University of Sunderland

Data Science Product Development

Exploring Diabetes Risk Factors: An In-depth Analysis and Prediction

A special thank to my professors for their invaluable guidance and support.

- **Dr Ashley Williamson**
- **Dr Ming Jiang**
- **Dr Ron Mo**

IBOURK Youssef

Table of Contents:

1. INTRODUCTION :

Diabetes is a long-standing defect and disorder that occurs as a result of a metabolic malfunction in carbohydrate metabolism, so it is a serious global health problem. Overall, early detection of diabetes can have a significant impact on the treatment of diabetic patients, leading to the elimination of its corresponding side effects.

Machine learning is a new technology that provides a high prognosis and deeper understanding of various groups of diseases, such as diabetes. And since there are not enough effective analysis tools to detect hidden relationships and trends in data, health information technologies have become a new technology in the health sector in a short period of time thanks to the use of data science, which is data-driven decision support system.

2. Dataset:

By 2040, researchers and statisticians expect that about 642 million adults will have diabetes. In addition, 46.5% of these adults were not diagnosed with diabetes. To reduce such a large number of diabetes-related deaths, it is important to develop many best practices and techniques that will help effectively diagnose diabetes at an early stage, since a large number of deaths among diabetic

patients are the result of late diagnosis of diabetes. To develop and implement advanced methods for early diagnosis of diabetes, we need to make extensive use of sophisticated information technology solutions, and data science, which are the right tools for such situations.

Data science and Machine learning techniques play a crucial role in the medical and healthcare sector due to the fact that data science is considered a broad category of methodologies, solutions and applications for capturing, collecting, maintaining, analysing and providing easy access to data to help users make successful and quick decisions. It also includes various actions and functions of decision support systems, such as queries and reports, online analytical processing, statistical analysis, forecasting, and data and text analysis.

And since the selection of a suitable data source and theme is crucial to the success of any data science product, the choice of data source should be based on its relevance to the problem at hand and the availability of data. The dataset used is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. And based on certain diagnostic measurements included in the dataset, the objective of the dataset is to diagnostically predict whether or not a patient has diabetes. Several constraints were placed on the selection of these instances from a larger database.

The dataset was obtained from Kaggle DATASET. There are a total of 768 observations and 9 variables in the dataset:

Pregnancies	Number of times pregnant
Glucose	The concentration of glucose in the blood plasma after a 2-hour oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm / HG)
SkinThickness	Triceps skinfold thickness (mm)
Insulin	2 hours of serum insulin (mu / ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	A function that determines the risk of developing diabetes based on a family history. The greater the function, the higher the risk of developing type 2 diabetes.
Age	Age in years
Outcome	Whether a person has been diagnosed with type 2 diabetes (1 = yes, 0 = no)

Figure 1: Dataset variables description

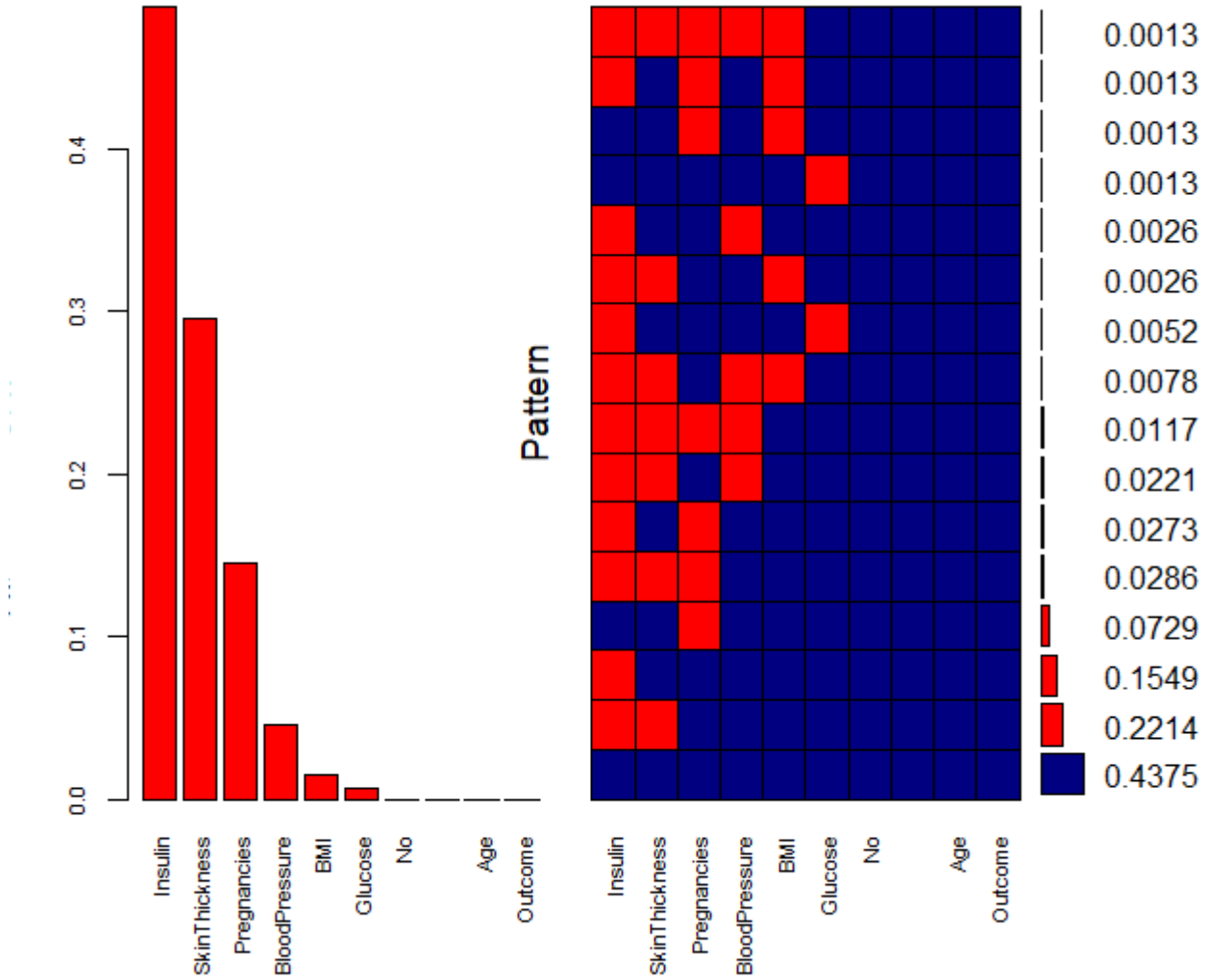

Figure 2: Histogram of missing data

Figure 3: visualization of the missing data patterns

3. Product Design:

To develop a successful data science product, it is essential to understand the needs and requirements of the end user, we first identified the target audience for our data science product and conducted extensive research on this field. In this case, the target audience was healthcare professionals and researchers who are interested in predicting the onset of diabetes based on patient data. I conducted research about several healthcare professionals to understand their requirements and expectations from a diabetes prediction model. Through this research, I identified the following requirements:

