**UNVEILING THE CODE OF LIFE : A GENETIC RISK PREDICTION MODEL**

**A PROJECT REPORT**

###### **Submitted by**

##### **Shreya Raj Gupta** (22BHI10143)

**Shashmita Sanyal** (22BHI10034)

**Jigisha Dharskar** (22BHI10050)

##### **Isha Choudhary** (22BHI10124)

*in partial fulfillment for the award of the degree*

*of*

**BACHELORS OF TECHNOLOGY**

*In*

**COMPUTING SCIENCE & ENGINEERING**

**(HEALTH INFORMATICS)**

****

**SCHOOL OF COMPUTING SCIENCE ENGINEERING AND ARTIFICIAL INTELLIGENCE**

**VIT BHOPAL UNIVERSITY**

**KOTHRIKALAN, SEHORE**

**MADHYA PRADESH - 466114**

##### MAY 2024

**VIT BHOPAL UNIVERSITY, KOTHRIKALAN, SEHORE**

**MADHYA PRADESH – 466114**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“UNVEILING THE CODE OF LIFE : GENETIC RISK PREDICTION”** is the bonafide work of “**Shreya Raj Gupta (22BHI10143), Shashmita Sanyal (22BHI10034), Jigisha Dharskar (22BHI10050), Isha Choudhary (22BHI10124)”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

**PROGRAM CHAIR PROJECT GUIDE**

Dr.Swagat Kumar Samantaray, Dr. Garima Jain ,

Assistant Professor Sr-I Senior Teaching Fellow

School of Computing Science School of Computing

and Engineering Science and Engineering VIT BHOPAL UNIVERSITY VIT BHOPAL UNIVERSITY

The Project Exhibition II Examination is held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENT**

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr. Pon Harshvardhan, Head of the Department, School of Computing Science Engineering and Artificial Intelligence for much of his valuable support and encouragement in carrying out this work.

I would like to thank my internal guide Ms.Garima Jain ,for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computing Science Engineering and Artificial Intelligence, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

**LIST OF ABBREVIATIONS**

**IBD** - Inflammatory Bowel Disease

**GWAS** - Genome-Wide Association Study

**SNP** - Single Nucleotide Polymorphism

**CD** - Crohn's Disease

**UC** - Ulcerative Colitis

**EO-IBD** - Early-onset Inflammatory Bowel Disease

**T2D** - Type 2 Diabetes

**EO-T2D** - Early-onset Type 2 Diabetes

**PRS** - Polygenic Risk Score

**OR** - Odds Ratio

**R**² - R-squared (coefficient of determination)

**LIST OF FIGURES AND GRAPHS**

| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
|  | **Functional architecture of our prediction model and its distribution.** | **21** |
|  | **Software Architecture Design of Type 2 Diabetes Prediction** | **23** |
|  | **Software Architectural Design Of IBD Crohn’s Disease** | **23** |
|  | **Relationship between different parameters and chances of having diabetes** | **25** |
|  | **Relationship between each pair of parameters** | **26** |
|  | **Genotype density distribution for first 25 sample patients** | **26** |
|  | **Distribution of SNPs after cleaning.** | **27** |
| **8** | **This is the correlation map which we get as an output of variables in a dataset.** | **29** |
| **9** | **Fig. 9 ROC and AUC Curve.** | **32** |
| **10.** | **Confusion matrix for Random Forest Classifier** | **35** |
| **11.** | **ROC and AUC Curve for Disease Risk Prediction with SVC** | **38** |

**LIST OF TABLES**

| **TABLE NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
| **1.** | **Testing and Validation Table for Different classifiers for Type-II Diabetes and IB Disease** | **40** |
| **2.** | **Performance analysis Table for T2D and IBD Disease** | **41** |
| **3.** | **Project outcome table of Diabetes and IBD** | **44** |

**ABSTRACT**

**[PURPOSE-METHODOLOGY-FINDINGS]**

This study explores the genetic underpinnings of complex diseases, specifically comparing the genetic architecture of early-onset (EO) versions of Inflammatory Bowel Disease (IBD) and Type 2 Diabetes (T2D) to their adult-onset counterparts in European populations. Immunochip genotyping was performed on EO-IBD cases and controls (IBD: n=1008, Controls: n=1633) to identify susceptibility genes. While known genes like NOD2 and IL23R were replicated for EO-IBD, only a small fraction of loci previously linked to adult-onset IBD were found (4% for Crohn's disease, 0.8% for ulcerative colitis). Interestingly, the NOD2 variant displayed a potentially different effect size in EO-IBD compared to adult-onset. A similar approach is planned for T2D to compare the genetic architecture of EO-T2D with adult-onset. Polygenic risk scores will also be investigated for both diseases to assess their ability to predict age of onset. These initial findings suggest a broadly similar genetic architecture between EO diseases and adult-onset, but highlight potential differences in specific gene effects and the need for further investigation to identify additional genetic factors contributing to EO-IBD and EO-T2D.

**TABLE OF CONTENTS**

| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
|  | List of Abbreviations  List of Figures and Graphs  List of Tables  Abstract | iii  iv  v  vi |
| 1 | **CHAPTER-1:**  **PROJECT DESCRIPTION AND OUTLINE** Introduction 1.2 Motivation for the work  1.3 Problem Statement  1.4 Objective of the work  1.5 Summary | 1  .  .  . |
| 2 | **CHAPTER-2:**  **RELATED WORK INVESTIGATION**  2.1 Introduction  2.2 Core area of the project  2.3 Existing Approaches/Methods  2.3.1 Polygenic Risk Scores (PRS) from GWAS 2.3.2 Functional Annotation Integration Methods  2.4 Pros and cons of the stated Approaches/Methods  2.4.1 PRS from GWAS  2.4.2 Functional Annotation Integration Methods  2.5 Issues/observations from investigation  2.6 Summary |  |
| 3 | **CHAPTER-3:**  **REQUIREMENT ARTIFACTS**  3.1 Introduction  3.2 Hardware and Software requirements  3.3 Specific Project requirements  3.3.1 Data requirement  3.3.2 Functions requirement  3.3.3 Performance and security requirement  3.4 Summary |  |
| 4 | **CHAPTER-4:**  **DESIGN METHODOLOGY AND ITS NOVELTY**  4.1 Methodology and goal  4.2 Functional modules design and analysis  4.3 Software Architectural designs  4.4 User Interface designs |  |
| 5 | **CHAPTER-5:**  **TECHNICAL IMPLEMENTATION & ANALYSIS**  5.1Outline  5.2 Technical coding and code solutions  5.3 Working Layout of Forms  5.4 User interface Prototype  5.5 Test and validation  5.6 Performance Analysis(Graphs/Charts)  5.7 Summary |  |
| 6 | **CHAPTER-6:**  **PROJECT OUTCOME AND APPLICABILITY**  6.1Outline  6.2 key implementations outlines of the System  6.3 Significant project outcomes  6.4 Project applicability on Real-world applications  6.4 Inference |  |
| 7 | **CHAPTER-7:**  **CONCLUSIONS AND RECOMMENDATION**  7.1Outline  7.2 Limitation/Constraints of the System  7.3 Future Enhancements  7.4 Inference |  |
|  | References |  |

# **CHAPTER-1:**

**PROJECT DESCRIPTION AND OUTLINE**

# **1.1 Introduction**

# Advancements in Genome-Wide Association Studies (GWAS) have unveiled a treasure trove of genetic variants associated with complex diseases like Type 2 Diabetes (T2D) and Inflammatory Bowel Disease (IBD). However, translating this knowledge into accurate risk prediction for individual patients remains a challenge. Current methods often fall short, hindering our ability to proactively identify those at high risk. This necessitates a new approach. We propose a novel framework that integrates diverse functional annotations from the genome with GWAS summary statistics. By incorporating this richer data landscape, we aim to surpass the limitations of existing models that neglect functional details and struggle with generalizability across populations. This framework has the potential to significantly improve the accuracy of disease risk prediction for both T2D and IBD.

# **1.2 Motivation for the work**

# Existing risk prediction models for type 2 diabetes and IBD often miss valuable insights by neglecting functional information about genetic variations and interactions. Additionally, limitations in generalizability due to population-specific data restrict their usefulness. Our approach tackles these issues by incorporating diverse functional annotations with GWAS summary statistics, aiming to improve prediction accuracy for both diseases.

# **1.3 Problem Statement**

# Current state-of-the-art methods for disease risk prediction using GWAS summary statistics alone are limited in accuracy. This restricts our ability to effectively identify individuals at high risk for diseases like Type II diabetes and IBD. There is a need for a novel framework that leverages various types of genomic annotations alongside GWAS data to improve the precision of risk prediction.

**1.4 Objective of the work**

The objective of this study is to develop and evaluate a novel genetic risk prediction framework, which integrates various types of genomic annotations with genome-wide association study (GWAS) summary statistics to improve the accuracy of disease risk prediction, in both extensive simulations and real data, in comparison with state-of-the-art risk prediction methods.

**1.5 Summary**

While Genome-Wide Association Studies (GWAS) have identified thousands of genetic variants linked to complex diseases like Type 2 Diabetes (T2D) and Inflammatory Bowel Disease (IBD), current methods for predicting individual risk remain limited. These limitations stem from a reliance solely on GWAS data, hindering our ability to proactively identify high-risk individuals. To address this challenge, we propose a novel framework that integrates diverse functional annotations from the genome with GWAS summary statistics. This approach aims to surpass existing models by capturing the functional significance of genetic variations often overlooked. By incorporating this richer data landscape, our framework has the potential to significantly improve the accuracy of disease risk prediction for both T2D and IBD.

