles, namely Random Forest, Multiboost, Logitboost, Adaboost, Stacked Generalization, and Bagging, were employed with the key features selected from each technique. They have concluded that this approach is a promising method that shows ensemble classifiers outperformed the decision trees for the classification of premalignant pancreatic cancer mass spectrometry data.

- Qiu et al. (2014) attempted a study on ‘The Prediction of Pancreatic Cancer Using SVM’. A feature space construction that accurately categorizes the patients can be seen in this study. This identifies unusual data points and predicts the condition of individuals diagnosed with pancreatic cancer using support vector machines (SVM) and multilayer perceptron (MLP) kernels.
- Hsieh et al. (2018) proposed a paper titled ‘Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models’. This study investigates the fact that patients with type 2 diabetes (T2DM) have a higher risk of pancreatic cancer by involving various risk factors for pancreatic cancer in prediction models. Logistic regression and artificial neural network (ANN) models were compared based on F1 score, precision, and recall, along with receiver operating characteristic (ROC) curves. This investigation finds that logistic regression is the best fit for predicting the likelihood of developing pancreatic cancer, with an area of 0.727 under the ROC curve.
- Muhammad et al. (2019) analyzed ‘Pancreatic Cancer Prediction through an Artificial Neural Network’. In this article, the prediction of the likelihood of an individual developing pancreatic cancer is done using an artificial neural network. Identifying high-risk individuals who require more concentrated monitoring and treatment was done using the results from the sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve by incorporating 18 key features.
- Tong et al. (2020) explored the paper ‘Development, Validation, and Comparison of Artificial Neural Network Models and Logistic Regression Models Predicting Survival of Unresectable Pancreatic Cancer’. They have conducted a study on 221 individuals diagnosed with incurable pancreatic cancer, analyzing the 32 clinical parameters. To predict the chance of survival, logistic regression and three ANN models were built to forecast survival rates for 8 months. The results showed that, in terms of accuracy and reliability, the ANN models performed better than the logistic regression.
- Rustam et al. (2021) delve into the study for the categorization of pancreatic cancer using logistic regression and random forest. To compare the effectiveness of each model, various metrics such as accuracy, precision, recall, and F1 score

were considered, and the results show that Random Forest, with a 99.38% accuracy score with 20% training data, is the best model, whereas Logistic Regression has only 96.48% accuracy with 30% training data.

- Chen et al. (2022) studied the ‘Prediction Model for the Detection of Sporadic Pancreatic Cancer (PRO-TECT) in a Population-Based Cohort Using Machine Learning and Further Validation in a Prospective Study using electronic health records (EHR)’. To establish a risk prediction model for pancreatic cancer, a cohort study is conducted utilizing over 500 variables. Researchers built a population model that detects PRO-TECT using EHR to help in the early detection of Sporadic Pancreatic Cancer.
- Keyl et al. (2022) researched ‘Multimodal survival prediction in advanced pancreatic cancer using machine learning’. Clinical data from 203 patients with pancreatic ductal adenocarcinoma (PDAC) was developed with random forest which is suitable for evaluating the clinical and molecular characteristics of the patients, in which unique models for subgroups were created. This study shows the capability of advanced data analysis techniques in identifying high and low-risk groups of patients than the mGPS and AJCC staging system to improve patient outcomes in the field of oncology.
- Lee et al. (2022) investigated ‘A machine learning approach was used to construct a predictive model for pancreatic cancer disease, according to a population-based study from the NHIRD’. This study aimed to develop a prediction model for pancreatic cancer that estimates the risk thereby early detection of the cancer within the curable period can be done. Comparison among logistic regression, ensemble learning, deep neural networks, and voting ensembles for the detection of pancreatic cancer are carried out with ROC curve and confusion matrix. DNN and stacking model in terms of first-factor combination showcased the best specificity and sensitivity measures while in the three-factor combinations, logistic regression exhibited the highest AUC in the external testing set.
- Placido et al. (2022) assess ‘Pancreatic cancer risk from disease trajectories using deep learning’. The goal of this study is to predict the chance of pancreatic cancer intruding into patients after risk assessment over the time interval of 3 to 60 months. A massive dataset of patient records from Mass General Brigham Healthcare System and the Danish National Patient Registry were collected, and

the analysis was done using advanced AI technologies. The machine learning models are trained with a sequence of diseases found in the patient history then they found that for the happening of cancer within 36 months disease trajectory model in the Danish dataset predicted better than the models without temporal information. Also, the temporal models performed better than the models without temporal information even after removing the chance of cancer within 3 months before the diagnosis of the cancer from the training dataset. Thus, this study demonstrated the domination of advanced AI technologies in providing improved cancer risk assessments and effective patient care.

- Acer et al. (2023) scrutinize ‘Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations’. In this study, the researcher's approach to establishing a technique for early detection of pancreatic ductile adenocarcinoma with noninvasive urine biomarkers and carbohydrate antigen 19-9 (CA19-9) can be seen. The performance of seven machine learning models, support vector machine (SVM), k-nearest neighbors (KNN), naive Bayes (NB), random forest (RF), AdaBoost, light gradient boosting machine (LightGBM), and gradient boosting classifier (GBC) are evaluated in determining the classes healthy controls, pancreatic disorders, and patients with PDAC to find the best model. Both binary and multiple classification approaches are carried out in this process and 5-10-fold cross-validation of data is conducted. In the classification of healthy controls and patients with PDAC, CV-10 is the best classification method in which GBC with 92.99% accuracy and AUC = 0.9761. The best method in classifying patients with pancreatic disorders and PDAC is CV-10 with LightGBM having an accuracy of 86.37% and AUC = 0.9348. In the classification of healthy controls, patients with pancreatic disorders, and patients with PDAC, CV-5 is the best method showing GBC has an accuracy of 72.91% and AUC = 0.8733.
- Karar et al. (2023) proposed the paper ‘Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks’. Making use of the 1D-CNNs and LSTM for the classification of individuals into the healthy pancreas, benign hepatobiliary disease, and PDAC the goal of researchers is to diagnose pancreatic cancer. Testing on a public dataset consisting of 590 urine samples this model achieved 97% accuracy and AUC =

98%. A further approach of evaluating the classifier by testing it in a lab setting on urinary biomarkers and this approach can improve diagnostic procedures for pancreatic cancer.

- Ingwersen et al. (2023) conducted research on ‘Machine learning versus logistic regression for the prediction of complications after pancreatoduodenectomy’. The prediction of delayed gastric emptying and pancreatic fistula after pancreatoduodenectomy by machine learning and logistic regression are compared. The evaluation based on the area under the curve of machine learning is 0.74, and logistic regression model is 0.73 for postoperative pancreatic fistula and the area under the curve for both machine learning and logistic regression for delayed gastric emptying is 0.59. Thus, logistic regression is better at predicting complications after pancreatoduodenectomy than the machine learning model.

Chapter 3

Materials and Methodology

3.1 Dataset

This study made use of the open-access website Kaggle for the Kaggle Urinary Biomarkers for Pancreatic Cancer (2020) dataset, which was collected by Debernardi et al. from BPTB: Barts Pancreas Tissue Bank, London, UK; ESP: Spanish National Cancer Research Centre, Madrid, Spain; LIV: Liverpool University, UK and UCL: University College London, UK. The dataset has 590 urine samples they are separated into three classes healthy patients (183 samples), benign cases (208 samples), and PDAC cases (199 samples). The features given in this dataset are patient cohort, sample origin, age, sex, diagnosis, stage, benign samples diagnosis, five urine biomarkers, and plasma CA19-9 among them the key features are age, sex, five urine biomarkers (LYVE1, REG1B, REG1A, TFF1, Creatinine), and Plasma CA19-9.