- **Accuracy:** To make informed decisions, healthcare professionals require accurate predictions. As a result, the model must be highly accurate.
- **Explainability:** To offer proper patient care, healthcare providers must understand how the model makes its predictions. As a result, the model must be explainable and produce clear and interpretable outcomes.
- **Ease of use:** Because the model's users may not be data science specialists, the model must be simple to use and have a user-friendly interface.

- **Robustness:** The model should be able to handle a broad variety of patient data while still producing accurate predictions.

Based on an examination of the data source and the application domain here are some functional and non-functional requirements for the product. Data cleaning, data preparation, model training, and prediction are among the functional requirements. Non-functional requirements included system performance, scalability, and dependability. We utilized these requirements to drive the design of the software architecture.

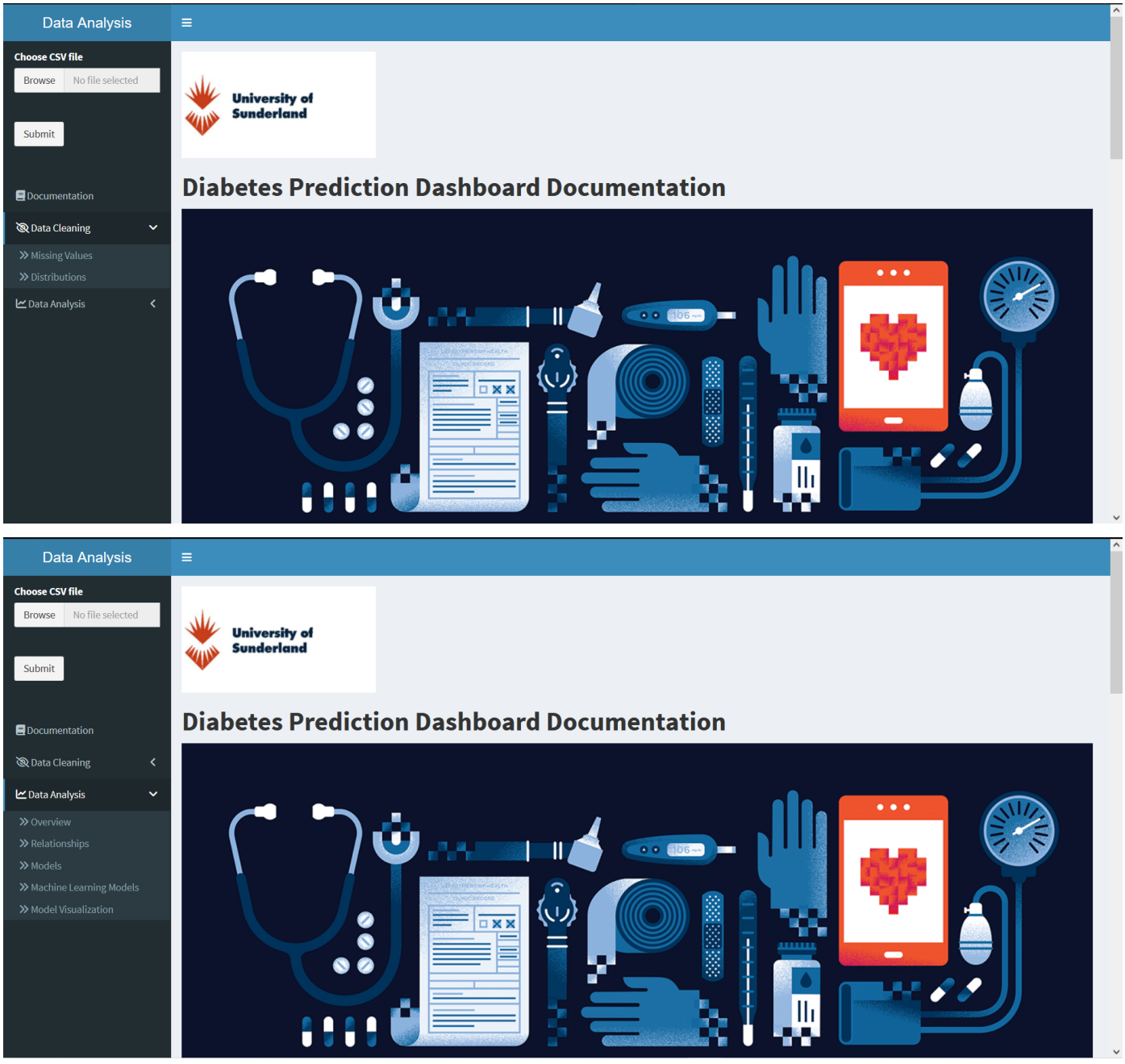

Figure 4: Dashboard design

The architecture of a data science product refers to the high-level structure of the system and how its components work together to achieve the desired functionality. For our data science project, we designed a software architecture that would allow us to process, clean, analyze, and visualize the

Dataset. The architecture consists of three main layers: documentation, data cleaning and data analysis.

The documentation layer is essential for promoting effective usage of the software, facilitating development, enabling support and maintenance, and ensuring a smooth user experience. The data cleaning layer is responsible for acquiring and cleaning the raw dataset. This layer includes sub-components for data cleaning, feature engineering, and data transformation. The data analysis layer is responsible for the statistical analysis and modelling of the cleaned dataset. This layer includes sub-components for exploratory data analysis, feature selection, and model training. Finally, the data cleaning and data analysis layers contains a data visualization which is responsible for presenting the results of the analysis in a meaningful and easy-to-understand format it includes sub-components for data visualization.