**CHAPTER-2:**

**RELATED WORK INVESTIGATION**

**2.1 Introduction**

Despite the identification of numerous disease-associated variants through Genome-Wide Association Studies (GWAS), current methods for predicting individual risk for complex diseases like Type 2 Diabetes (T2D) and Inflammatory Bowel Disease (IBD) remain limited in accuracy. This investigation focuses on existing approaches that integrate functional annotations from the genome with GWAS data to improve disease risk prediction.

**2.2 Core area of the project**

This project proposes a novel framework for disease risk prediction that leverages the power of GWAS summary statistics and phenotypic records alongside diverse functional annotations from the genome. This approach aims to overcome limitations of current methods by incorporating richer information and capturing the functional significance of genetic variations often overlooked.

**2.3 Existing Approaches/Methods**

**2.3.1 Polygenic Risk Scores (PRS) from GWAS:**

GWAS have been instrumental in identifying thousands of genetic variants associated with complex diseases like T2D and IBD. These variants are often used to construct Polygenic Risk Scores (PRS) that estimate an individual's overall genetic susceptibility based on the cumulative effect of their variants. While straightforward to implement and leveraging established GWAS findings, PRS typically explain a modest proportion of disease risk and do not account for the functional aspects of the individual variants.

**2.3.2 Functional Annotation Integration Methods:**

Recognizing the limitations of PRS based solely on GWAS data, researchers have explored methods that integrate functional annotations with GWAS data for risk prediction. These annotations can encompass various aspects of the genome, such as regulatory elements that influence gene expression, protein-protein interactions, or variant locations within genes. While these methods offer the potential for improved accuracy and biological interpretability compared to PRS alone, they can be complex to implement. Challenges arise in data integration due to the vast amount and diverse nature of functional annotations, and choosing the most informative annotations for a specific disease remains an active area of research.

**2.4 Pros and cons of the stated Approaches/Methods**

**2.4.1 PRS from GWAS**

**Pros:** Simple to implement, leverages established GWAS findings.

**Cons:** Limited accuracy, lacks biological interpretability.

**2.4.2 Functional Annotation Integration Methods**

**Pros:** Integrates functional information, potentially improves accuracy and biological understanding.

**Cons:** Complex to implement, data integration challenges, chosen annotations may not be most informative.

**2.5 Issues/observations from investigation**

The current landscape of disease risk prediction methods necessitates more robust approaches for effectively integrating diverse functional annotations with GWAS data. Existing methods often focus on a limited set of annotations or employ generic analysis techniques, potentially overlooking important functional information.

**2.6 Summary**

The investigation of existing approaches highlights the limitations of current disease risk prediction methods for T2D and IBD. While PRS from GWAS offer a basic framework, their accuracy is restricted. Functional annotation integration methods hold promise for improved accuracy and biological understanding; however, challenges remain in data integration and selecting the most informative annotations. This investigation underscores the need for a novel framework that effectively leverages diverse functional annotations with GWAS data to achieve more accurate and biologically meaningful disease risk prediction for T2D and IBD.

**CHAPTER-3:**

**REQUIREMENT ARTIFACTS**

**3.1 Introduction**

This section outlines the hardware, software, and specific project requirements for developing the novel disease risk prediction framework for Type 2 Diabetes (T2D) and Inflammatory Bowel Disease (IBD).

**3.2 Hardware and Software requirements**

Hardware:-

* Computer or Server
* Storage: Adequate storage space is essential for storing datasets, model files, and intermediate results. Depending on the size of our data, we may need terabytes of storage space.
* GPU (Graphics Processing Unit)

Software:- We will need Python 2.7 and several packages to run it:

* h5py
* plinkio
* scipy
* numpy
* sklearn
* pandas == 0.18.1 (required for munge step in LDSC),etc

**3.3 Specific Project requirements**

**3.3.1 Data requirement**

GWAS Summary Statistics: The framework will utilize GWAS summary statistics for IBD & phenotypic pathological records for T2D.

* Disease: Identifying whether the statistics pertain to T2D or IBD cases.
* Variants: Specifying the genetic variants (SNPs) included in the GWASanalysis.
* Effect Sizes: Indicating the effect size of each variant on disease risk (e.g., beta coefficients).
* P-values: Including the p-values associated with each variant to assess statistical significance.
* **Source:** Specifying the source of the GWAS summary statistics (e.g., public databases like GWAS Catalog).

Functional Annotation Data: The framework will integrate various types of functional annotations alongside the GWAS data to improve prediction accuracy. The specific types of annotations will depend on available resources and their relevance to T2D and IBD. Some potential examples include: Regulatory elements (e.g., promoters, enhancers) that influence gene expression. Protein-protein interaction data to understand cellular pathways potentially disrupted by genetic variants. Gene annotations including gene location, function, and pathway membership. The data format for functional annotations will depend on the chosen tools. Common formats include BED files or databases like Gene Ontology (GO). It's crucial to specify the source of this data (e.g., public databases like ENCODE).

**3.3.2 Functions requirement**

* Data Pre-processing:  
  This stage involves cleaning, filtering, and formatting both GWAS summary statistics and functional annotation data to ensure compatibility with subsequent analysis steps. This may involve handling missing data, standardizing formats, and quality checks.
* Integration of GWAS and Functional Annotation Data: A core function of the framework will be to effectively combine the GWAS data with the chosen functional annotations. This may involve mapping variants to relevant annotations or developing weighting schemes based on the importance of specific annotations for T2D and IBD risk prediction.
* Model Development for Risk Prediction:  
   The framework will utilize machine learning or statistical modeling approaches to build a predictive model that incorporates the integrated GWAS and functional annotation data. The specific choice of model will depend on the data characteristics and desired performance metrics.
* Generation and Interpretation of Risk Scores: The framework will generate individual risk scores for T2D and IBD based on the developed model. The functionality should also allow for visualization and interpretation of these risk scores for better understanding and communication with healthcare professionals.

**3.3.3 Performance and security requirement**

* Performance: The framework should be computationally efficient, especially when handling large datasets. Aim for acceptable processing times for data pre-processing, integration, and model training/testing. Consider memory usage optimization to ensure scalability for future data growth.
* Security: If the framework handles sensitive patient data (indirectly through GWAS data), prioritize data security. Implement secure data storage practices (e.g., encryption) and access control protocols to restrict unauthorized access.

**3.4 Summary**

This section has outlined the specific project requirements for developing the disease risk prediction framework. By defining the data needs, functionalities, and performance/security considerations, we ensure a clear roadmap for building a robust and informative framework for improved T2D and IBD risk prediction. These requirements, coupled with the hardware and software specifications from Section 3.2, provide a solid foundation for project development.

**CHAPTER-4:**

**DESIGN METHODOLOGY AND ITS NOVELTY**

|  |
| --- |

**4.1 METHODOLOGY AND GOAL**

Methodology:-

This approach utilizes Genome-Wide Association Studies (GWAS) to estimate an individual's susceptibility to complex diseases like celiac disease, inflammatory bowel disease (IBD), and diabetes. It includes

1. **Leveraging GWAS Data:** GWAS investigate associations between genetic variations (Single Nucleotide Polymorphisms - SNPs) and specific diseases. The model utilizes **condensed GWAS summary statistics**.
2. **Polygenic Risk Score (PRS) Calculation:** The model calculates a PRS for each individual. This involves:

* Obtaining the individual's genotype data.
* Referencing the effect sizes from the GWAS summary statistics.
* Weighting the effect sizes based on the individual's alleles (higher weight for risk allele)
* Summing the weighted effect sizes across all included SNPs to create the PRS.

Goals:-

Our genetic risk prediction model aims to unlock the message within our DNA. It calculates a Polygenic Risk Score (PRS) that estimates how your genes might influence your risk for complex diseases. This score allows healthcare professionals to potentially recommend earlier interventions or more personalized treatment options based on the unique genetic structure. Ultimately, it's a tool to translate our genes into useful information for our health.

**4.2 FUNCTIONAL MODULES DESIGN AND ANALYSIS**

****

**Fig. 1 Functional architecture of our prediction model and its distribution.**

**Analysis of the functional model:**

So we have tested our prediction model on various classifiers such as logistic regression, random forest and SVM(Support Vector Machine). So some of the pros and cons of the classifiers are:

**Logistic Regression:**

**Pros:**

(i) **It is simple to implement and interpret**: It is relatively straightforward to understand and implement, which makes it a popular choice for binary classifications.

(ii)**Efficient with small datasets:** It performs well when the dataset is small, making it suitable for scenarios with limited data availability.

**Cons:**

(i) **Assumes linearity:** Logistic regression assumes that the relationship between the independent variables and the log odds of the dependent variable is linear. It may not perform well if the relationship is highly non-linear.

(ii)**High reliance on proper feature selection:** The performance of logistic regression heavily depends on the selection of relevant features. Including irrelevant or redundant features can degrade its performance.

**Random Forest:**

**Pros:**

(i) **High accuracy:** Random Forest generally yields high accuracy by combining predictions from multiple decision trees. It reduces overfitting compared to individual decision trees, leading to better generalization.

(ii) **Robust to overfitting:** The ensemble nature of Random Forest, where multiple trees are built independently and then aggregated, helps reduce overfitting, especially when compared to deep learning models.

**Cons:**

(i) **Lack of interpretability:** While Random Forest can provide feature importance rankings, the individual decision trees within the ensemble are not easily interpretable, making it challenging to understand the reasoning behind specific predictions.

(ii) **Not suitable for very imbalanced datasets:** Random Forest may not perform well on highly imbalanced datasets where one class is significantly more prevalent than others.

**SVM(Support Vector Machine)**

**Pros:**

(i) **High Accuracy:** Random forests combine multiple decision trees, each making predictions, and the final output is an average (for regression) or vote (for classification) of these predictions.

(ii) **Handles Missing Data:** Random forests can handle missing data points relatively well without requiring extensive pre-processing.

(iii) **Can Handle Complex Relationships:** Random forests can effectively capture both linear and non-linear relationships between features and the target variable.