3.2 Machine Learning

Machine learning, a branch of computer science and artificial intelligence (AI), specializes in modeling human learning processes using data and algorithms, subsequently gaining accuracy. Because of his studies on the game of checkers, Arthur Samuel is credited with coining the term "machine learning." In 1962, Robert Nealey, who considered himself a master of checkers, faced out against IBM's 7094 computers and lost. Thus, in light of current capabilities, his accomplishments fade in comparison, but they are still regarded as key moments in the history of artificial intelligence. Advancements in storage and processing capacities have led to the development of machine learning-based products, including Netflix recommendation algorithms and self-driving automobiles.

Machine learning is one of the key elements in the rapidly expanding disciplines of data science. For gaining import insights in data mining projects, algorithms are trained using statistical techniques for building the classification and predictive model. The decisions made as a result of these insights serve as a key growth indicator in enterprises and applications. Data Scientists are more in demand as big data is developing and flourishing. They were expected to assist in determining business questions and with information needed to address them. The accelerating solution development frameworks like TensorFlow and PyTorch are utilized for machine learning algorithm development.

Types of Machine Learning

3.2.1 Supervised Learning

Supervised learning in machine learning built on supervision and trained with labeled datasets. i.e., each input data is associated with a corresponding target or label. The goal of supervised learning is to learn a mapping from inputs to outputs providing the models to make accurate predictions or classifications on the new unseen dataset. It requires a labeled dataset as input and will give labeled output is one of the key features of supervised learning. The ultimate aim of this type is to minimize the difference between the predicted and actual labels. Versatility, achieving high accuracy with sufficient labeled data, and applications across various domains are some of the advantages of supervised machine learning.

Supervised machine learning is widely used in predictive demographics such as population growth or health metrics utilizing regression and in image classification for recognizing images and identifying objects in the images, in sentimental analysis to determine the sentiment expressed in a piece of text, in diagnosing diseases based on patient data, language translation for translating text from one language to another, medicine, etc. Linear regression, support vector machines, neural networks, decision trees, etc. are examples of algorithms that can be used in supervised learning. Versatility, achieving high accuracy with sufficient labeled data, and applications across various domains are some of the advantages of supervised machine learning.

3.2.2 Unsupervised Learning

Unsupervised learning involves training the model on an unlabeled dataset., without any target labels. It is built without any supervision thus the algorithm is provided with an

input dataset trained to group objects by common characteristics instead of rewarding or optimizing to a specific output. Models are trained using unclassified, unlabeled data and they behave autonomously on the data. The objective is to find hidden patterns, structures, and relationships in classifying or grouping the unsorted dataset. This is useful in discovering hidden patterns and for exploratory data analysis.

A clustering technique for grouping the data points with similarity and dimensionality reduction technique for preserving the important features in reducing the number of features. Examples of the algorithms used in Unsupervised learning include hierarchical clustering, K-means clustering, autoencoders, principal component analysis (PCA), etc. Some of the applications of this method are customer segmentation for grouping customers based on their purchasing behavior, extracting underlying themes from the collection of documents, anomaly detection in identifying unusual patterns in data, and topic modeling.

3.2.3 Semi-Supervised Learning

Semi-supervised learning is a combined form of both Supervised learning and Unsupervised learning. It includes a large amount of unlabeled data and a small amount of labeled data for training the model. This became different from other types since unsupervised and supervised learning is dependent on the lack or presence of labels. This aims to address the drawbacks of both unsupervised and supervised learning methods. In this method the dataset is first clustered with an unsupervised learning technique then it can be labeled based on the labeled data making use of all the accessible data. The model performance can be improved and the generalization can be enhanced by making use of the unlabeled data even if the labeled data is scarce.

Using the semi-supervised learning in speech analysis for speech recognition utilizing a large amount of unlabeled speech data along with a small amount of transcribed speech data, document classification training a model on a larger unlabeled text corpus and a smaller labeled text dataset, sentimental analysis, spam detection, etc. can be employed. Generic semi-supervised algorithms, the algorithms specially designed for semi-supervised learning, and the traditional algorithms that can be employed by incorporating the unlabeled data into the dataset are the algorithms used in semi-supervised learning.

3.2.4 Reinforcement Learning

Reinforcement learning for maximizing the reward signal by making a sequence of decisions within an environment with the help of training agents. By trailing, striking, and acting the AI agent (software component) explores the surroundings thus it can learn from the experience and can increase its performance. The algorithms in these methods are trained by many trial and error experiments and they learn from the environment on interaction and receive feedback in the form of rewards or penalties. The goal of reinforcement learning is to maximize the reward since the agent is rewarded for good action and penalized for bad action. It works on the experiences of agents and that can be compared with the thoughts of people.

The reinforcement learning method is used in NLP for translation, machine translation, question answering, text summarization, and dialogue generation, in video games and games like chess, in Robotics for teaching robots to fly, walk, or manipulate objects, and in Autonomous driving for developing self-driving cars, etc. Q-learning, Proximity Policy Optimization (PPO), and Deep Q Networks (DQN) are the algorithms used in reinforcement learning. This method is suitable for sequential decision-making problems, for scenarios where actions have delayed consequences and can handle environments with changing uncertainty and dynamics.

3.3 Exploratory Data Analysis (EDA)

EDA is a statistical approach that uses data visualizations to understand the characteristics of the given dataset. In 1970, John Tukey promoted the EDA method which helps in the formulation of hypotheses and to explore the data. It is an analysis technique employed to gain information regarding the dataset such as the summary of the main characteristics and provides data visualization after the analysis. EDA helps in gathering unexpected data discoveries, selecting the necessary statistical tools and techniques, removing unnecessary values and irregularities, and helps in reducing the occurrence of errors in future analysis. The data visualization techniques in EDA to understand the distribution of data include quantile-quantile (Q-Q) plots, boxplots, histograms, etc.

3.4 Classification Models

There are several classification models in the machine learning algorithm that help in classifying the dataset into different classes, based on their characteristics when trained

with the training data and tested with the unseen test data. Various classification models used in this study are Support Vector Machine (SVM), Random Forest, Logistic Regression, and K Nearest Neighbor (KNN).

3.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a Supervised Learning technique in the machine learning algorithm used for both Regression and Classification tasks. The purpose of SVM is to create the decision boundary that separates different classes in n-dimensional space. The algorithm is named a Support Vector Machine since the support vectors or the extreme points help in creating the hyperplane.

In linear classification problems, SVM finds the best boundary or hyperplane that separates data points into different classes in an N-dimensional space. The maximum distance between the boundary and the closest data point in each class is the maximum-margin hyperplane or the optimal hyperplane. The closest data points are called the support vectors. In non-linearly separable data, the kernel trick is used, which maps the data into higher-dimensional space thus the data becomes linearly separable. An orthogonal set of vectors defines the hyperplane, where the dot product of data points and the set of vectors in that space are constants.

