- **Have you selected a topic for Assignment 3 that is different from the one that you used for Assignment 1 (i.e., have you rewrote the first two sections of the report)?**

  Ans: The topic for assignment 3 is the same as assignment 1. And I had modified the first two sections of the report based on the feedback that I received from Week 7 applied class.

Feedback from Assignment 1

- **What is the feedback given by your tutor for Assignment 1?**
  The feedback I received from my tutor is listed below.
  - I did not provide any reference in my paper.
  - Misleading heading (e.g. 1.0 or 2.0).
  - Grammatical errors.

- **Please briefly describe how you act upon the provided feedback to prepare Assignment 3 (no more than 150 words).**
  To improve my paper based on the feedback, I first put some references to support my opinions. Second, I replaced the misleading heading by clearer heading. For instance, I replaced 1.0 Project Description by 1. Project Description. Finally, I reviewed my paper line by line and fixed the grammatical errors.

Diabetes Diagnosis Using Decision Trees and Regression Models

1. Project Description

1.1 Introduction

Diabetes is a chronic medical condition which might result in serious health problems (Saeedi et al., 2020). Akil et al. (2021) stated that the traditional method that diagnoses diabetes could cost a great amount of time and waste a lot of resources. Thanks to the significant advancement of data science and machine learning, diagnostic procedures could be improved by developing predictive models which could analyse patient data rapidly and accurately, as mentioned by Fregoso-Aparicio et al. (2021). This project aims to develop decision trees and regression models to support healthcare providers in diabetes diagnosis. The models are going to use key health parameters such as gender, age and other relevant factors to predict the likelihood of patients having diabetes.

1.2 Objective

The objective of this project is to develop a system using decision trees and regression models to predict the probability of diabetes in patients based on their relevant body information.

1.3 Data Description

This dataset was collected from the laboratory of Medical City Hospital in Iraq. This dataset consists of 1000 entries, each entry records patients' medical information. Detailed information of recorded features is provided below.

- **ID:** Identifier of the patient (Categorical)
- **No_Pation:** Number of the patient (Continuous)
- **Gender:** Gender of the patient (Categorical, e.g., Male, Female)
- **AGE:** Age of the patient (Continuous)
- **Urea:** Blood urea concentration (Continuous)
- **Cr:** Serum creatinine level (Continuous)
- **HbA1c:** Glycated hemoglobin (Continuous)
- **Chol:** Total cholesterol level (Continuous)
- **TG:** Triglycerides level (Continuous)
- **HDL:** High-density lipoprotein level (Continuous)
- **LDL:** Low-density lipoprotein level (Continuous)
- **VLDL**: Very-low-density lipoprotein level (Continuous)
- **BMI:** Body Mass Index (Continuous)
- **CLASS:** The presence of diabetes (Categorical; N for non-diabetic, Y for diabetic, and P for predict-diabetic)

1.4 Data Science Roles and Responsibilities

- **Data Scientist** (Simplilearn, 2024)
    - **Description:** Data scientist is the core role in developing the models for diabetes diagnosis.
    - **Responsibilities**
        - **Exploratory Data Analysis (EDA):** Understand the relationships and patterns in the dataset.
        - **Model Development and Evaluation:** Use decision trees and regression to develop models and evaluate the accuracy of models.
        - **Interpretation:** Ensure the results of models can be correctly delivered to healthcare providers.
- **Data Engineer** (Chia, 2024)
    - **Description:** Data engineer focuses on gathering and preprocessing the data, ensuring all data formats are correct for analysis and modelling.
    - **Responsibilities**
        - **Data Collection:** Collect data from various sources.
        - **Data Cleaning:** Deal with missing values, outliers and inconsistencies of data formats.
- **System Architect** (Shiff, 2022)
    - **Description:** System Architect designs the system framework and ensures final models could be used in existing healthcare IT infrastructure.
    - **Responsibilities**
        - **Integration:** Ensure the final models could be connected to existing systems.
        - **User Interface:** Develop user interface so that healthcare providers could interact with models and view results.