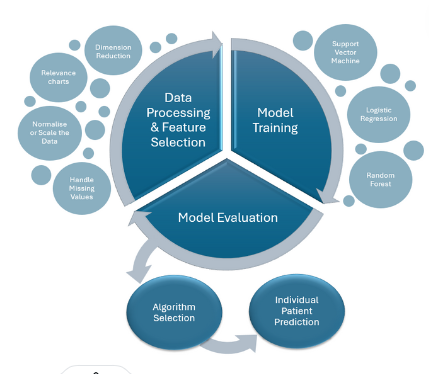
**Cons:**

(i) **Computationally Expensive:** Training a random forest with a large number of trees or using a massive dataset can be computationally demanding in terms of processing power and time.

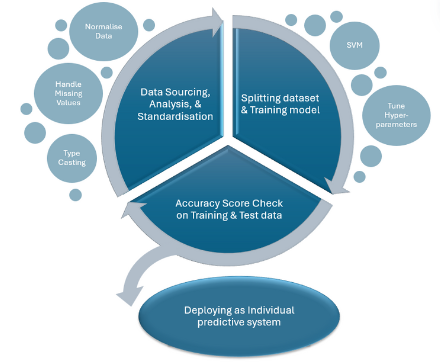
(ii) **High Memory Usage:** Random forest models, especially those with many trees or deep decision trees, can consume a significant amount of memory.

(iii) **Slower Predictions:** While training can be fast, making predictions with a random forest model can be slower compared to some simpler algorithms.

**4.3 SOFTWARE ARCHITECTURE DESIGN**

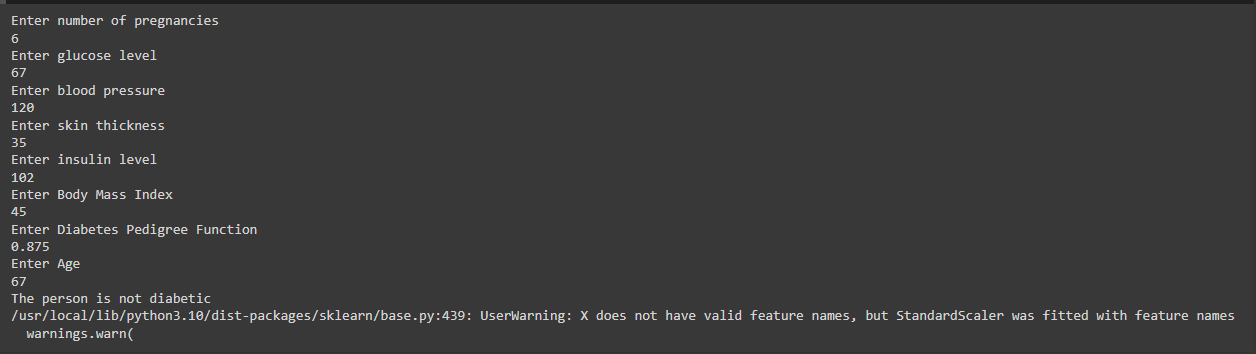
****

**Fig 2. Software Architecture Design of Type 2 Diabetes Prediction**

****

**Fig. 3 Software Architectural Design Of IBD Crohn’s Disease**

**4.4 User Interface Design**

****

**CHAPTER-5**

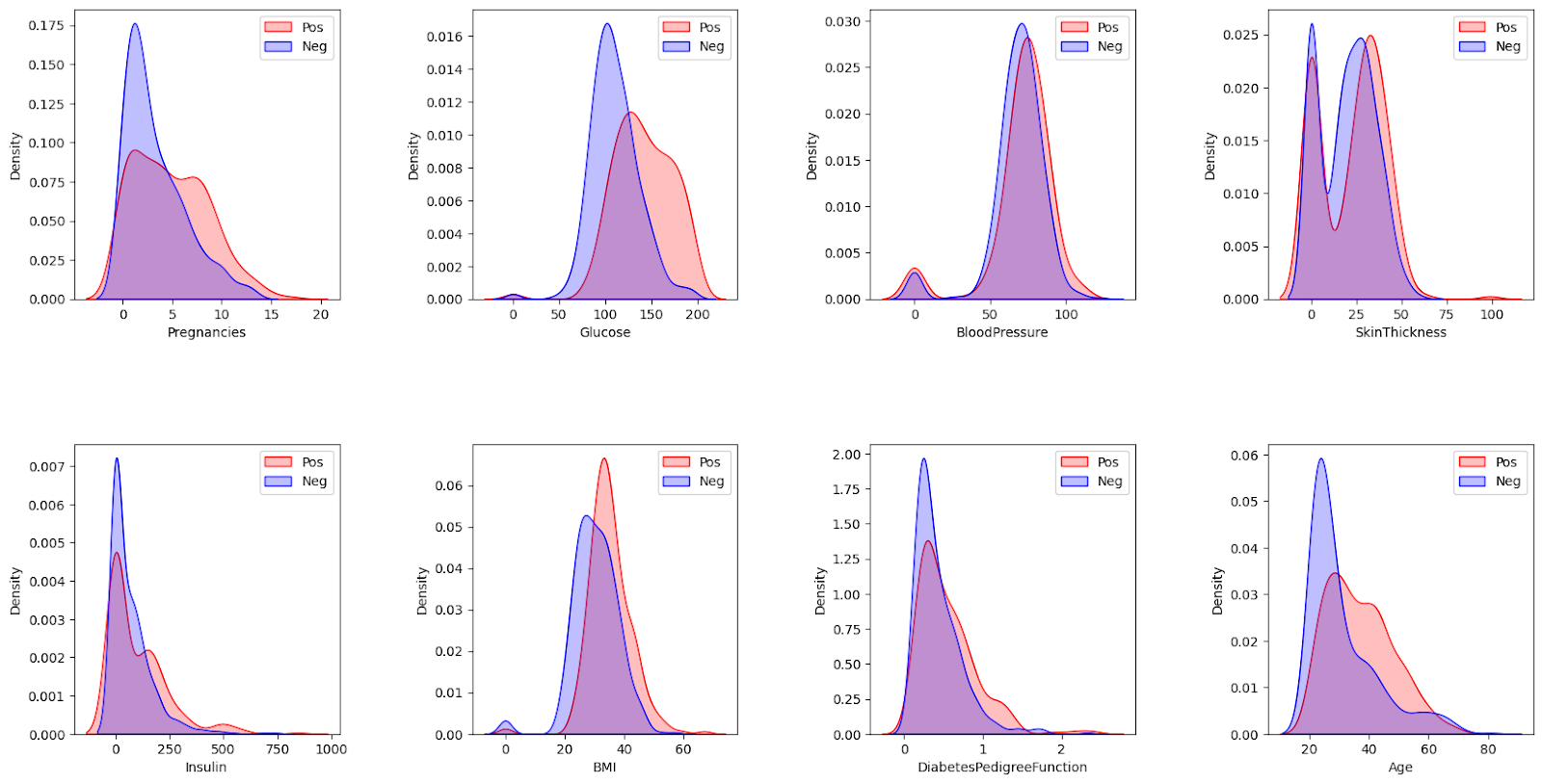
**TECHNICAL ASSESSMENT AND ANALYSIS**

**5.1 Outlines:**

In the first stage, we'll go into how we prepared the data to train our prediction model. This involves gathering the data from its original sources. Then we'll explain the methods used to bring this data into our system for processing. Then, we'll focus on cleaning the data, ensuring it's free of errors and inconsistencies. This might involve handling missing information or unusual data points. We'll also discuss any adjustments we made to the data's format or structure to improve the model's ability to learn from it. Finally, we'll mention the specific tools or libraries that helped us clean and prepare the data for training.