4. Product Development:

Any data science project requires the selection of appropriate software tools/platforms and hardware approaches. The appropriate combination of hardware and software tools can have a major influence on the system's performance, scalability, and adaptability.

We picked RStudio as our Integrated Development Environment for our project because of its simplicity of use, interactive console, project management, and data analysis.

Furthermore, this project's code goes through a variety of data processing and visualization processes. The server function is defined first, along with its inputs and outputs. When the submit button is pressed, the observeEvent method is called to activate the code.

Next, the code reads a CSV file specified by the user and assigns it to the variable "data". In this process, any zero values in columns 2 to 9 are replaced with NA. Additionally.

The summary function is then used to compute summary statistics for the data, which are saved in the variable "summary_data." The resultant summary table is displayed as an output.

The code also computes the percentage of missing values in each column and row of data, which is saved in the variables "col_missing" and "row_missing" accordingly. Various plots related to visualizing missing data are generated using different functions, and they are rendered as outputs for visualization.

In addition to the missing data plots, the gg_miss_var function generates and renders a particular plot named "vim_plot2" as an output.

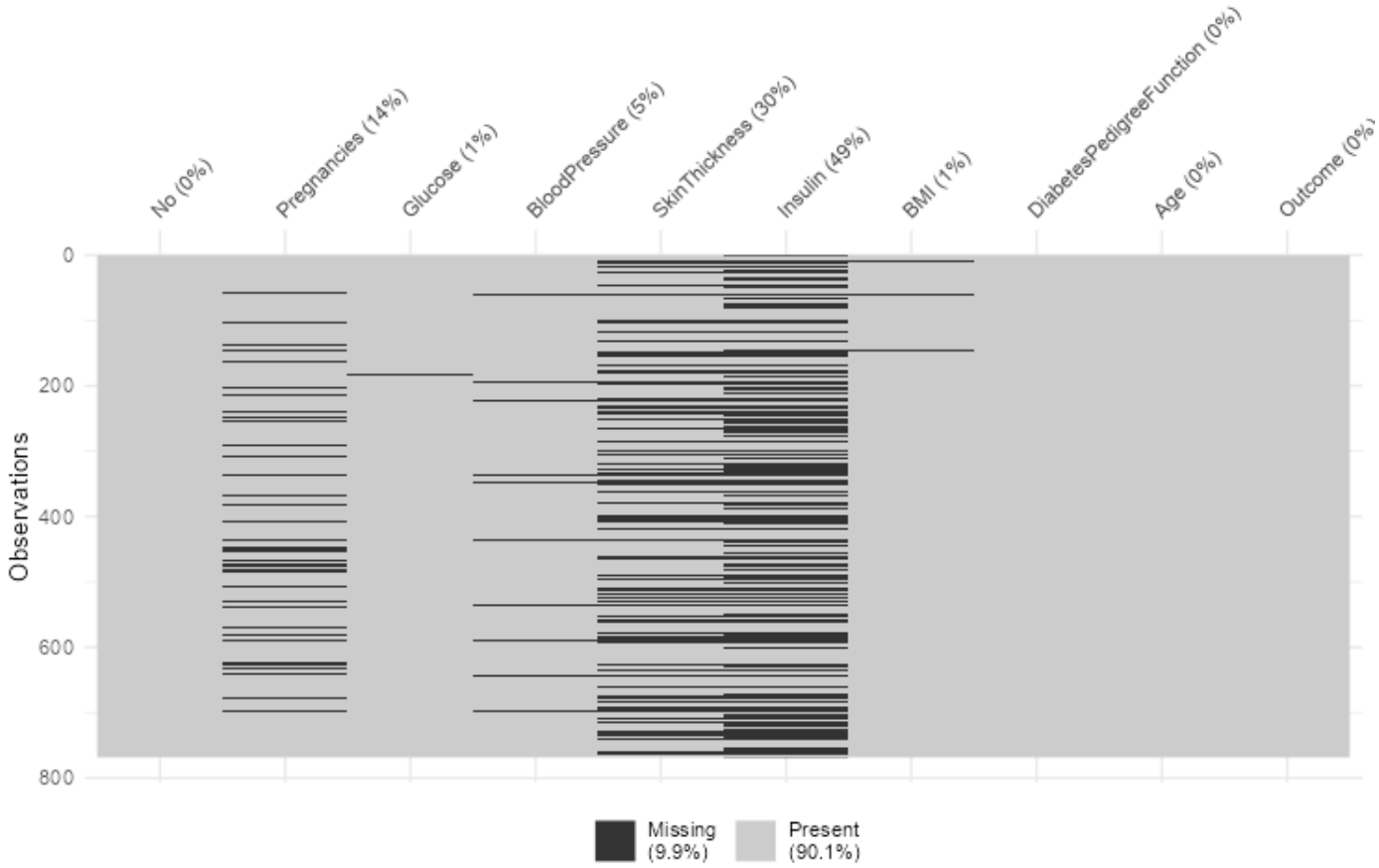

Figure 5: missing data pattern

The marginplot function is used to construct several plots, especially “multi_vim_plot1” to “multi_vim_plot8” to compare variables with the first variable in the dataset.

To fix missing data, the mice function is used to do multiple imputations on the dataset, and the imputed data is saved in the variable “tempData”. And to indicate the variables that require imputation, a vector named “vars_to_impute” is defined.

A for loop iterates over the variables specified in “vars_to_impute” and accesses the imputations stored in “tempData$imp”. However, the results of this loop are not explicitly assigned or saved.

To ensure consistency, the “Outcome” column is transferred from the original data to “impdata”.

Furthermore, the “impdata” dataframe is categorized by “Age” in order to compute the mean, median, and standard deviation of the variables for each age group. The computed findings are presented as different outputs to give insights into the correlations between factors and age.

Boxplots are constructed to depict the distributions of individual variables with respect to “Age” to further investigate these correlations. These boxplots are displayed as outputs for visual inspection.

Overall, these steps encompass data processing, missing data imputation, summary statistics computation, missing data visualization, result category analysis, and variable association research, all of which provide useful insights into the dataset.