3.4.2 Random Forest

Random Forest is the most popular machine learning algorithm that can be used for both Regression and Classification problems belonging to Supervised Learning techniques. It is a classifier that contains several decision trees on different subsets of the dataset, each tree will make predictions and the final output will be predicted based on the majority votes of predictions. When the number of trees in the forest is higher, the accuracy will be higher as well as prevents overfitting. For a better Random Forest classifier assume that the predictions from each tree have fewer correlations, and the feature variable of the dataset should contain some actual values thus, a classifier can predict accurate results instead of guessing results.

The random forest working process includes, from the given dataset N number of decision trees built from the K data points selected at random from the training dataset then a new unseen dataset or the test dataset is put into the model where each decision tree will make

predictions. Based on these predictions the category that has majority votes will be assigned to test data points.

3.4.3 Logistic Regression

Logistic regression is a machine learning algorithm that comes under the supervised learning technique for solving classification problems. It uses the given set of independent variables for the prediction of categorical dependent variables. The outcomes will be categorical values either cancerous or non-cancerous, 0 or 1, true or false, etc. Sometimes the categorical values will be likelihood values that lie between 0 and 1. Logistic regression identifies the most effective variables used for the classification and it is used for classifying observations using different types of data. An 'S' shaped logistic function which predicts two values is fitted in this model.

Mathematically it is a sigmoid function, likelihood of the classes whether they are cancerous or not. The predicted probabilities will be between 0 and 1 in such cases the concept of the threshold value, 0.5 is considered, values less than the threshold are taken as a negative class, and that greater than the threshold are taken as a positive class. Assumptions of logistic regression include dependent variables should be categorical and independent variables should not have multi-collinearity.

3.4.4 K Nearest Neighbor (KNN)

In 1951 KNN was proposed by Fix and Hodges, which comes under the Supervised machine learning algorithm. It is a simple and essential classification algorithm that can be used for data mining, pattern recognition, and intrusion detection. KNN is known for its ease of implementation, flexibility, and simplicity. This model can handle both categorical and numerical data, does not make assumptions regarding the distribution of the data. It is flexible for both regression and classification tasks, when compared with the other algorithm it is less sensitive to outliers and makes predictions based on the closeness of the given dataset as if it is a non-parametric method.

The working of KNN follows first the selection of the optimal value of k , which is the closest neighbor considered in making predictions. Any of the distance measures is used to calculate the distance between the training and target data points. Nearest neighbors are assigned as the data points that have the closest distance to the target point and the classification is made based on the majority voting.

3.5 Model Evaluation

To ensure optimal and correct performance of a model it is necessary to evaluate the performance of the model. The model evaluation techniques test the performance of and help in identifying the best model for the problem. Two methods for model evaluation used in this study are the Confusion matrix and the ROC-AUC curve. These methods evaluate the performance of the model as per the expectation or not when put into action.

3.5.1 Confusion Matrix

The confusion matrix is used for determining the performance of the classification models in the given test data. Performance can be determined if there are true values for the test data. Also called an error matrix since the errors in model performance are displayed in the form of a matrix. The matrix has two dimensions one is predicted values and the other is actual values with the total number of predictions. Actual values are the true values for the given observations and predicted values are the values predicted by the model.

The confusion matrix for this study can be interpreted as:

		<i>Predicted values</i>	
		<i>N</i>	<i>P</i>
<i>Actual values</i>	<i>N</i>	True Negative (TN)	False Positive (FP)
	<i>P</i>	False Negative (FN)	True Positive (TP)

True Positive (TP) – The model predicted Yes and the actual classification was Yes.

False Negative (FN) – The model predicted No and the actual classification was Yes, it is called a Type-II error.

False Positive (FP) – The model predicted Yes and the actual classification was No, it is called a Type-I error.

True Negative (TN) – The model predicted No and the actual classification was No.

Various calculations are done using this confusion matrix for evaluating the performance of the model such as:

Accuracy: This determines the accuracy of the classification problem and defines how the model predicts the correct output is the number of correct predictions to all the number of predictions made by the model.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision: The number of actual true predictions to all the positive class predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is defined as how the model predicted correctly out of total predictions.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: It is difficult to compare the models if has high recall and low precision or vice versa. In this situation evaluation of recall and precision is done at the same time using F1-score. If recall is equal to precision, then the F1-score is maximum.

$$F1 - score = \frac{2 * Recall * precision}{Recall + Precision}$$

3.5.2 ROC-AUC Curve

The receiver operating characteristic (ROC) curve is a graphical plot that plots the varying threshold values and depicts the performances of a binary classifier or multi-class classifier. This plots the False Positive Rate (FPR) against the True Positive Rate (TPR) at each threshold. Plots the statistical power as a function of Type-II Error in the decision rule thus, this curve is sensitivity or recall as a function of False Positive rate. If the probability distribution of False Positive and True Positive is known then the Cumulative Distribution Function (CDF) is shown for the ROC curve. From the ROC curve, we can identify the optimal model independently from the class distribution or the class context.

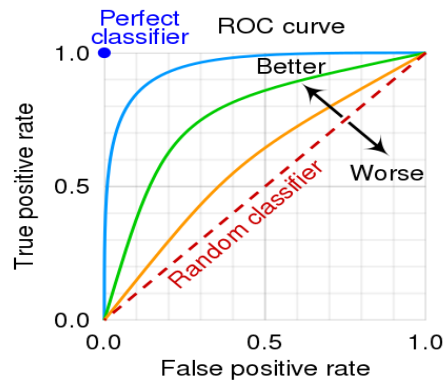


Figure 3.1

The plot of a ROC curve is shown in Figure 3.1, to portray this only the True Positive Rate (TPR) and the False Positive Rate (FPR) are needed. Plotting FPR along the x-axis and TPR along the y-axis demonstrates the relative trade-off between true positives and false positives. The ROC curve is sometimes called the sensitivity vs specificity plot, where TPR represents sensitivity and FPR represents specificity. Each point in the ROC space is the prediction result of the confusion matrix.

When there is a point in the ROC space with coordinate (0,1) this is the best prediction method possible called perfect classification. The line through the diagonal of the ROC space connecting the bottom left corner and the top right corner is the line of no discrimination. This diagonal divides the space into better and worse classification spaces. The prediction results above the diagonal show better classification and those below the diagonal show worse classification.

3.6 Feature Importance

A feature importance plot is a graphical plot that exemplifies the significance of different variables in a dataset in predicting outputs using a machine learning model. This plot commonly used in data analytics, statistics, and machine learning helps in understanding which features in the given dataset have relative importance in the model's prediction. Feature importance plot helps in gaining insights about the relationship between features and the target variables and also for explaining the results of a model.