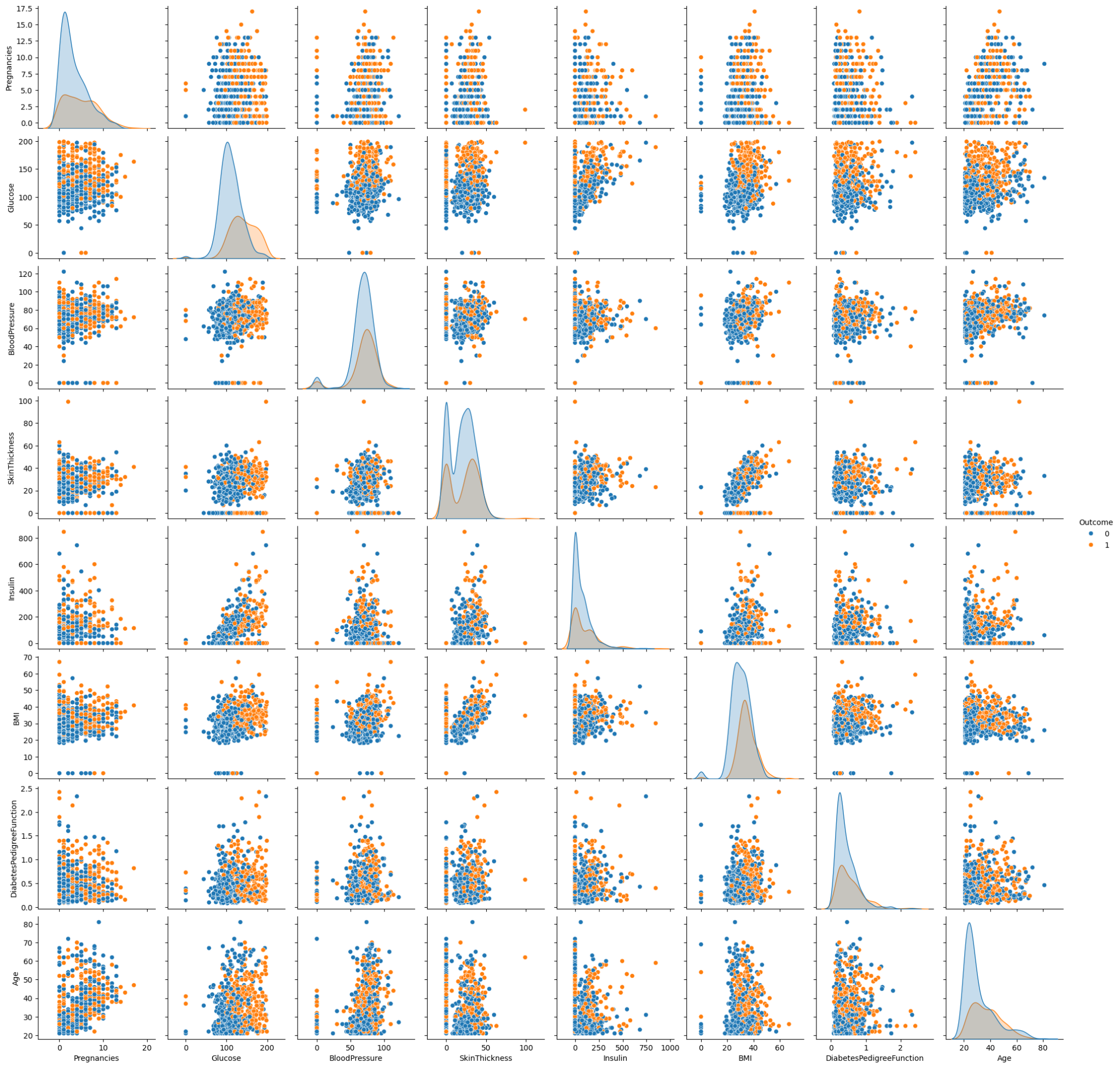
**5.2 Exploratory Data Analytics:**

## **Fig. 4 Relationship between different parameters and chances of having diabetes**



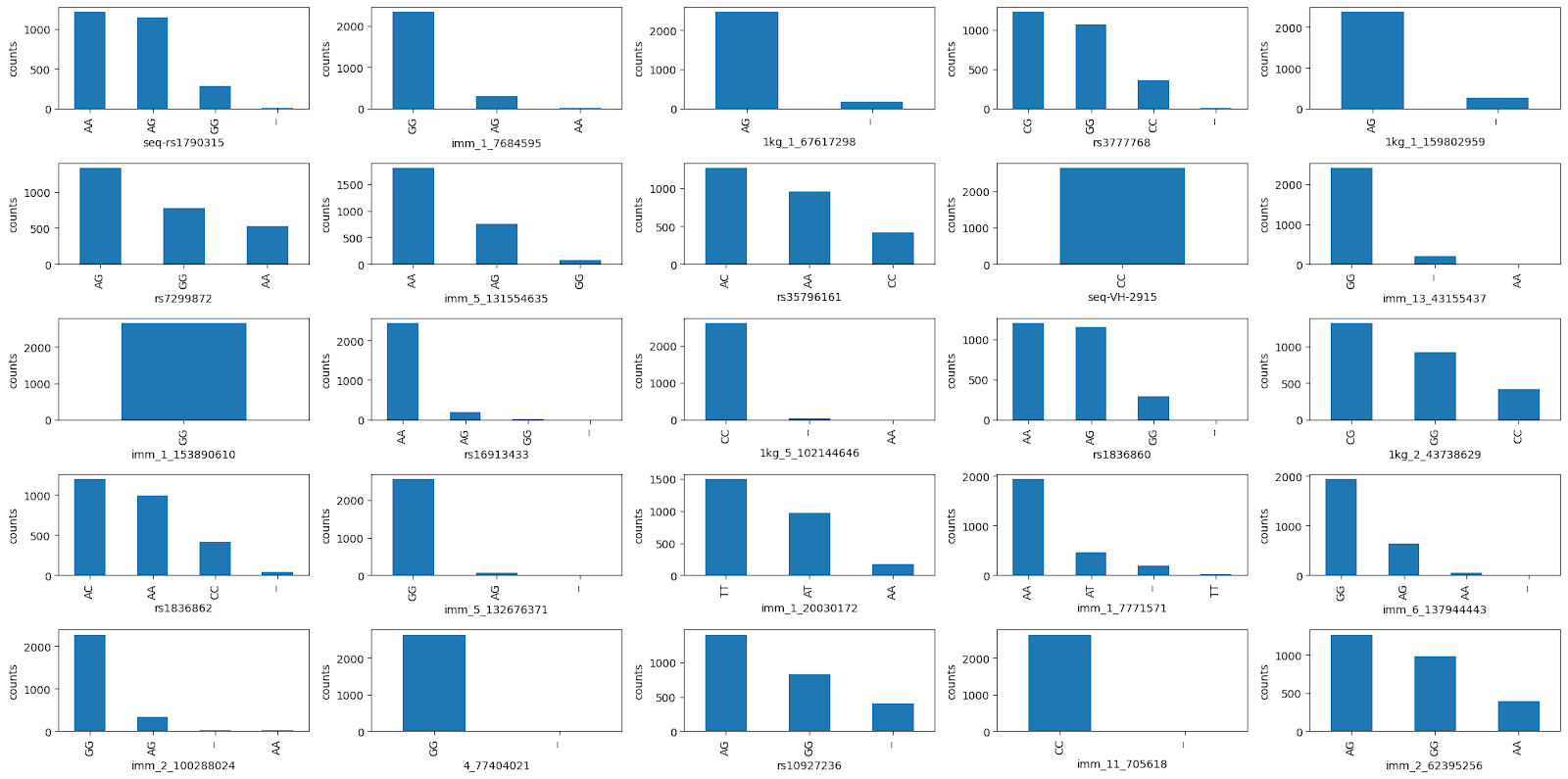
This is a density plots graphs where each plot represents the distribution of a specific variable, divided into two categories, labeled as "Pos" (positive) and "Neg" (negative), which likely correspond to different outcomes such as the presence or absence of diabetes.

## **Fig.5 Relationship between each pair of parameters**

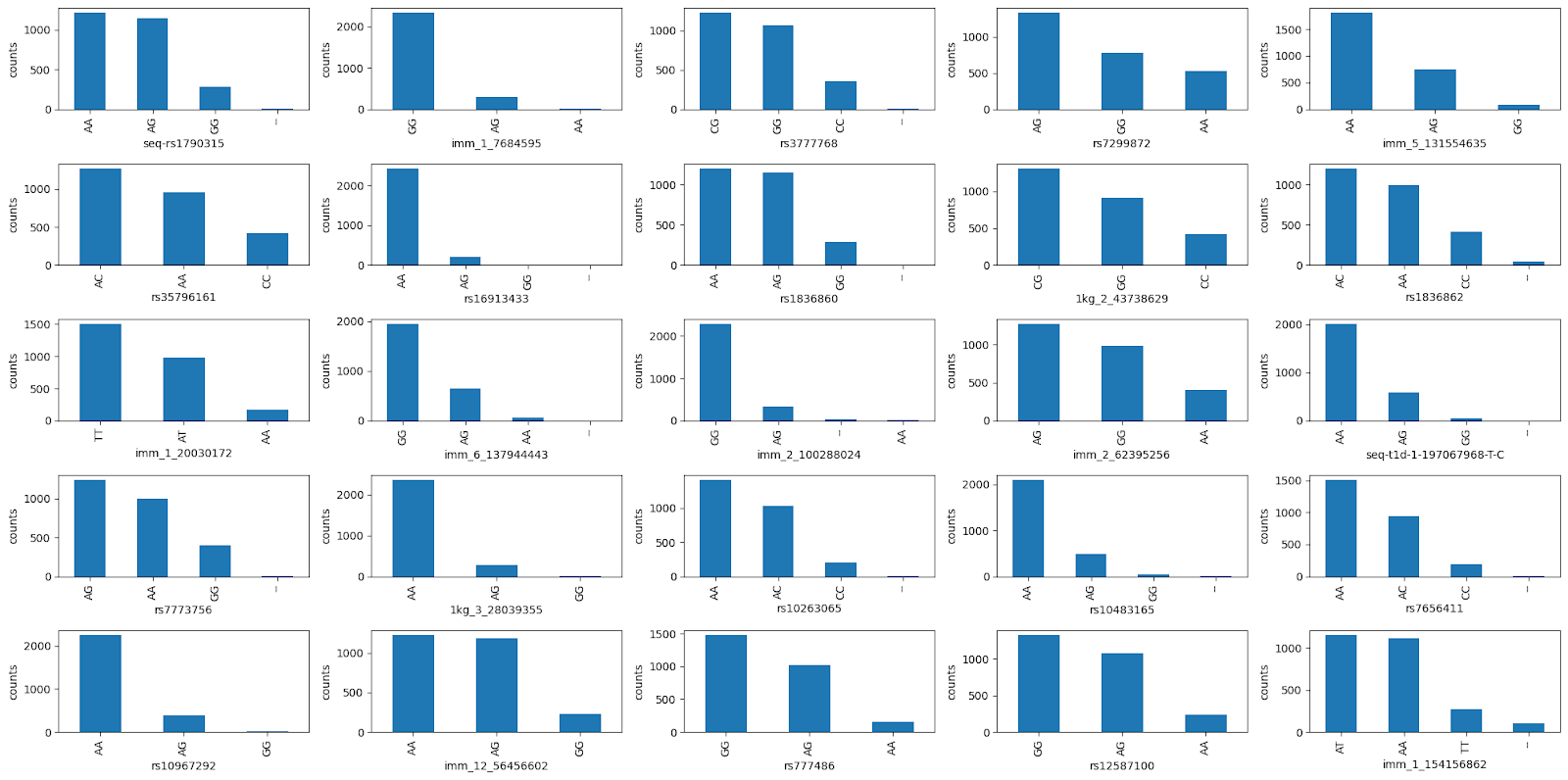


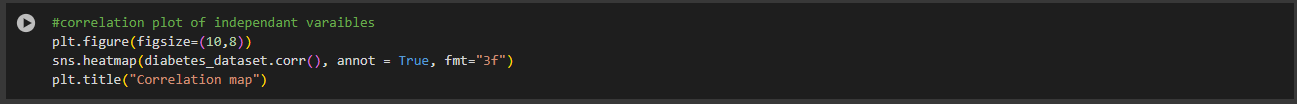
This is an output which shows the scatter plot matrix (or pair plot) consisting of multiple scatter plots and histograms. Each scatter plot represents the relationship between two of the variables in the dataset, while each histogram represents the distribution of a variable.