2. Business Model

2.1 Application Area

This project can be applied to medical areas, specifically for disease diagnosis. Early and accurate diagnosis is crucial since patients could be treated effectively.

2.2 Business Benefits

- **Accuracy of diagnosis:** The predictive models are able to significantly raise the accuracy of diabetes diagnosis by analysing detailed information that might be neglected.
- **Time Efficiency:** The automatic process reduces analysis time. The models allow healthcare providers to have more time focusing on treatments for patients.

2.3 Beneficiaries

- **Healthcare Providers:** Hospitals, clinics and medical professionals can all benefit from advanced diagnostic tools which are able to support providers to make correct decisions and save more time.
- **Patients**: Patients can receive faster, accurate diagnoses and immediate treatments.

2.4 Challenges

- **Privacy and Security of Data:** Medical data is more sensitive than other data. Therefore, how to properly use sensitive data and ensure data is stored securely is crucial (Modak, 2024).
- **Quality of Data:** The successful predictive models are trained by data with good quality and comprehensive data. Inconsistent or biassed data can affect the accuracy of predictive models negatively (Modak, 2024).
- **Integration with Existing System:** The new diagnostic models should be integrated properly with the existing system so that it can be used efficiently. Hence, the integration with the existing system should be planned carefully.

3. Characterising and Analysing Data

3.1 Potential Sources & Characteristics of the Data & Required Tools

- **Potential Sources**
  Due to this project involving the medical data which is sensitive and relatively harder to collect through open data websites than the general dataset. To access the diabetes data for training models, some websites contain open datasets for diabetes analysis.
    - **Kaggle (https://www.kaggle.com/):** Kaggle offers various data for data scientists to analyse and train models.
    - **Australian Institute of Health and Welfare (https://www.aihw.gov.au/):** This website provides a wide range of datasets which cover a variety of health and welfare topics.
    - **World Health Organisation (https://www.who.int/):** WHO website contains comprehensive resources for global health information.

  The dataset used for demonstration for this project is collected from Kaggle. The data in this dataset was extracted from the laboratory of Medical City Hospital in Iraq. The link to download the dataset is provided below.
    - **Diabetes Dataset:** https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset
- **Characteristics of the data**
  4Vs in big data represent Volume, Variety, Velocity and Veracity (The Four V's of Big Data, 2023). The list below is the explanation of 4Vs in big data (Yadav, S., 2022).
    - **Volume:** Volume means the size of the dataset which is going to be used to analyse and process. With large amounts of data, it improves the accuracy of machine learning models. Furthermore, compared to a small dataset, it allows us to discover patterns easily.
    - **Variety:** Variety represents the different types of the same data. In general, structured, semi-structured and unstructured data are the most common types of data nowadays. Different types of data provide various aspects of the data.
    - **Velocity:** Velocity represents the speed about how fast the data is generated and consumed. The higher velocity of the data, the faster the response to address issues.

    - **Veracity:** Veracity represents the reliability of the data. The high data quality and accuracy provides better decision making from the data.
- **Required Tools**
    - **Programming Languages**
        - **Python (https://www.python.org/):** A high-level, general purpose programming language.
        - **R (https://www.r-project.org/):** A programming language for statistical computing and graphics.
    - **Data Storage**
        - **Oracle Developer (https://www.oracle.com/au/database/sqldeveloper/):** It provides an integrated development environment for relational databases which can store a large amount of data.
        - **MongoDB (https://www.mongodb.com/):** MongoDB is a NoSQL database which allows users to handle a large amount of unstructured data.
    - **Data Processing & Analysis & Modelling**
        - **Jupyter Notebooks (https://jupyter.org/):** An integrated development environment for Python.
        - **R Studio (https://posit.co/download/rstudio-desktop/):** An integrated development environment for R.
    - **Data Visualisation**
        - **Jupyter Notebooks:** Python contains powerful libraries for data visualisation.
        - **R Studio:** R Studio contains powerful libraries for data visualisation.
        - **Tableau (https://www.tableau.com/):** Tableau provides a simple drag and drop interface for users to create various data visualisations.