3.7 Methodology of the study

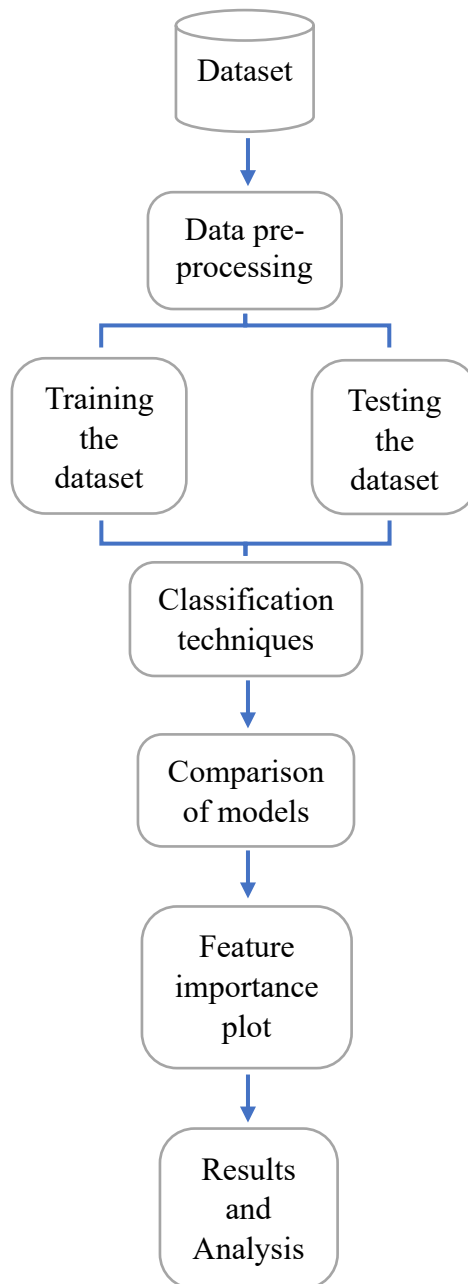


Figure 3.2

Chapter 4

Data Description And Exploratory Data Analysis

4.1 Attributes

There are 14 variables in the Kaggle Urinary Biomarkers for Pancreatic Cancer dataset. The target variable in this study is 'diagnosis' and the rest 13 variables are 'sample id', 'patient cohort', three demographic variables ('sample origin', 'age', 'sex'), 'stage', 'benign sample diagnosis', 'plasma CA19-9', and five urine biomarkers ('creatinine', 'LYVE1', 'REG1B', 'TFF1', 'REG1A').

1. Sample Id: The unique labels for each sample in the dataset, there are 591 samples in total.
2. Patient cohort: The groups or different classes of samples based on any specific criteria. Here the groups are A and B, that is the samples used before and the newly added samples.
3. Sample origin: The places where the samples in the dataset are collected. The sample origins are Barts Pancreas Tissue Bank (BPTB), London, UK, Spanish National Cancer Research Centre (ESP), Madrid, Spain, Liverpool University (LIV), UK, University College London (UCL), UK.
4. Age: The age of the patient samples in years ranges from 26 to 89.
5. Sex: Gender of patient samples, Male or Female.
6. Diagnosis: The diagnosis of the patient samples based on their medical condition is a categorical variable. In this dataset, the patient samples were diagnosed as Healthy patients, Benign cases, and PDAC cases.
7. Stage: The stage of samples diagnosed as PDAC cases to identify the patients are in which stage among IA, IB, IIA, IIIB, III, IV.
8. Benign sample diagnosis: The diagnosis of samples with benign cases such as patients with Abdominal pain, Biliary disorders, Gallbladder conditions, Gastric

conditions, Pancreatic conditions, Premalignant lesions, Serous cystadenoma, Simple benign liver cysts, Duodenal disorders, Choledocholithiasis.

9. Plasma CA19-9: The levels of CA19-9 monoclonal antibodies in the blood plasma of the patient.
10. Creatinine: The levels of plasma creatinine in the blood of patients for indicating kidney function.
11. LYVE1: The lymphatic vessel endothelial hyaluronan receptor 1 protein level in the urine plays a role in the lymphatic system maintenance and development of malignant tumors.
12. REG1B: The level of Regenerating islet-derived 1 beta of protein in the urine that is involved in the differentiation, cell growth, and tissue repair process and associated with pancreatic beta-cell function.
13. TFF1: The level of Trefoil factor 1 protein in the urine plays a role in the repair of the gastrointestinal tract and mucosal protection, also known as pS2 protein.
14. REG1A: The level of Regenerating islet-derived 1 alpha protein in the urine is similar to REG1B associated with pancreatic beta-cell function responsible for the production and secretion of insulin.

The actual diagnosis classes in the dataset are healthy patients (183 samples), benign cases (208 samples), and PDAC cases (199 samples). The main aim of this study is to diagnose whether the sample is cancerous or non-cancerous. Thus, the samples of healthy patients and benign cases are combined to make it a non-cancerous class of 391 samples. This will help in the classification of diagnoses focusing more on the samples with cancer, PDAC cases which may be treated as a crucial approach for the early detection of Pancreatic Ductile Adenocarcinoma.

4.2 Sample Dataset

	sample_id	patient_cohort	sample_origin	age	sex	diagnosis	stage	benign_sample_diagnosis	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
0	S1	Cohort1	BPTB	33	F	Non Cancerous	NaN	NaN	11.7	1.83222	0.893219	52.94884	654.282174	1262.000
1	S10	Cohort1	BPTB	81	F	Non Cancerous	NaN	NaN	NaN	0.97266	2.037585	94.46703	209.488250	228.407
2	S100	Cohort2	BPTB	51	M	Non Cancerous	NaN	NaN	7.0	0.78039	0.145589	102.36600	461.141000	NaN
3	S101	Cohort2	BPTB	61	M	Non Cancerous	NaN	NaN	8.0	0.70122	0.002805	60.57900	142.950000	NaN
4	S102	Cohort2	BPTB	62	M	Non Cancerous	NaN	NaN	9.0	0.21489	0.000860	65.54000	41.088000	NaN

Figure 4.1

	sample_id	patient_cohort	sample_origin	age	sex	diagnosis	stage	benign_sample_diagnosis	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
585	S549	Cohort2	BPTB	68	M	Cancerous	IV	NaN	NaN	0.52026	7.058209	156.241000	525.178000	NaN
586	S558	Cohort2	BPTB	71	F	Cancerous	IV	NaN	NaN	0.85956	8.341207	16.915000	245.947000	NaN
587	S560	Cohort2	BPTB	63	M	Cancerous	IV	NaN	NaN	1.36851	7.674707	289.701000	537.286000	NaN
588	S583	Cohort2	BPTB	75	F	Cancerous	IV	NaN	NaN	1.33458	8.206777	205.930000	722.523000	NaN
589	S590	Cohort1	BPTB	74	M	Cancerous	IV	NaN	1488.0	1.50423	8.200958	411.938275	2021.321078	13200.0

Figure 4.2

Figure 4.1 and Figure 4.2 shows the head and tail of the dataset.