## **Fig.6 Genotype density distribution for first 25 sample patients**



## **Fig.7 Distribution of SNPs after cleaning.**



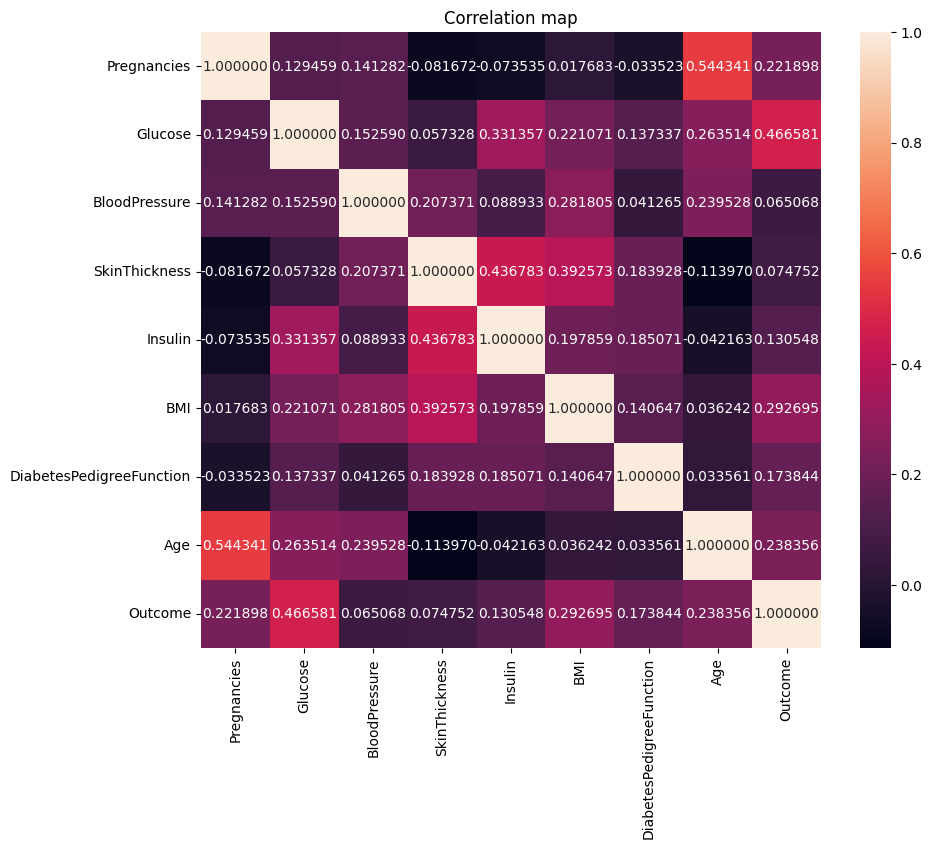


In this code, it is designed to create a heatmap for visualizing the correlation between variables in a dataset using the matplotlib library.

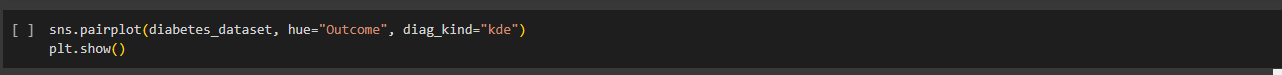


This code visualizes the distributions of selected features in a dataset, distinguishing between positive ('Pos') and negative ('Neg') outcomes. It utilizes Matplotlib to create a figure with subplots, each displaying Kernel Density Estimate (KDE) plots for specific columns along with the 'Outcome' column.

**5.3 Technical Coding and Code Solution:**

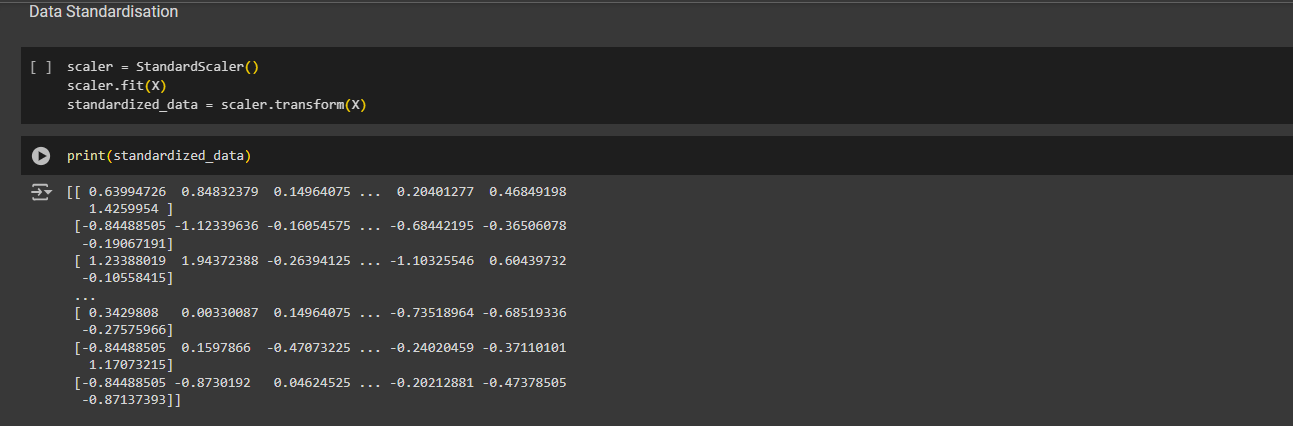


**Fig.8 This is the correlation map which we get as an output of variables in a dataset.**



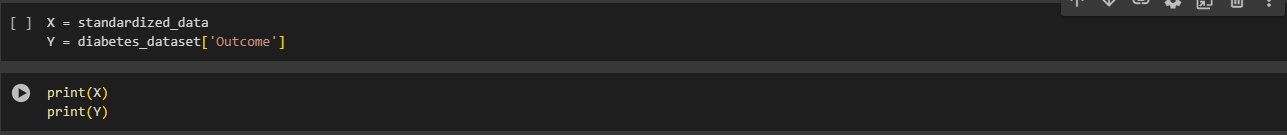
This code is using the Matplotlib library to create a simple plot.

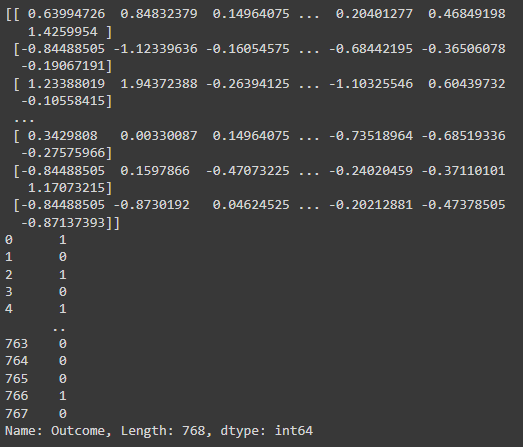
Now, in this output, it shows the Y values in the diabetes dataset.

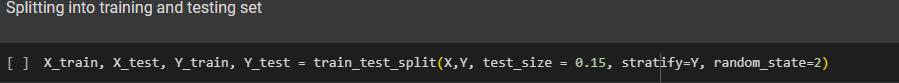


Here in this code, it shows the process of data preprocessing phase commonly used in machine learning workflows to normalize the features in a dataset.

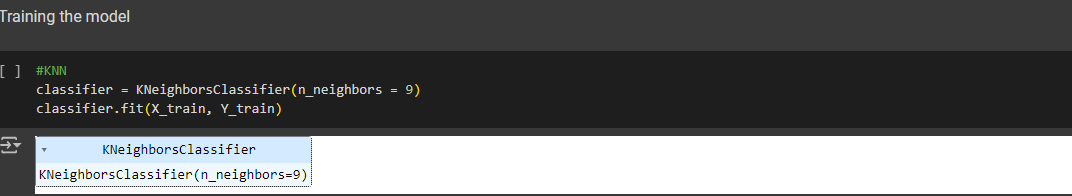
And the output displays the standardized features used to train the diabetes prediction model.