3.2 Data Analysis & Statistical Methods

This section will discuss the methodology of data analysis and statistical methods including decision tree and regression models that will be used in the demonstration in the next section.

3.2.1 Correlation Analysis

According to Andreev et al. (2019), correlation analysis describes the dependency between each pair of variables. Moreover, the outcome of models can be improved by removing the variables which are too similar to others. Hence, I analyse the dependency of each pair of variables by generating the confusion matrix and attempting to remove similar variables from the diabetes dataset.

3.2.2 Bootstrapping

Bittmann (2021) mentions that the purpose of bootstrapping is to collect more information by resampling the training data. Bootstrapping can not only prevent models from overfitting but also making the most out of limited data. Furthermore, bootstrapping produces multiple models so that those models can vote for the final answer. This way can also enhance the accuracy of predictions.

3.2.3 Decision Tree

Since this project will use multiple models to vote the final prediction, I decided to use different split methods to build models, which can enhance the accuracy. Therefore, in this project, I will use entropy and gini index split method to perform the decision tree models.

3.2.4 Regression Model

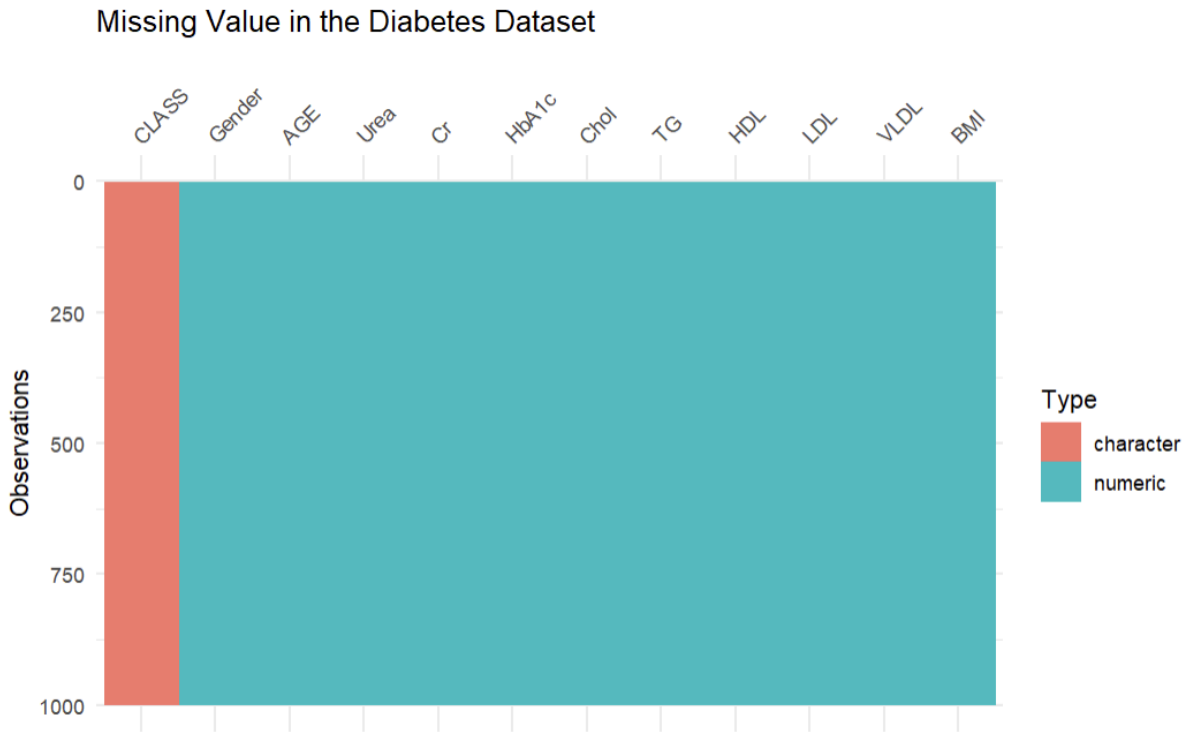
As the final target column in the dataset is categorical, logistic regression will be used for this project. However, the final output of the logistic regression is binary and the models from this project are multi-class classification. Alternatively, I will use multinomial logistic regression to achieve the same purpose since multinomial logistic regression is useful to classify more than two different subjects according to the data (Multinomial Logistic Regression, 2024).