4.3 Data Pre-processing

In the dataset, there is the presence of unwanted features. Figure 4.3 is the dataset after dropping the unwanted features and maintaining the 9 key features such as ‘age’, ‘sex’, ‘diagnosis’, ‘plasma CA19-9’, ‘creatinine’, ‘LYVE1’, ‘REG1B’, ‘TFF1’, and ‘REG1A’.

	age	sex	diagnosis	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
183	32	F	Non Cancerous	12.0	1.164930	5.417692	21.135000	445.725000	NaN
179	65	F	Non Cancerous	34.0	1.244100	0.228900	17.199000	273.424000	NaN
258	55	M	Non Cancerous	NaN	0.746460	1.488295	2.872698	227.598800	46.980
348	61	M	Non Cancerous	NaN	0.305370	2.321889	13.833203	0.023819	63.387
297	42	M	Non Cancerous	116.0	0.316680	0.304016	11.457000	672.536000	NaN

Figure 4.3

The dataset is checked for null values then Figure 4.4 shows the presence of null values in the features ‘plasma_CA19_9’ and ‘REG1A’ of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 590 entries, 0 to 589
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  -
0   age             590 non-null   int64
1   sex             590 non-null   object
2   diagnosis       590 non-null   object
3   plasma_CA19_9   350 non-null   float64
4   creatinine      590 non-null   float64
5   LYVE1          590 non-null   float64
6   REG1B          590 non-null   float64
7   TFF1           590 non-null   float64
8   REG1A          306 non-null   float64
dtypes: float64(6), int64(1), object(2)
memory usage: 41.6+ KB
```

Figure 4.4

These null values are handled by the imputation of median values of corresponding features to the null space.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 590 entries, 0 to 589
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   age                    590 non-null   int64
1   sex                    590 non-null   object
2   diagnosis              590 non-null   object
3   plasma_CA19_9         590 non-null   float64
4   creatinine            590 non-null   float64
5   LYVE1                 590 non-null   float64
6   REG1B                 590 non-null   float64
7   TFF1                  590 non-null   float64
8   REG1A                 590 non-null   float64
dtypes: float64(6), int64(1), object(2)
memory usage: 41.6+ KB

```

Figure 4.5

Now Figure 4.5 displays that the dataset has no null values. Then the sample dataset after handling the null values is shown in Figure 4.6

	age	sex	diagnosis	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
273	68	F	Non Cancerous	28.000000	0.62205	0.307446	6.124705	301.795000	50.0070
458	68	M	Cancerous	1475.000000	0.27144	7.056607	389.514560	1632.246564	1315.3200
398	77	F	Cancerous	55.600000	0.57681	2.081535	22.244200	54.804740	1186.0000
313	78	M	Non Cancerous	26.500000	1.02921	1.777817	465.267000	2962.862000	208.5385
372	67	F	Non Cancerous	11.000000	0.80301	0.003212	3.751000	336.579000	208.5385

Figure 4.6

4.4 Exploratory Data Analysis

Conducting exploratory data analysis for ideas regarding the dataset from some of the basic results obtained. The description of the dataset for getting a preliminary understanding of the dataset is given in Figure 4.7 below

	age	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
count	590.000000	590.000000	590.000000	590.000000	590.000000	590.000000	590.000000
mean	59.079661	398.747509	0.855383	3.063530	111.774090	597.868722	481.730488
std	13.109520	1896.028213	0.639028	3.438796	196.267110	1010.477245	1095.179818
min	26.000000	0.000000	0.056550	0.000129	0.001104	0.005293	0.000000
25%	50.000000	17.000000	0.373230	0.167179	10.757216	43.961000	195.201000
50%	60.000000	26.500000	0.723840	1.649862	34.303353	259.873974	208.538500
75%	69.000000	41.750000	1.139482	5.205037	122.741013	742.736000	224.007000
max	89.000000	31000.000000	4.116840	23.890323	1403.897600	13344.300000	13200.000000

Figure 4.7

The description shows that the age of patients under the study ranges from 26 to 89 years with a mean of 59 years among this range majority of the individuals are between the age group 50 and 69 years. The plasma CA19-9 levels with a mean of 654 units and a standard deviation of 2430.32 units that ranges from 0 to 31000 units show a large variation between the mean value and the maximum value thus, there are some extreme values in these levels. Creatinine levels range from 0.056 to 4.117 with a mean of 0.855 and a low standard deviation of 0.639 and most of the values fall below the third quartile which is 1.14. From the other urine biomarkers LYVE1, REG1B, TFF1, and REG1A levels their means showed 3.06, 111.77, 597.87, and 735.28 with considerable variability.

For deeper insights and interpretation look for the data visualizations.

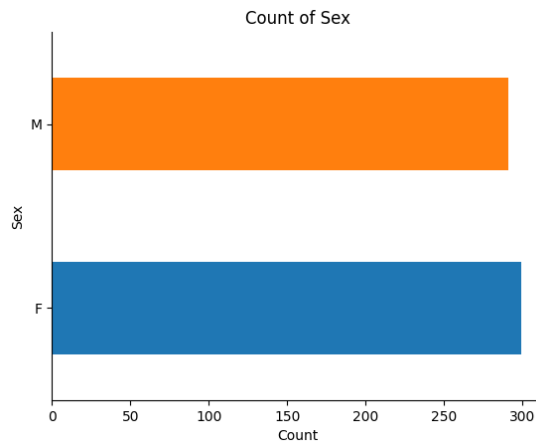


Figure 4.8

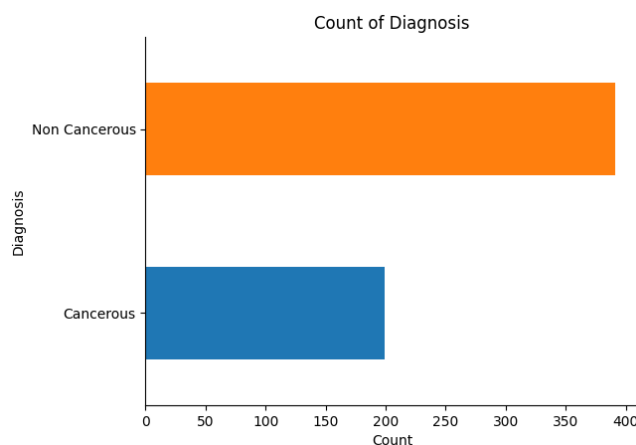


Figure 4.9

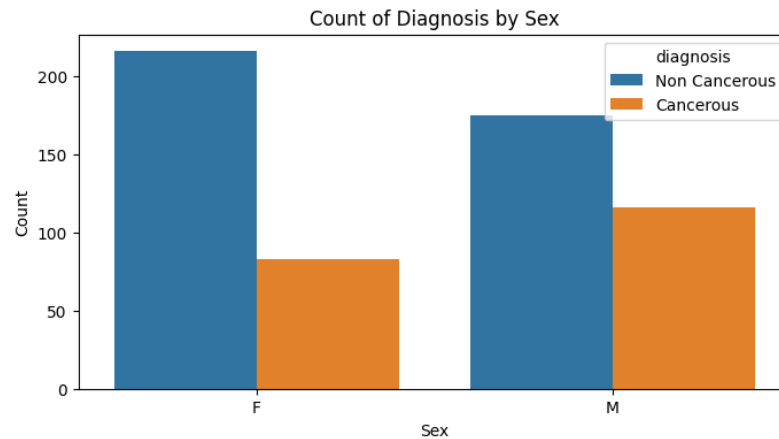


Figure 4.10

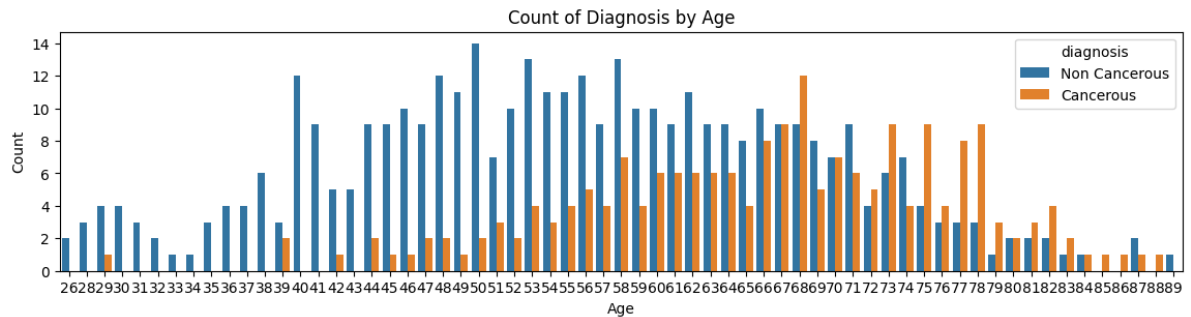


Figure 4.11

Figure 4.8 is the count plot of sex which shows that the count of females is slightly higher than the count of males and figure 4.9 shows that the dataset is imbalanced with the non-cancerous cases being more than the cancerous cases.