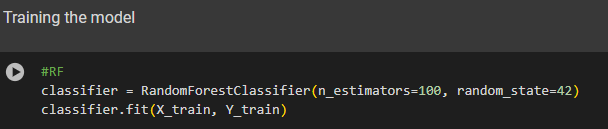




Now , in this code, we have splitted the dataset into training and testing dataset(Using KNN)

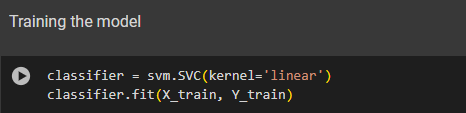


Splitting the dataset and training the dataset using Random Forest





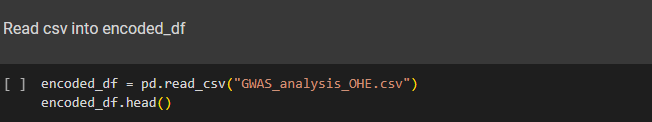
Splitting the dataset and training the model using SVM





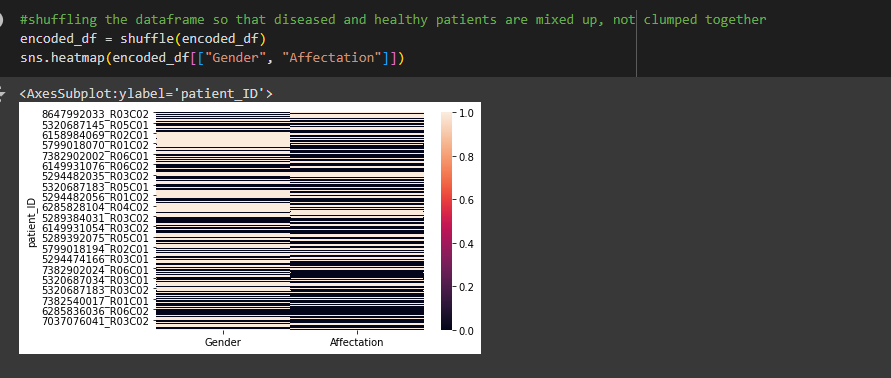
**Now, Technical code for IBD(Crohn’s Disease):**

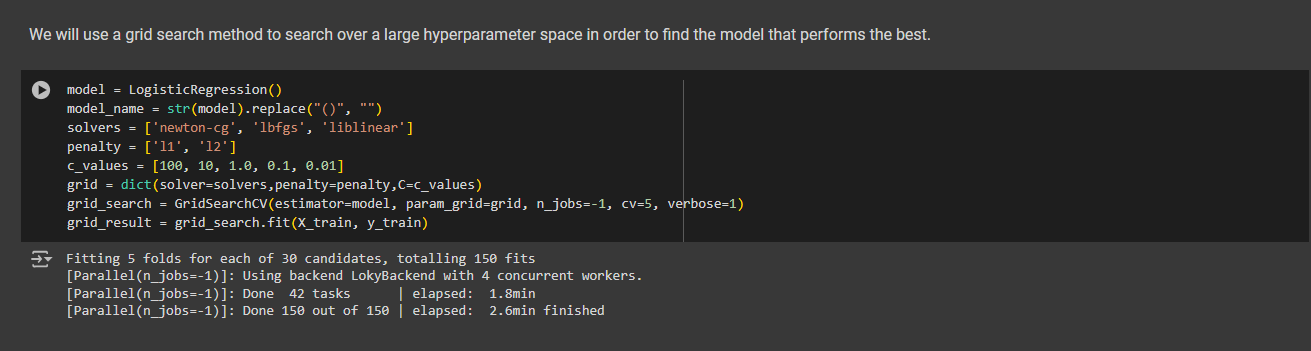
**IBD(Logistic Regression)**



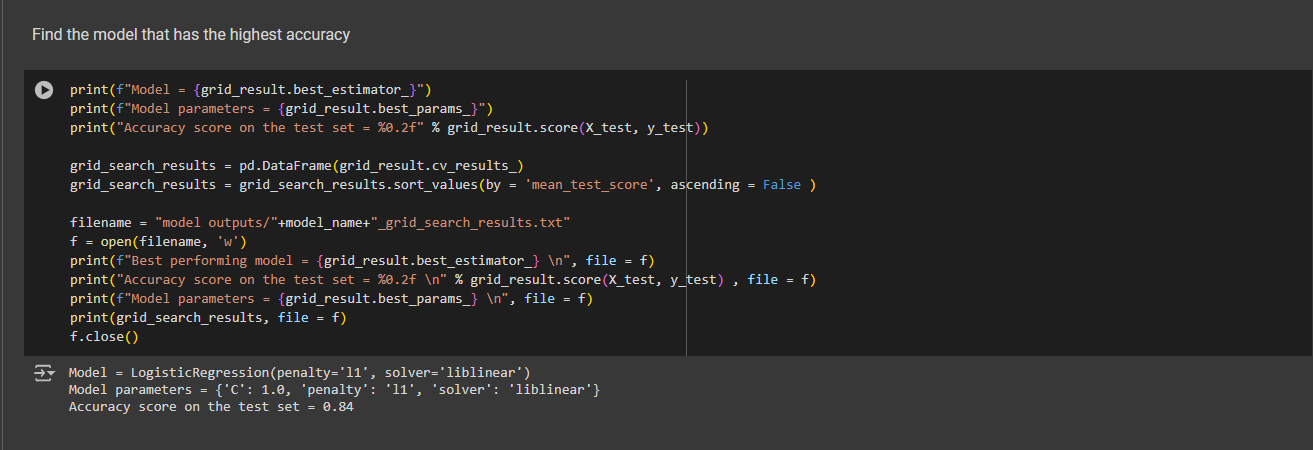
Here, we have imported necessary packages which makes our code functional and efficient.

And in the next part of the code, we read the csv file of GWAS Analysis.

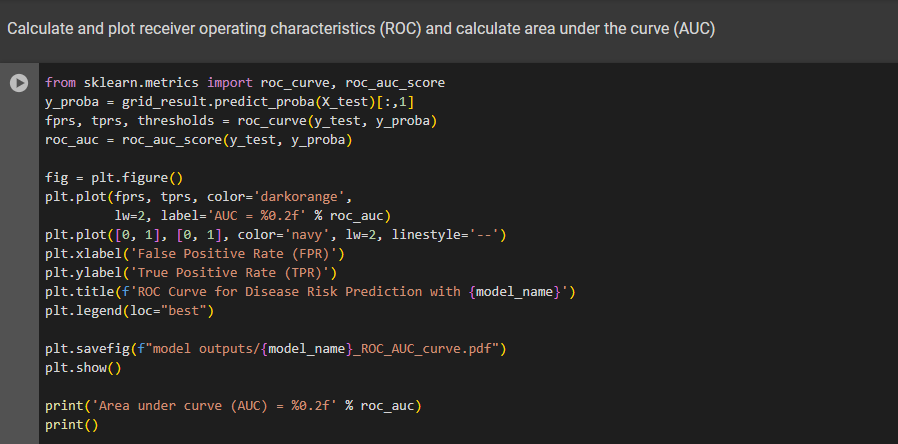




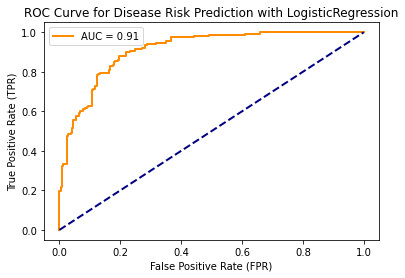
Now, using the grid search method to search a large hyperparameter space to see which model performs the best.



Here, we found the model with the highest accuracy.



Here is the code for plotting the graph of ROC Curve(Receiver Operating Curve) and AUC Curve(Area Under the Curve).





**Fig. 9 ROC and AUC Curve.**

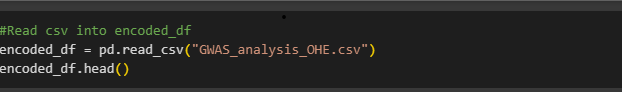


So, this is the confusion matrix generated.

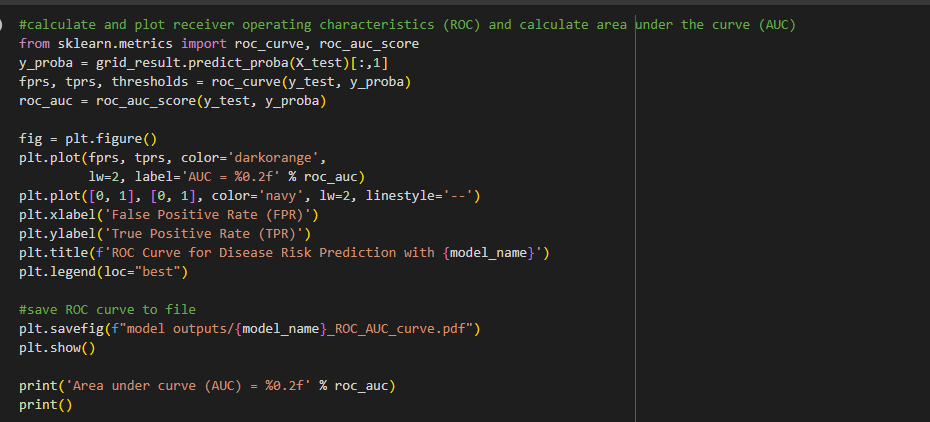


So, this is the precision, recall, F1-score and support values.