3.3 Demonstration

To perform the basic analysis of this project, I will use the diabetes dataset from Kaggle to build decision trees and regression models step by step.

3.3.1 Missing Value

First, to examine whether there is any missing value in this dataset, I use vis_data() and miss_summary() functions from “visdat” and “naniar” libraries respectively. Figure 1 and figure 2 illustrate that there is no missing value in this dataset.



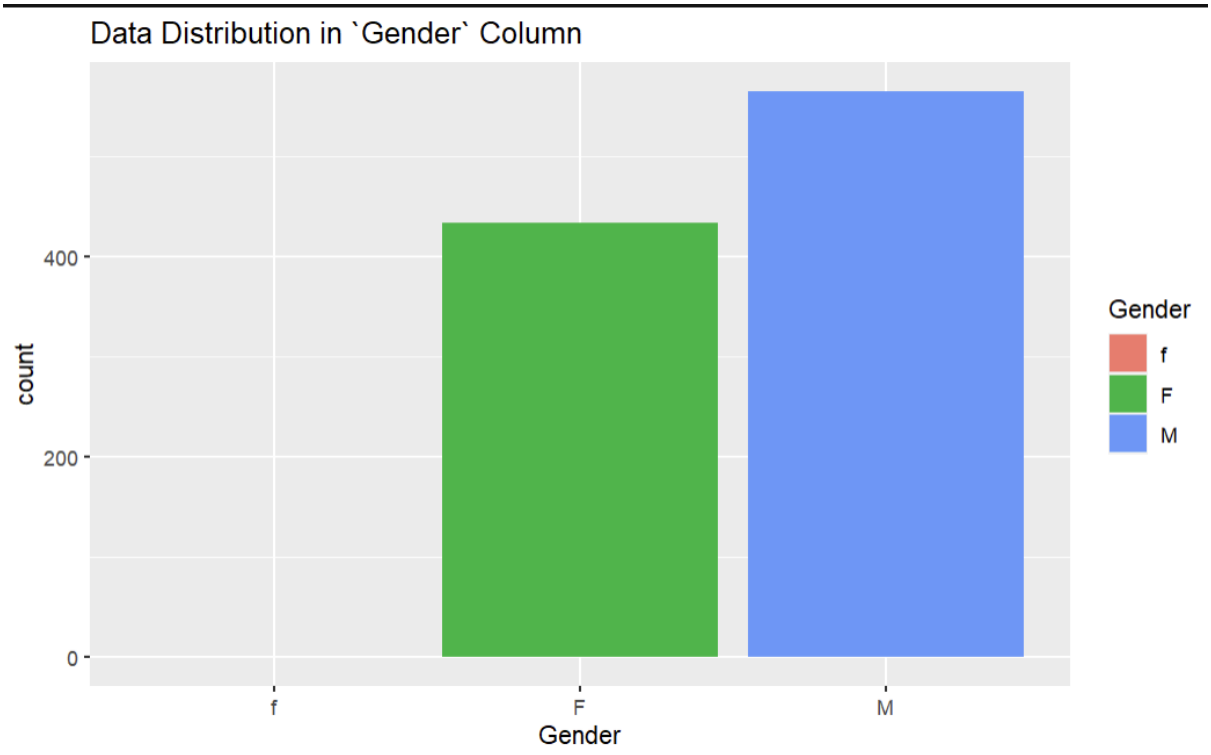
F*igure 1*. Missing Values in the Diabetes Dataset (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