Now from Figure 4.10, the count plot for sex by diagnosis the count of the males with cancer is higher than the females with cancer. Cancer cases are seen among individuals aged 40 to 84 years, which is high between the age of 58 to 78 years can be explored from the Figure 4.11.

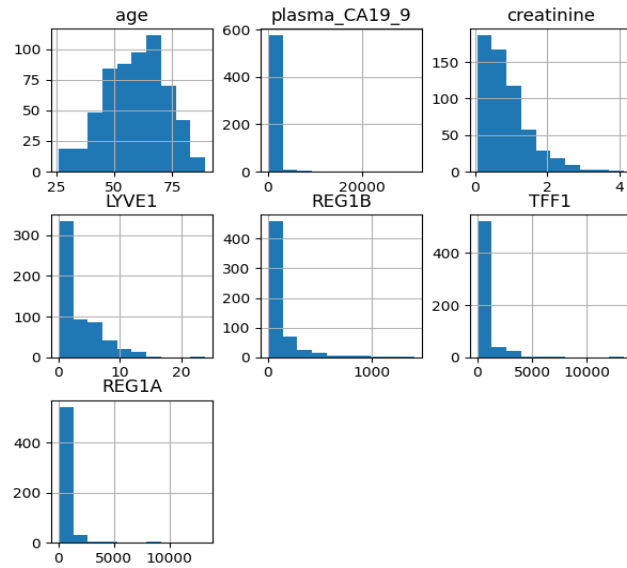


Figure 4.12

Figure 4.12 is the description of the numerical variables their age is symmetric and is normally distributed also the creatinine levels are relatively normally distributed with closer mean and median values. Then the other urine biomarkers LYVE1, REG1B, TFF1, REG1A, and plasma CA19-9 exhibits a right-skewed distribution with a mean much lower than the maximum value.

Chapter 5

Results And Discussions

5.1 Model Building

5.1.1 Support Vector Machine (SVM)

The Support Vector Machine model has 86 percent accuracy in classifying 70 and 32 cases as non-cancerous and cancerous cases. Of the samples, 6 were classified as non-cancerous and 10 were classified as cancerous samples incorrectly are shown in Figure 5.1. Since this study includes imbalanced classes looking at other metrics the model has 87 percent precision, 86 percent recall then 87 percent F1 score, and has an AUC value = 0.94 depicted in Figure 5.2 of ROC curve.

Support Vector Machine (SVM) Metrics:

Accuracy: 0.864406779661017
 Precision: 0.8698016227008197
 Recall: 0.864406779661017
 F1 Score: 0.866058235549761

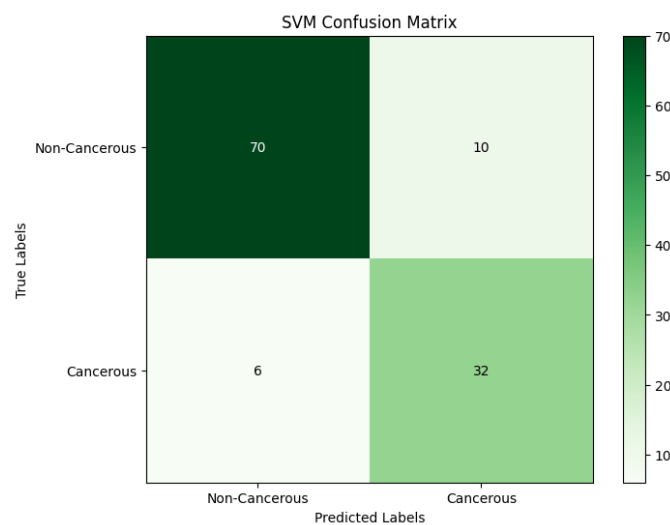


Figure 5.1

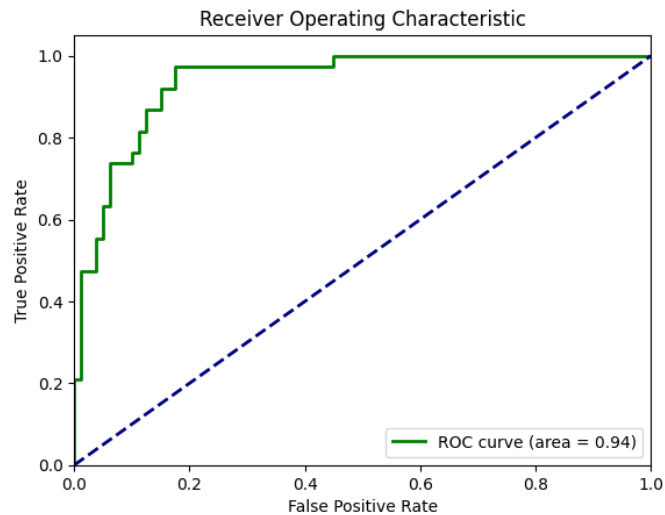


Figure 5.2

5.1.2 Random Forest

The Random Forest model has an accuracy of 90 percent by classifying 73 cases into non-cancer and 33 cases into cancer. 7 cases were wrongly classified as cancerous and 5 were wrongly classified as non-cancerous can be seen in Figure 5.3. The metrics such as precision, recall, and F1 score are considered because of the classes are imbalanced and each score is 90 percent, in Figure 5.4 the ROC curve has a 0.96 AUC value.

Random Forest Metrics:

Accuracy: 0.8983050847457628
 Precision: 0.9001847023033464
 Recall: 0.8983050847457628
 F1 Score: 0.8989652270064199