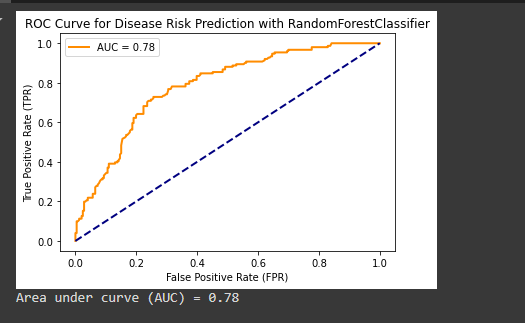
**IBD(Random Forest):**

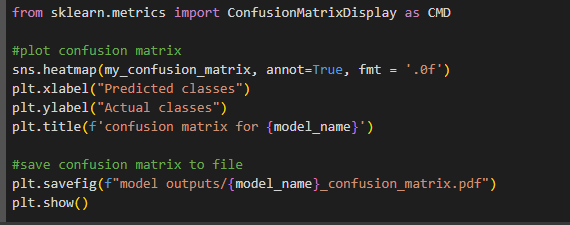


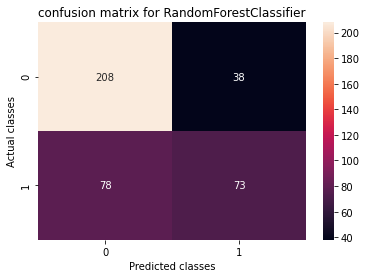




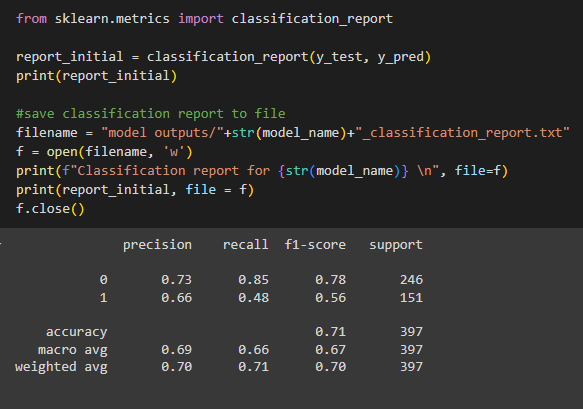
This is the code for ROC and AUC curve





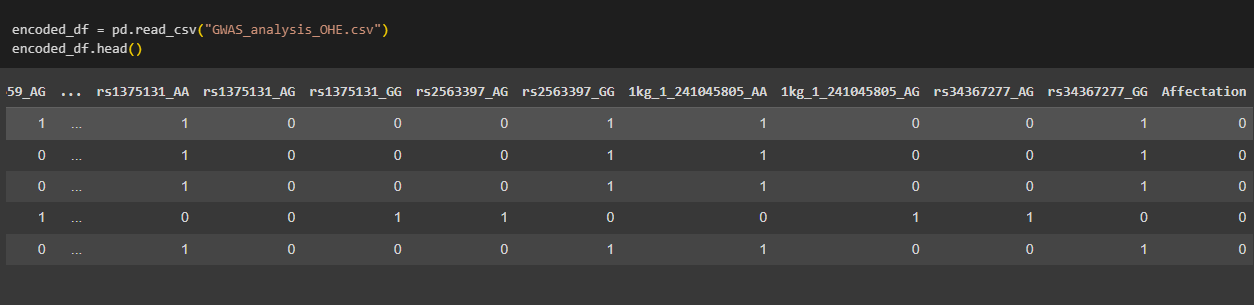


**Fig.10 Confusion matrix for Random Forest Classifier**

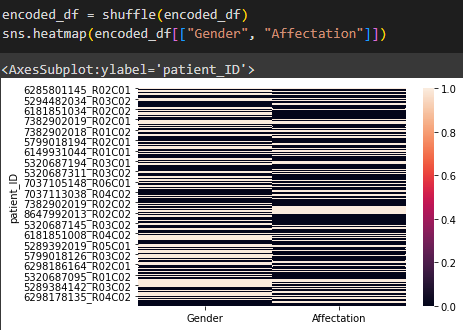


This is the precision, recall, f1-score, support

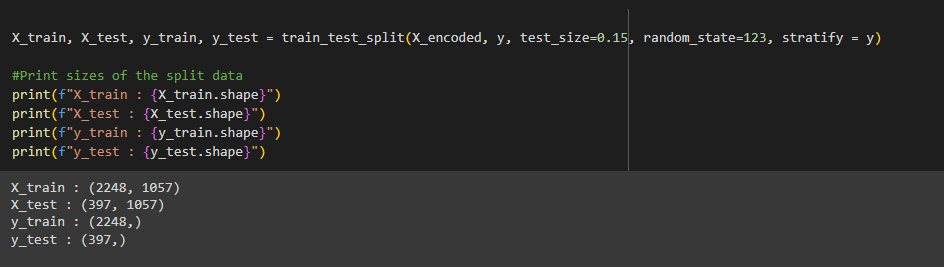
**Technical Code of IBD(SVM):**

****

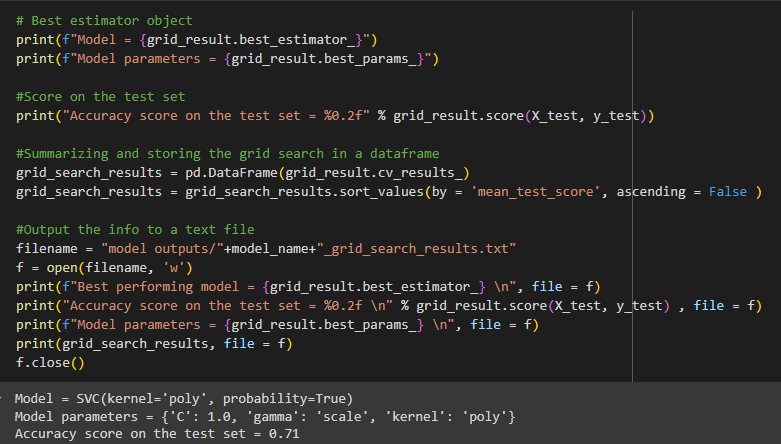
Here, in this code we import our dataset csv file into the code and display the output as shown.



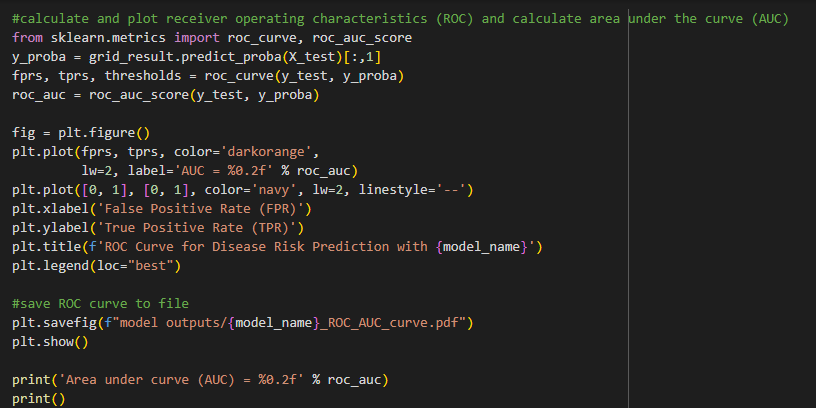
This output is a combined visualization featuring a barcode on the left and a heatmap on the right, which is used to analyze data distribution or relationships between variables.



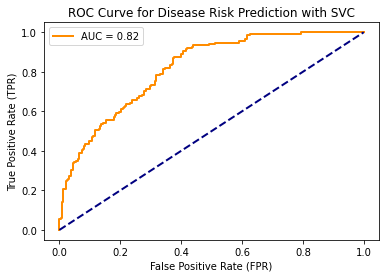
Here, in this code, we will split our dataset into training and testing dataset, and we will validate it, the output is shown here.



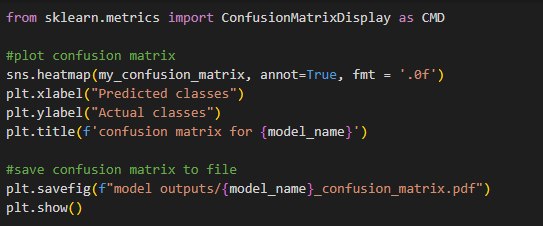
Here, we have found the best estimator which is SVC which has an accuracy score of 0.71.



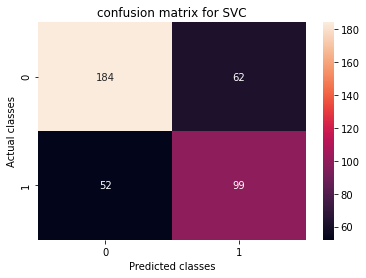
So this is the code for the ROC and AUC Curve.



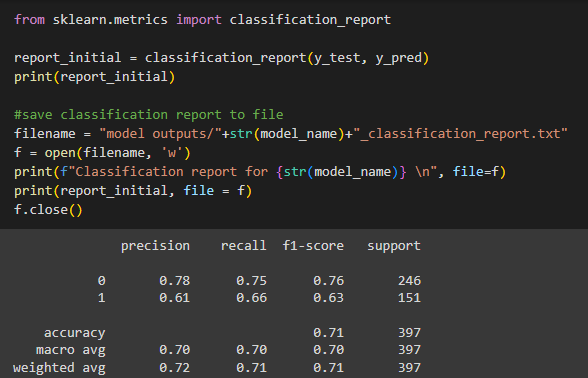
**Fig.11 ROC and AUC Curve for Disease Risk Prediction with SVC**



So, here in this code, we have to create a confusion matrix.



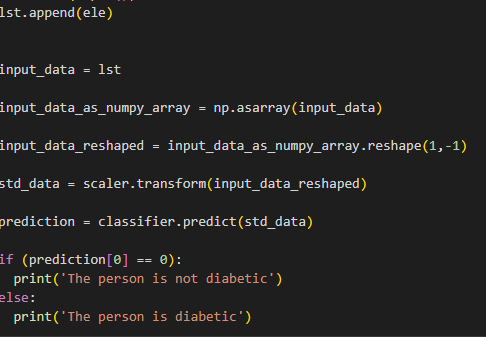
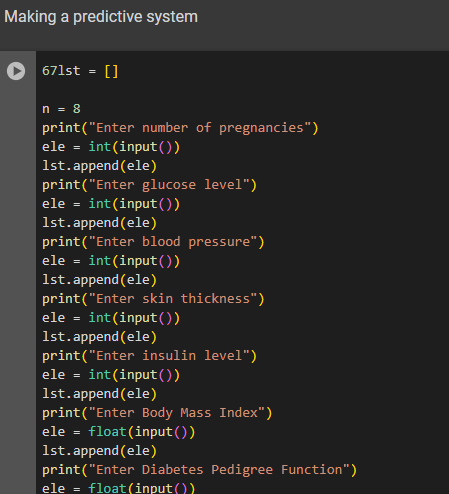
This is the confusion matrix which we have got.