variable <chr>	n_miss <int>	pct_miss <dbl>
ID	0	0
No_Pation	0	0
Gender	0	0
AGE	0	0
Urea	0	0
Cr	0	0
HbA1c	0	0
Chol	0	0
TG	0	0
HDL	0	0

F*igure 2*. The Table of Missing Values in the Diabetes Dataset (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

3.3.2 Data Wrangling

There is a capital letter issue in the `Gender` column. Figure 3 is the visualisation of the issue.



F*igure 3*. Data Distribution in `Gender` Column (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

To address this issue, I modified the error data to the correct format. Furthermore, I transfer this column from categorical to continuous to support the further actions. The graph below is the result after the modification.
Figure 4 is the result after the modification.

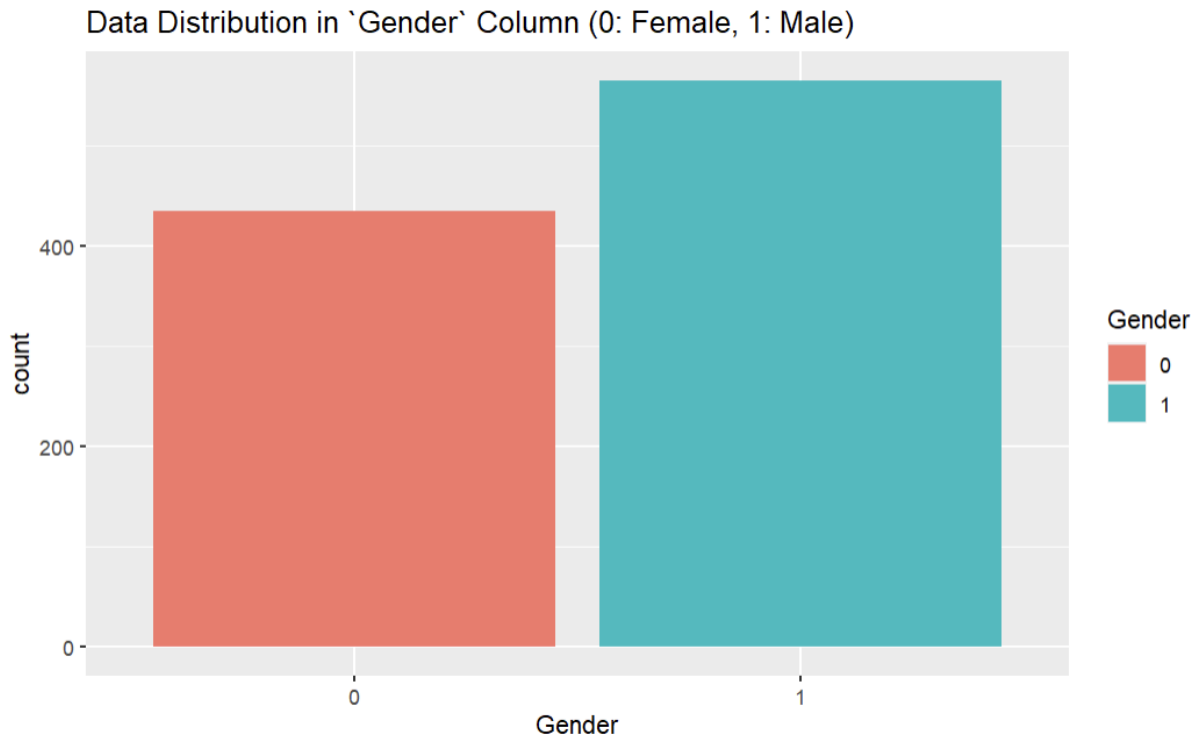


Figure 4. Data Distribution in `Gender` Column - Modified (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

3.3.3 Correlation Analysis

To analyse the relationship between attributes and filter attributes with high relationship, I use Pearson correlation to produce the correlation matrix since all attributes except the target attribute, CLASS, are all continuous. Figure 5 is the correlation matrix based on the Pearson correlation.