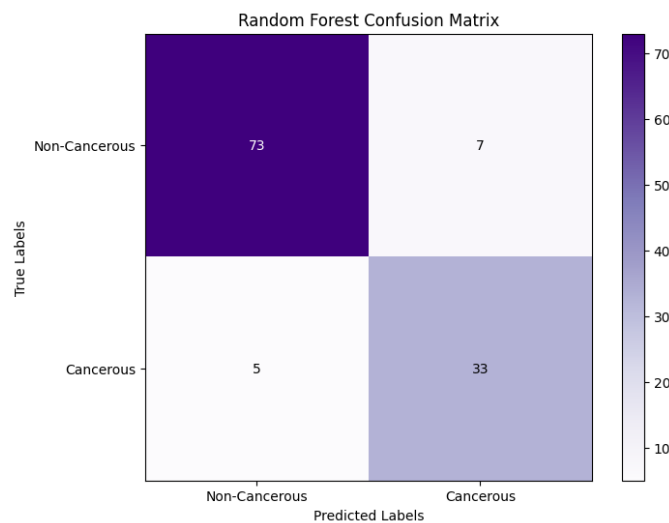


Figure 5.3

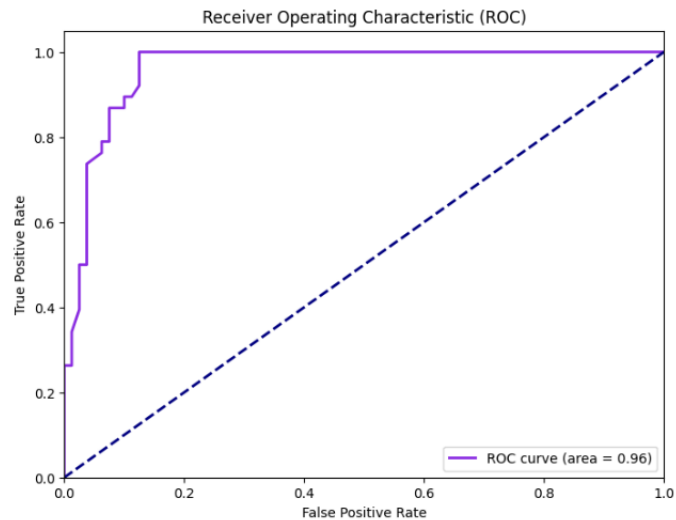


Figure 5.4

5.1.3 K Nearest Neighbor

From Figure 5.5, the confusion matrix of K Nearest Neighbor, the model wins an 81 percent accuracy on classifying 69 cases as non-cancerous and 27 cases as cancerous. Along with this 11 cases were classified as non-cancerous and cancerous each faulty. It has 81 percent precision, recall, and F1 score each for the model in which 0.89 is the AUC value displayed in Figure 5.6 of the ROC curve.

K Nearest Neighbor (KNN) Metrics:

Accuracy: 0.8135593220338984
 Precision: 0.8135593220338984
 Recall: 0.8135593220338984
 F1 Score: 0.8135593220338984

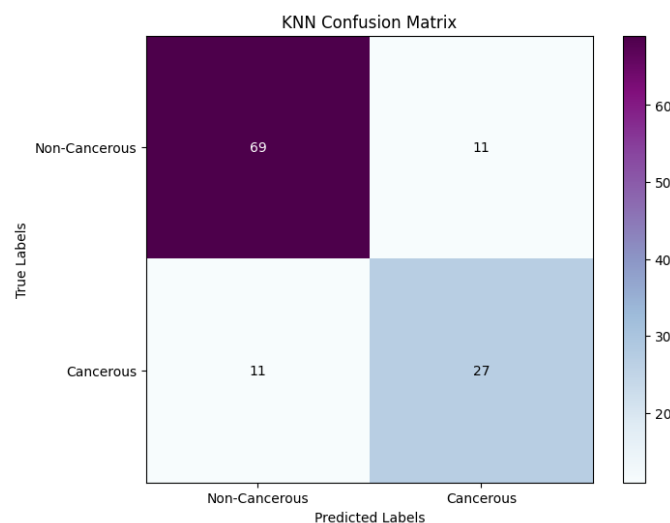


Figure 5.5

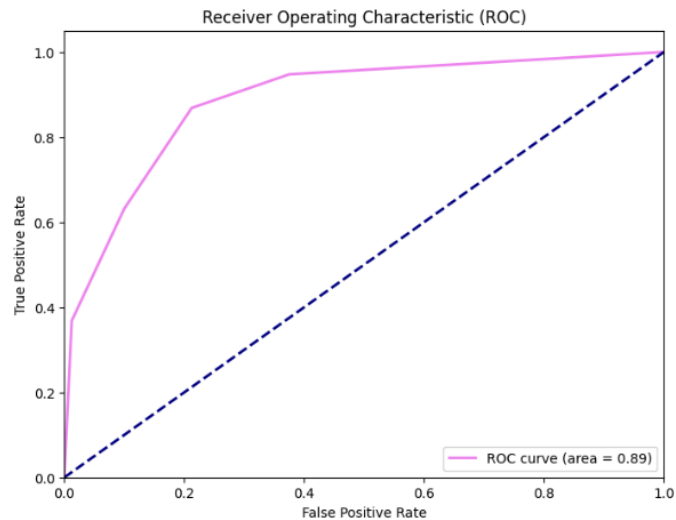


Figure 5.6

5.1.4 Logistic Regression

The logistic regression model gains an 84 percent accuracy in the classification of 69 cases and 30 cases into non-cancerous and cancerous classes in Figure 5.7. 11 cases were assigned as non-cancerous and 8 cases as cancerous by mistake. Thus, the model has 84 percent precision and recall and F1 score of 84 percent with AUC value = 0.92 in the ROC curve in Figure 5.8.

Logistic Regression Metrics:

Accuracy: 0.8389830508474576
 Precision: 0.8431626248906116
 Recall: 0.8389830508474576
 F1 Score: 0.840501956197038

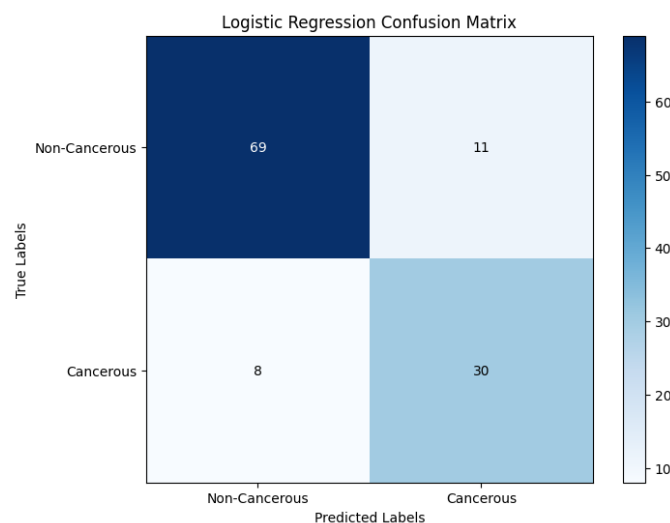


Figure 5.7

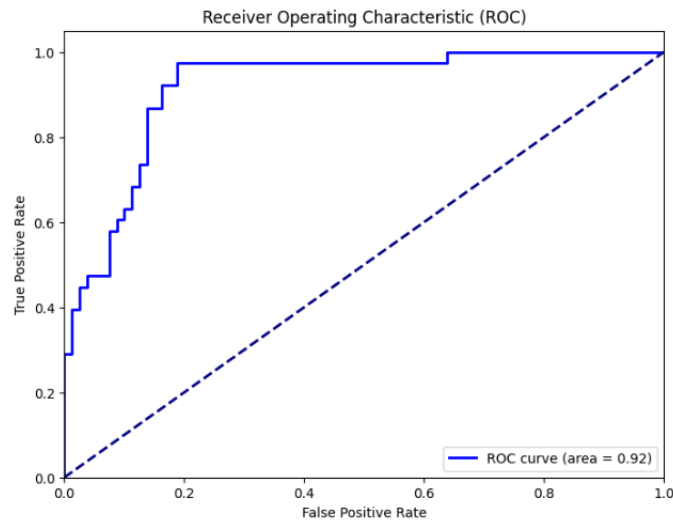


Figure 5.8

5.2 Comparison of Models

For the early diagnosis of pancreatic cancer, we build different models for classifying the samples into non-cancerous and cancerous. In Table 5.1 the Random Forest model has the highest F1 score of 90 percent, when compared to the other models.