Following these steps, the code does further analysis such as histogram plots, boxplots, outlier detection and removal, PCA, K-means clustering, data splitting, standardization, SOM training, and SOM training visualization. Other machine learning techniques, such as random forest, rpart tree, K Nearest Neighbors (KNN), Logistic Regression, and Linear Discriminant Analysis (LDA), are then implemented in the code.

In terms of hardware, I have used my personal computer with the recommended specifications for running RStudio and Shiny. The selection of appropriate software tools and platforms is important to ensure efficient development and implementation of the data science product.

In summary, the selection of appropriate software tools, platforms and hardware methodologies is critical to maximize performance, scalability, and flexibility while ensuring reproducibility and ease of use.

5. System Testing Methods :

To guarantee that the product satisfied functional and non-functional criteria, we employed a combination of manual and automated testing. We created test cases for each use scenario, which were then manually run. We also made advantage of automated testing technologies.

In the context of our project, **unit testing** involved testing specific functions and algorithms that are implemented. **Integration Testing,** we tested the integration between the data pre-processing module, machine learning model module, and the user interface module. **Functional Testing** involved validating that the data import, pre-processing, visualization, model building, evaluation, selection, and deployment functionalities are functioning correctly. In the **performance testing** we evaluated the system's response time and resource usage during data processing, model training, and prediction tasks. And as for the **Regression Testing**, it was necessary when modifying or adding new features to the data pre-processing, machine learning model, or user interface.

6. Project Management:

A Gantt chart is an efficient approach to organize and schedule jobs throughout the project lifetime. That is why we will first decompose each of the major tasks in our project into sub-tasks,

identifying deliverables and deadlines.

Task Id	Task Name	Objective	Hours	Deliverable
1	Data Collection and Preparation	Gathering relevant data for analysis and pre-processing it for modelling	1 week	Collected and organized dataset and a pre-processed dataset
2	Exploratory Data Analysis	Exploring and understanding the dataset to gain insights about the data's characteristics.	1 week	Summary statistics of variables, Data visualizations as plots and patterns and trends identification in the data.
3	Feature Engineering	Creating new features or transform the existing ones to enhance the power of the models	1 week	Feature selection and reduction of dimensionality techniques applied
4	Model Selection and Training	Evaluating different machine learning algorithms and select the best-performing ones	2 weeks	Trained models and model evaluation metrics for each model.
5	Model Deployment and User Interface Development	Creating a user-friendly interface to get predictions	2 weeks	Deployed model with an interactive user interface

Figure 6: schedule table

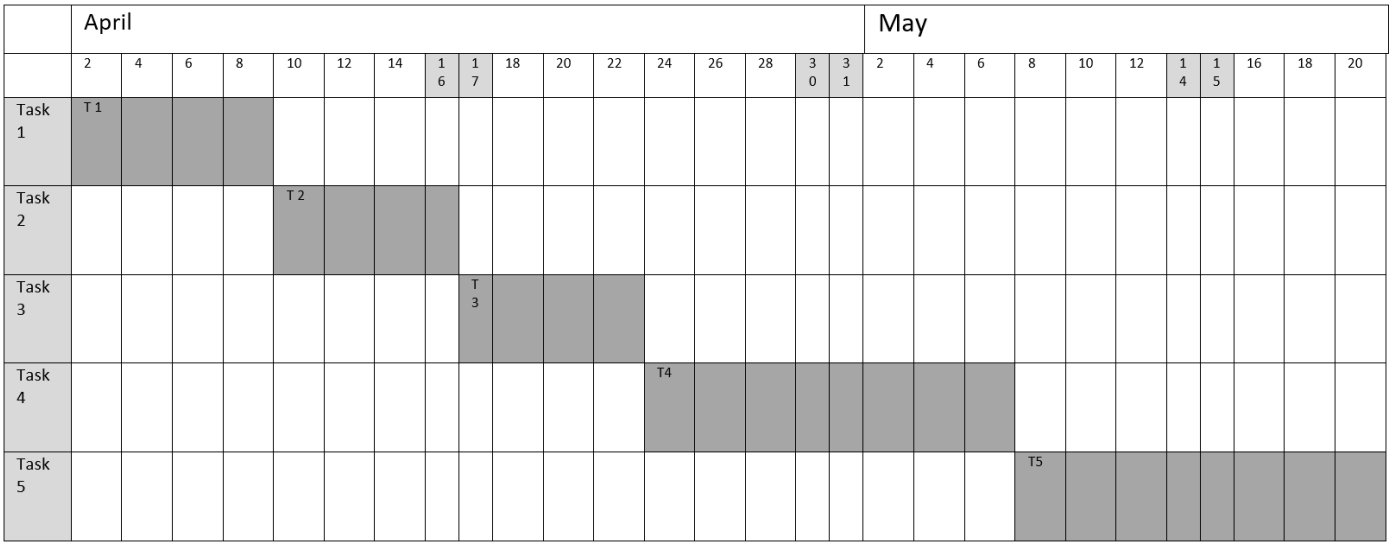

Figure 7: GANTT chart

Personal Information Protection And Data Security/Governance

Every data science product must consider personal information protection, data security, and data governance. It is critical to identify the risks connected with these regions and apply suitable risk-mitigation procedures. Data encryption, access limits, and regular security audits are examples of such methods. That is why we must **Identify Potential Risks** like unauthorized access to data, data breaches, loss of sensitive information. **Evaluating Likelihood** of each risk occurring in the product.

Data Protection Measures the RStudio offers for the files used, **HTTPS** and **SFTP** (SSH File Transfer Protocol) are two examples. These protocols encrypt communication between our browser and the RStudio server, ensuring that files are safely sent across the network. RStudio supports the encryption of files, which means that the files are kept in an encrypted manner on the underlying storage system. Encryption gives an extra degree of security, guaranteeing that even if unwanted access is gained to the storage, the contents remain encrypted and unreadable.

One more measure is files to be backed up on a regular basis. This can defend against data loss due to device failure, software difficulties, or human mistake by backing up information.

Some or all the above measures can be included as potential improvements and future research directions based on experiments and findings.

7. Results Inspection:

7.1 K-Nearest Neighbours:

For KNN, we calculate the result for each test case by comparing this case with the “nearest neighbours” in the training set. The assigned result depends on how many of these neighbours we choose to look at; the class of most of the three nearest neighbours may differ from the class of most of the five nearest neighbours.