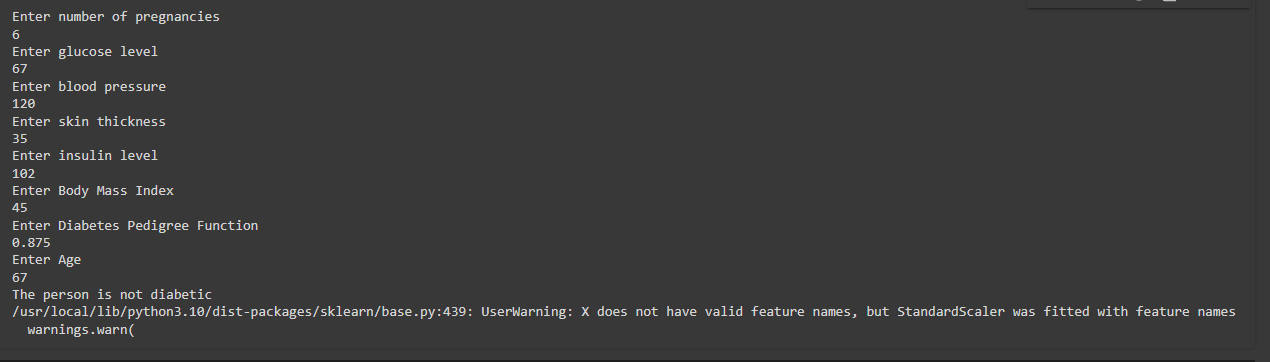


Here, we got the precision, recall, f1-score and support values.

**5.4 User Interaction Prototype**

So, here we tried to make a predictive system, so the code for it is given below:





This is the output of the predictive system.

|  |
| --- |

**5.5 Testing and Validation:**

**Fig. 1 Testing and Validation Table for Different classifiers for Type-II Diabetes and IB Disease**

| **Disease** | **Algorithm** | **Testing Data Percentage** | **Training Data Percentage** |
| --- | --- | --- | --- |
| Diabetes Prediction | KNN(K-Nearest Neighbors) | 15% | 85% |
|  | Random Forest | 20% | 80% |
|  | SVM (Support Vector Machine) | 20% | 80% |
| Crohn’s Disease (IBD) | KNN(K-Nearest Neighbors) | 15% | 85% |
|  | Random Forest | 15% | 85% |
|  | SVM (Support Vector Machine) | 15% | 85% |

**5.6 Performance Analysis(Graphs/Charts) Fig.2 Performance analysis Table**

| T2D | IBD |
| --- | --- |
| For KNN |  |
| For Random forest |  |
| SVM: |  |

**Summary:**

Testing and validation are critical phases in the development of machine learning models, ensuring their effectiveness and reliability. In the context of disease prediction, such as for diabetes and Crohn's Disease, different algorithms like KNN, Random Forest, and SVM are employed, each with varying percentages of testing and training data. By systematically validating these models, researchers can determine their accuracy and suitability for real-world applications, ultimately improving diagnostic and prognostic capabilities in healthcare.

**CHAPTER-6**

**PROJECT OUTCOMES AND AVAILABILITY**

**6.1 Outline:**

Comparing the performance of different algorithms is crucial in machine learning for several reasons:

(i) Optimization of Performance: Comparing algorithms allows researchers to identify the most effective approach for a given task. By determining which algorithm yields the best results, they can optimize performance and achieve higher accuracy in predictions.

(ii) Understanding Algorithm Behavior: Each algorithm has its strengths and weaknesses. Through comparison, researchers gain insights into how different algorithms behave under various conditions, such as with different types of data or problem domains.

(iii) Resource Efficiency: Comparing algorithms helps in selecting the most resource-efficient option. Some algorithms may require less computational power or memory, making them more suitable for deployment in resource-constrained environments.

|  |
| --- |

**6.2 Significant Project Outcomes: Fig.3 Project outcome table of Diabetes and IBD**

|  | **Outcome0** | **Outcome1** |
| --- | --- | --- |
| **Type II Diabetes** |  |  |
| **IBD Disease(Crohn’s Disease)** |  |  |

**6.3 Project Applicability on Real World Applications:**

Implementing machine learning (ML) disease prediction models in India to tackle real-world problems in real-time can have significant benefits in healthcare. Here's how it can be done effectively:

**(i)** **Data Collection and Integration:** Collaborate with healthcare institutions and organizations across India to gather diverse and comprehensive datasets encompassing various diseases prevalent in different regions. Integration of electronic health records (EHRs), diagnostic reports, and demographic data can provide a holistic view of patients' health status.

**(ii) Mobile Health (mHealth) Solutions:** Utilize mobile applications and wearable devices to collect real-time health data from individuals. These data streams can include vital signs, activity levels, dietary habits, and symptoms, which can be integrated into ML models for continuous monitoring and early detection of diseases.

**(iii) Telemedicine and Remote Healthcare:** Integrate ML disease prediction models into telemedicine platforms to enable remote consultations and diagnostics. This approach can improve access to healthcare services, especially in rural and underserved areas where there is a shortage of healthcare professionals.

**Inference:**

The implementation of machine learning (ML) disease prediction models in India has yielded significant project outcomes with far-reaching implications for healthcare. By harnessing the power of ML algorithms, early detection and diagnosis of diseases have vastly improved, leading to timely interventions and better treatment outcomes. This has been particularly impactful in extending healthcare access to underserved populations, including rural and remote areas, through integration with telemedicine platforms and community health initiatives. Moreover, the localization of ML models has enabled the development of tailored healthcare solutions that account for regional variations in disease prevalence and socio-economic factors.

**CHAPTER-7**

**CONCLUSIONS AND RECOMMENDATION**

**7.1 Outline:**

This study delved into the genetic basis of complex diseases, specifically examining the genetic architecture of early-onset (EO) versions of Inflammatory Bowel Disease (IBD) and Type 2 Diabetes (T2D) in comparison to their adult-onset counterparts within European populations. Immunochip genotyping was utilized on EO-IBD cases and controls, revealing some notable insights. While known genes like NOD2 and IL23R were replicated for EO-IBD, the presence of loci previously associated with adult-onset IBD was limited, indicating potential differences in genetic susceptibility between early and adult-onset forms. A similar analysis is planned for T2D to further understand the genetic underpinnings of early-onset

versus adult-onset forms. Additionally, the study will explore polygenic risk scores to predict age of onset for both diseases.

**7.2 Limitations/Constraints of the System:**

Despite the valuable insights gained, this study has its limitations. One major constraint is the sample size, particularly for EO-IBD, which may have impacted the ability to detect all relevant genetic factors. Additionally, the focus on European populations may limit the generalizability of the findings to other ethnic groups. Furthermore, the use of polygenic risk scores, while promising, may not fully capture the complexity of genetic predisposition to these diseases.

**7.3 Future Enhancements:**

To address the limitations, future research should aim to expand the sample size, possibly through multi-center collaborations or meta-analyses, to improve the statistical power and generalizability of the findings. Moreover, including diverse populations will provide a more comprehensive understanding of the genetic architecture across different ethnicities. Additionally, integrating other omics data, such as transcriptomics and epigenetics, could provide further insights into the molecular mechanisms underlying EO-IBD and EO-T2D.

**7.4 Inference:**

In conclusion, while this study provides valuable insights into the genetic underpinnings of EO-IBD and EO-T2D, further research is warranted to elucidate the full spectrum of genetic factors involved in these diseases. By addressing the limitations and incorporating diverse approaches, future studies can contribute to a better understanding of the complex interplay between genetics and early-onset disease pathogenesis, ultimately informing more effective prevention and treatment strategies.

**REFERENCES**

1. N. Chatterjee, J. Shi, and M. Garcia-Closas, "Developing and evaluating polygenic risk prediction models for stratified disease prevention," Nat. Rev. Genet., advance online publication, 2016. [Online]. Available: https://doi.org/10.1038/nrg.2016.27. .
2. M. I. McCarthy et al., "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," Nature reviews genetics, vol. 9, no. 5, pp. 356–369, 2022. [Online]. Available: https://doi.org/10.1038/nrg2344. .
3. Z. Wei et al., "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," Am. J. Hum. Genet., vol. 92, no. 6, pp. 1008–1012, 2013. [Online]. Available: https://doi.org/10.1016/j.ajhg.2013.04.014. .
4. D. J. Cutler et al., "Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping," PLOS ONE, vol. 10, no. 6, p. e0128074, 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0128074. .
5. X. Zhou, P. Carbonetto, and M. Stephens, "Polygenic modeling with Bayesian sparse linear mixed models," PLoS Genet., vol. 9, no. 2, p. e1003264, 2013. [Online]. Available: https://doi.org/10.1371/journal.pgen.1003264. .
6. D. Speed and D. J. Balding, "MultiBLUP: improved SNP-based prediction for complex traits," Genome Res., vol. 24, no. 9, pp. 1550–1557, 2014. [Online]. Available: https://doi.org/10.1101/gr.169375.113. .
7. J. Minnier et al., "Risk classification with an adaptive naive Bayes kernel machine model," J. Am. Stat. Assoc., vol. 110, no. 509, pp. 393–404, 2015. [Online]. Available: https://doi.org/10.1080/01621459.2014.948195. .
8. N. Nguyen et al., "Machine learning-based prediction models for diagnosis and prognosis in inflammatory bowel diseases: a systematic review," J. Crohn's Colitis, vol. 16, no. 3, pp. 398–413, 2022. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjab134. .
9. S. Kraszewski et al., "Machine learning prediction model for inflammatory bowel disease based on laboratory markers. Working model in a discovery cohort study," J. Clin. Med., vol. 10, no. 20, p. 4745, 2021. [Online]. Available: https://doi.org/10.3390/jcm10204745. .
10. M. Hasan et al., "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2984985. .
11. J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, 2021. [Online]. Available: https://doi.org/10.1016/j.icte.2021.03.013. .