From the performances of each model using a Receiver Operating Curve (ROC) in Table 5.1, the Random Forest has the maximum Area Under the Curve (AUC) value that is 0.96 which is closer to 1. Hence the Random Forest model is the best model to classify the dataset into non-cancerous and cancerous classes.

Models	F1 Score	AUC Value
Support Vector Machine	87%	0.94
Random Forest	90%	0.96
K Nearest Neighbor	81%	0.89
Logistic Regression	84%	0.92

Table 5.1

5.3 Feature Importance Plot

Random Forest is the best model for classifying the dataset into non-cancerous and cancerous so that the feature importance plot is built to capture the important features that help the Random Forest classify accordingly. Thus, Figure 5.9 illustrates the feature ranking based on the classification in this model. From that LYVE1 has the highest rank and thereafter comes the Plasma CA19-9, TFF1, and REG1B levels. While the LYVE1

level plays a major role in the development of malignant tumors and lymphatic vessel maintenance, Plasma CA19-9 is the level of CA19-9 monoclonal antibodies in the blood plasma of patients.

Feature Ranking:

-
1. LYVE1
 2. plasma CA19-9
 3. TFF1
 4. REG1B
 5. age
 6. creatinine
 7. REG1A
 8. sex
-

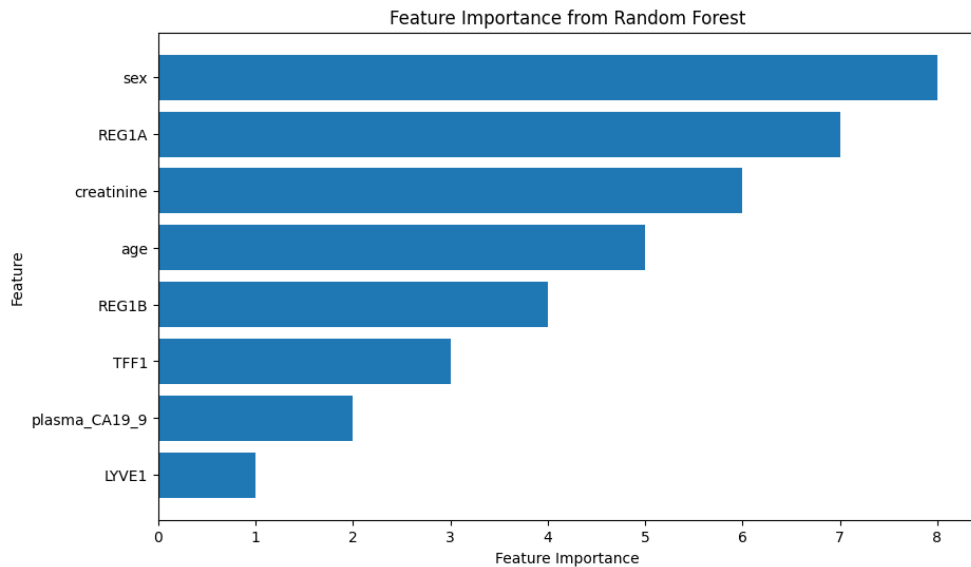


Figure 5.9

Chapter 6

Conclusion

Through this research, it has been found that the need for early detection of pancreatic cancer both in the PanNETs and in the PDAC cases is crucial. The study was done using the Kaggle Urinary Biomarkers for Pancreatic Cancer dataset. In this dataset, the four machine learning models such as Support Vector Machine (SVM), Random Forest, K Nearest Neighbor (KNN), and Logistic Regression were tested. From these models using the best model, the feature importance plot is employed for detecting the feature that is crucial for the classification of the cancerous and non-cancerous cases. The Random Forest model detected as the best model for classifying the cancerous and non-cancerous cases with a 90 percent of F1 score. From that with the help of the feature importance plot the LYVE1 levels are detected as critical in classifying the classes for the Random Forest model.

Chapter 7

References

1. Acer, İ., BULUCU, F. O., İÇER, S., & LATİFOĞLU, F. (2023). Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations. *Turkish Journal of Electrical Engineering and Computer Sciences*, 31(1), 112-125.
2. Chen, W., Zhou, Y., Xie, F., Butler, R. K., Jeon, C. Y., Luong, T. Q., ... & Wu, B. U. (2022). Prediction Model for Detection of Sporadic Pancreatic Cancer (PROTECT) in a Population-Based Cohort Using Machine Learning and Further Validation in a Prospective Study. *medRxiv*, 2022-02.
3. Floyd, S. (2007). Data mining techniques for prognosis in pancreatic cancer (Doctoral dissertation, Worcester Polytechnic Institute).
4. Ge, G., & Wong, G. W. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, 9, 1-12.
5. Hsieh, M. H., Sun, L. M., Lin, C. L., Hsieh, M. J., Hsu, C. Y., & Kao, C. H. (2018). Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer management and research*, 6317-6324.
6. Ingwersen, E. W., Stam, W. T., Meijs, B. J., Roor, J., Besselink, M. G., Koerkamp, B. G., ... & Dutch Pancreatic Cancer Group. (2023). Machine learning versus logistic regression for the prediction of complications after pancreatoduodenectomy. *Surgery*.
7. Karar, M.E., El-Fishawy, N. & Radad, M. Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks. *J Biol Eng* **17**, 28 (2023). <https://doi.org/10.1186/s13036-023-00340-0>

8. Keyl, J., Kasper, S., Wiesweg, M., Götze, J., Schönrock, M., Sinn, M., ... & Kleesiek, J. (2022). Multimodal survival prediction in advanced pancreatic cancer using machine learning. *ESMO open*, 7(5), 100555.
9. Lee, H. A., Chen, K. W., & Hsu, C. Y. (2022). Prediction model for pancreatic cancer—A population-based study from NHIRD. *Cancers*, 14(4), 882.
10. Muhammad, W., Hart, G. R., Nartowt, B., Farrell, J. J., Johung, K., Liang, Y., & Deng, J. (2019). Pancreatic cancer prediction through an artificial neural network. *Frontiers in Artificial Intelligence*, 2, 2.
11. Placido, D., Yuan, B., Hjaltelin, J. X., Haue, A. D., Chmura, P. J., Yuan, C., ... & Sander, C. (2021). Pancreatic cancer risk predicted from disease trajectories using deep learning. *BioRxiv*, 2021-06.
12. Qiu, Y., Jiang, H., Shimada, K., Hiraoka, N., Maeshiro, K., Ching, W. K., ... & Furuta, K. (2014). Towards prediction of pancreatic cancer using SVM study model. *Journal of Clinical Oncology and Research*.
13. Tong, Z., Liu, Y., Ma, H., Zhang, J., Lin, B., Bao, X., ... & Zhao, P. (2020). Development, validation, and comparison of artificial neural network models and logistic regression models predicting survival of unresectable pancreatic cancer. *Frontiers in Bioengineering and Biotechnology*, 8, 196.
14. Vaiyapuri, T., Dutta, A. K., Punithavathi, I. H., Duraipandy, P., Alotaibi, S. S., Alsolai, H., ... & Mahgoub, H. (2022, April). Intelligent deep-learning-enabled decision-making medical system for pancreatic tumor classification on CT images. In *Healthcare* (Vol. 10, No. 4, p. 677). MDPI.
15. Website: (<https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer>)