To make sure that we use the number for which gives the best performance of the model, I performed a two-part cross-validation. First, I changed the possible values from 2 to 10; second, I repeated splitting the data into training and test sets 100 times to provide a reliable estimate of the model performance for each, I used `knn` function within the limits of `class` Accuracy of packing and calculate model on the test set for each fold.

The black line indicates the average of all 100 folds for each value of k.

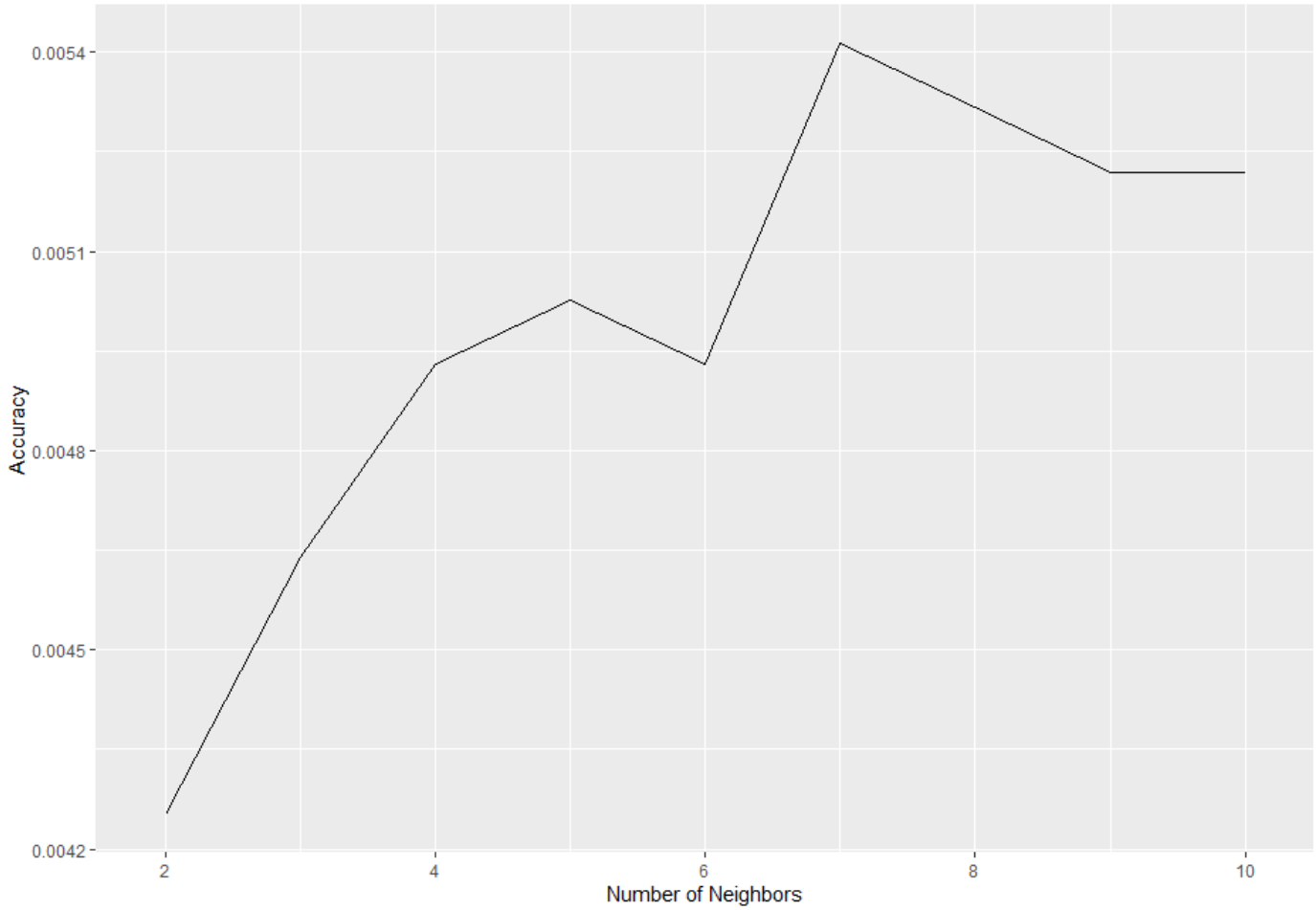

Figure 8: KNN accuracy vs number of neighbours

From this analysis, we can see that KNN perform better for slightly larger values_,_with a performance that reaches a maximum of about 71% classification accuracy. While there is still some difference depending on the exact separation of the data, using 9 or 10 neighbours seems to produce fairly stable model scores on the test set.

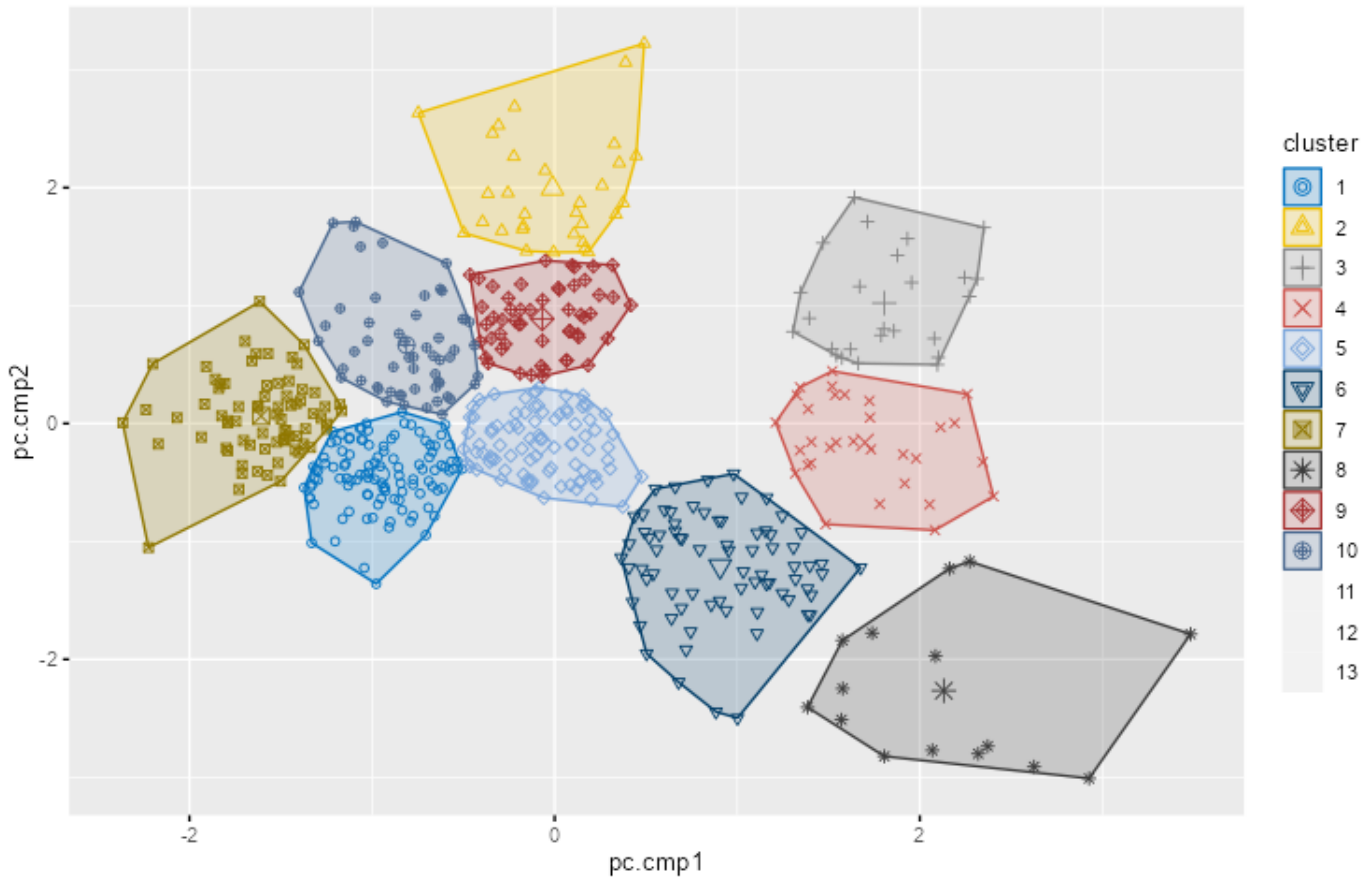

Figure 9: Clustering analysis

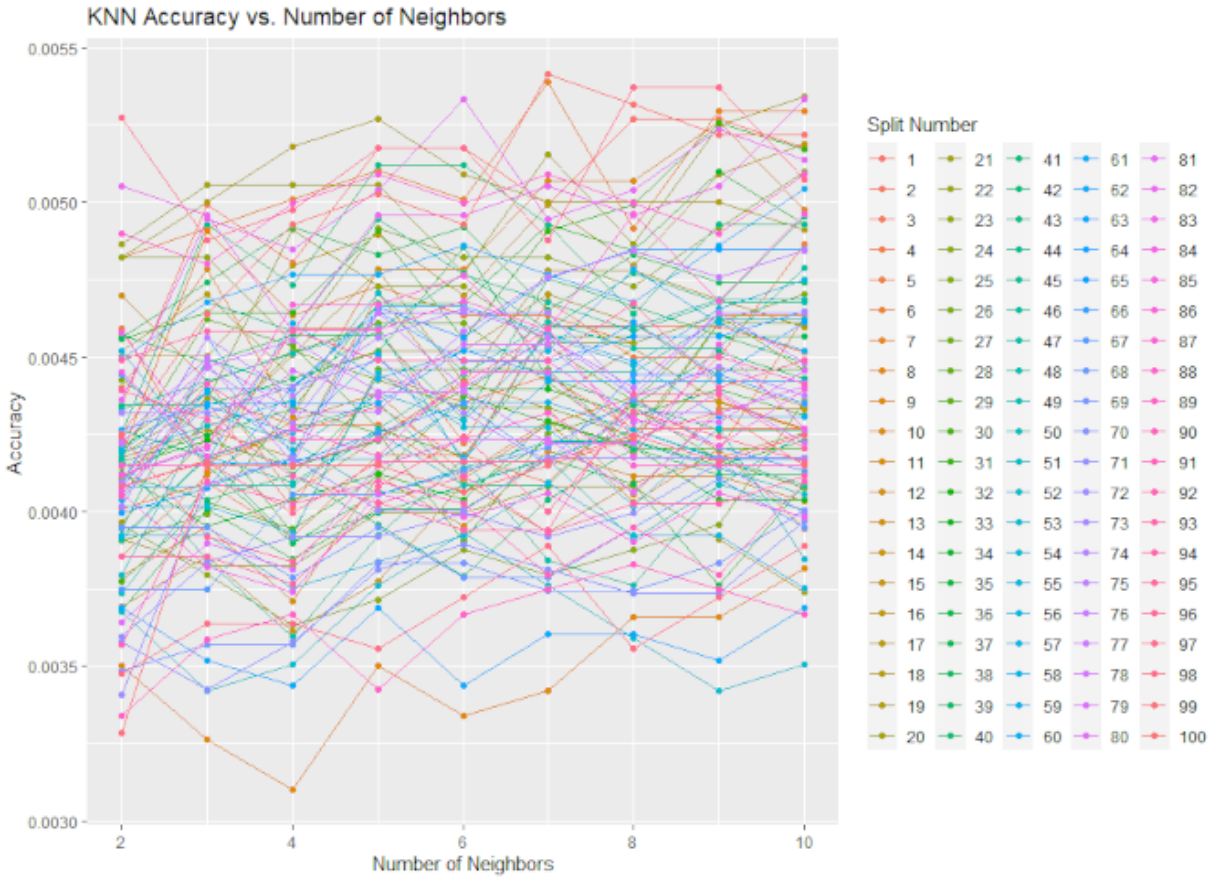

Figure 10: KNN accuracy vs Number of neighbours

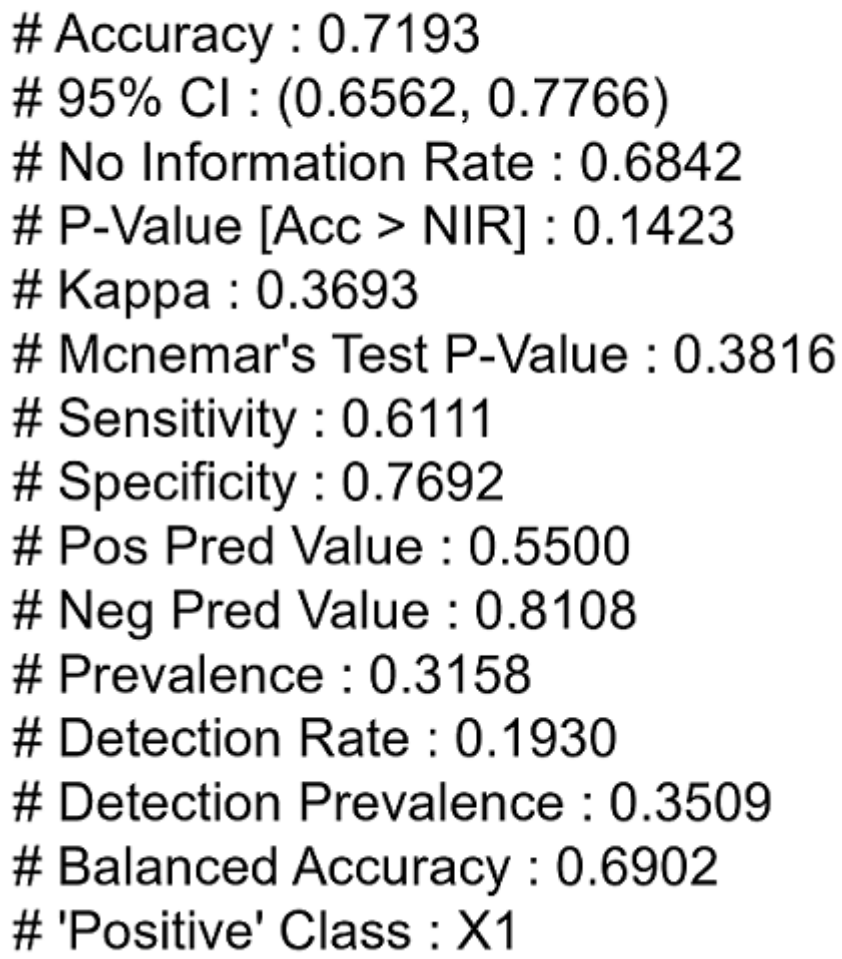

Figure 11: Result from KNN

7.2 Logistic Regression:

Next, we'll use another workhorse from the machine learning toolkit: regression. For this dataset, where we predict a binary outcome. Again, I will cross-validate the logistic regression model by repeatedly splitting the data into different training and test sets.

For all folds, we achieve an average model accuracy of 77%, and performance varies from 65-77% depending on the exact division of the training test. Logistic regression appears to be somewhat more accurate in this data set than the KNN, even with optimal selection.

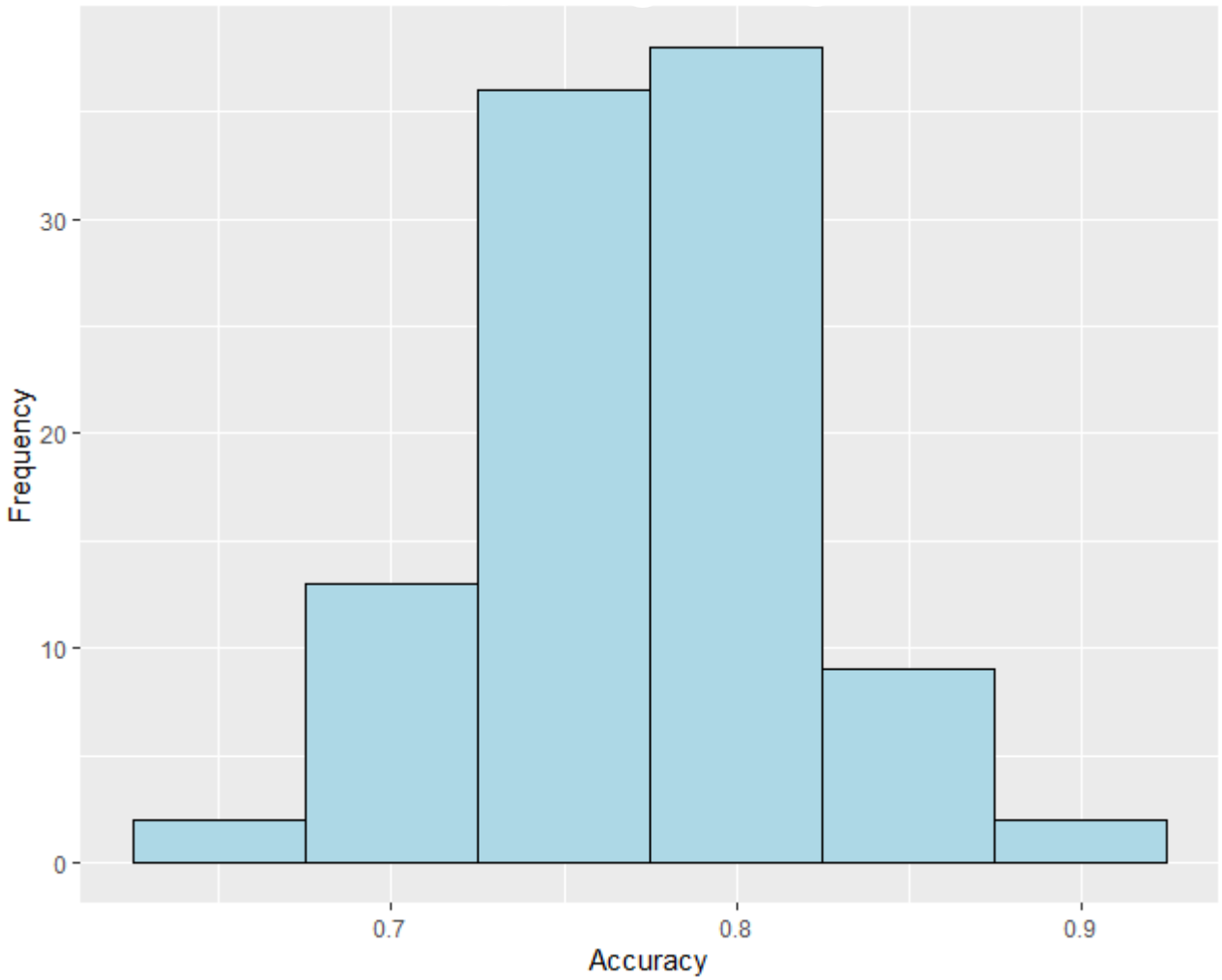

Figure 12: Accuracy distribution of Logistic Regression model

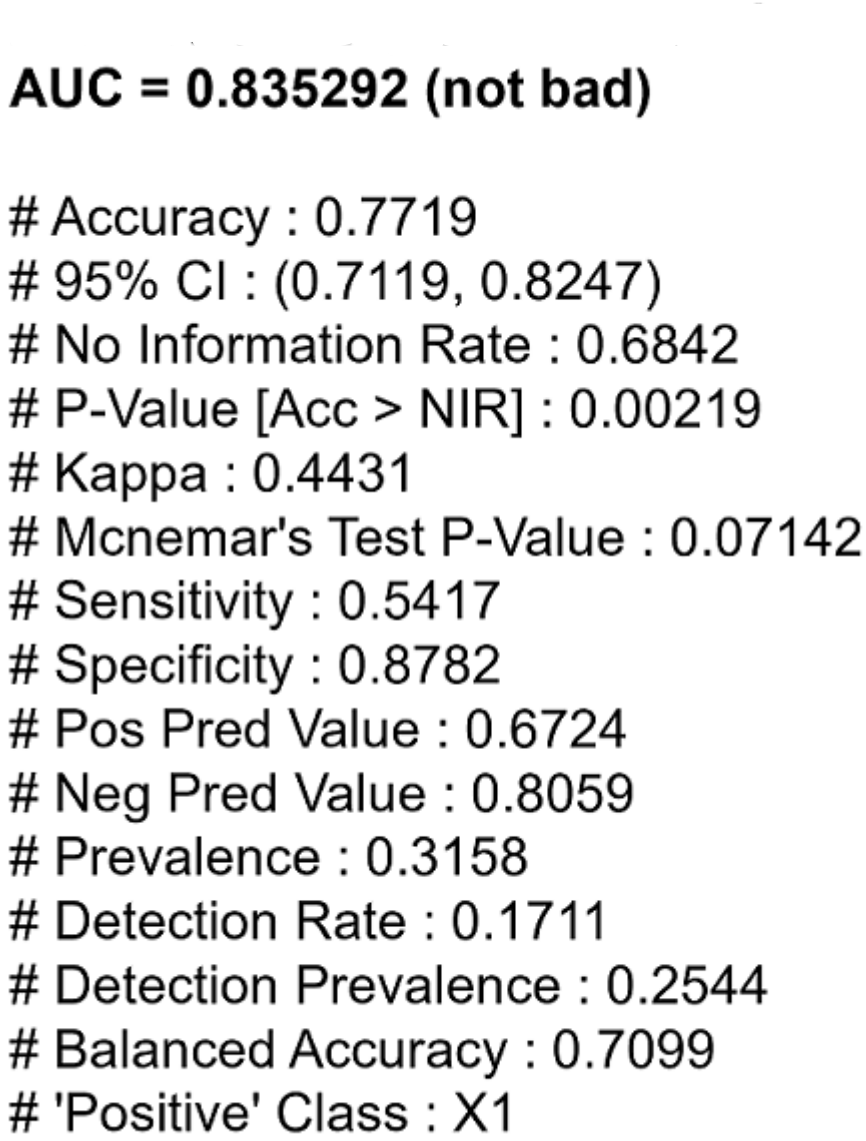

Figure 13: Result from Logistic Regression

8. Conclusion:

Based on patient data, we successfully produced a data science product that can help healthcare practitioners and researchers in forecasting the onset of diabetes. To acquire insights into the dataset, we used a combination of data processing, visualization, summary statistics computation, and data analysis approaches. To predict diabetes onset, various machine learning techniques, including KNN and Logistic Regression, were used, with logistic regression showing somewhat greater accuracy in this dataset.

Finally, this project illustrates the potential of data science and machine learning in diabetes diagnosis and gives significant insights for medical professionals and researchers. examining more

machine learning techniques, improving data protection measures, and performing trials to maximize model performance might be future advancements and research paths.

9. References:

- Basha, S.M., Balaji, H., Iyengar, N.C.S. and Caytiles, R.D., 2017. A soft computing approach to provide recommendation on PIMA diabetes. *Heart*, *106*, pp.19-32.
- Chang, V., Bailey, J., Xu, Q.A. and Sun, Z., 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, pp.1-17.
- Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L. and Bellazzi, R., 2018. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, *12*(2), pp.295-302.
- Garcia-Carretero, R., Vigil-Medina, L., Mora-Jimenez, I., Soguero-Ruiz, C., Barquero-Perez, O. and Ramos-Lopez, J., 2020. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Medical & biological engineering & computing*, *58*, pp.991-1002.
- Huang, Y., Shi, Q., Zuo, J., Pena-Mora, F. and Chen, J., 2021. Research status and challenges of data-driven construction project management in the big data context. *Advances in Civil Engineering*, *2021*, pp.1-19.
- Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, *8*, pp.1-14.
- Mujumdar, A. and Vaidehi, V., 2019. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, pp.292-299.
- Roglic, G., 2016. WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, *1*(1), p.3.
- Sarker, I.H., Faruque, M.F., Alqahtani, H. and Kalim, A., 2020. K-nearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(26), pp.e4-e4.
- Selvakumar, S., Kannan, K.S. and GothaiNachiyar, S., 2017. Prediction of diabetes diagnosis using classification based data mining techniques. *International Journal of Statistics and*

Systems, *12*(2), pp.183-188.

- Skyler, J.S., Bakris, G.L., Bonifacio, E., Darsow, T., Eckel, R.H., Groop, L., Groop, P.H., Handelsman, Y., Insel, R.A., Mathieu, C. and McElvaine, A.T., 2017. Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*, *66*(2), pp.241-255.

- Wei, S., Zhao, X. and Miao, C., 2018, February. A comprehensive exploration to the machine learning techniques for diabetes identification. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)* (pp. 291-295). IEEE.

- World Health Organisation, 2016. Global report on diabetes. *World Health Organisation*.

- Zhu, C., Idemudia, C.U. and Feng, W., 2019. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, p.100179.