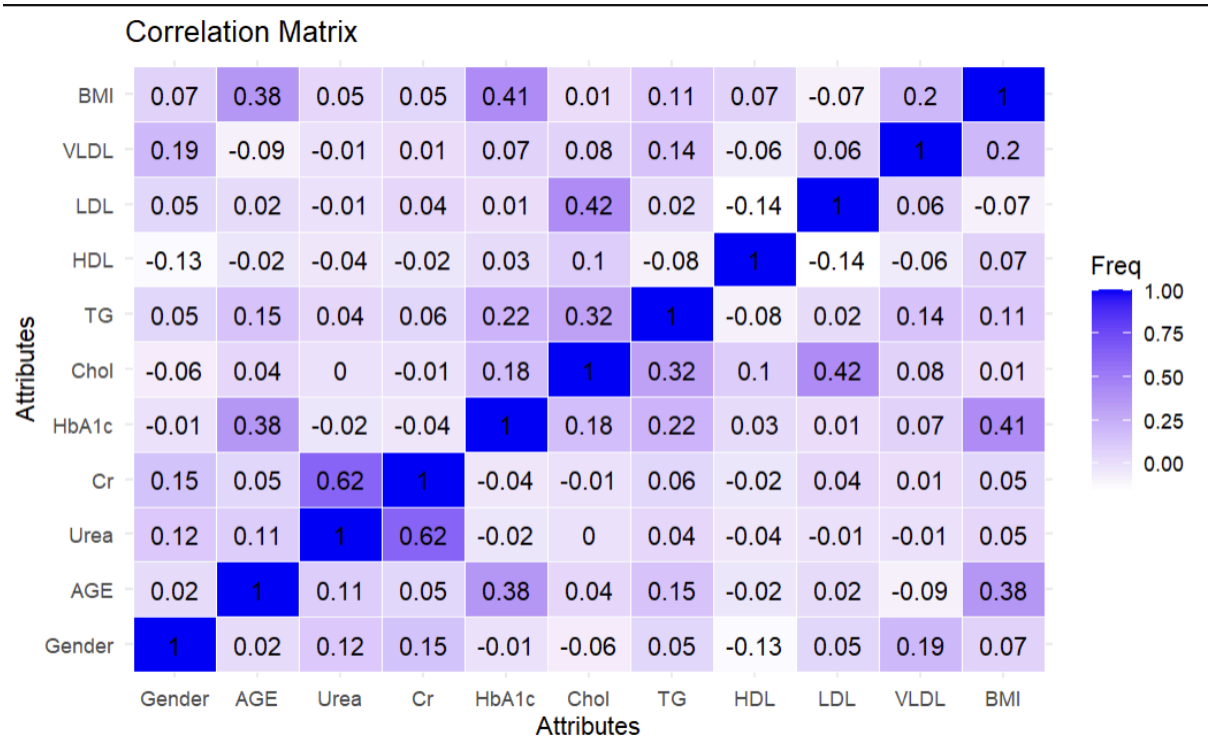


Figure 5. Correlation Matrix (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

3.3.4 Splitting of train/test sets & Bootstrapping

To test the accuracy of models, I split the whole dataset into train and test sets and the proportion of these two sets is 8:2. Furthermore, I use bootstrapping to build multiple models for decision trees and regression models to vote on the final prediction. To successfully regenerate the result, I use the set.seed() function to fix the random seed. Afterward, I use the createDataPartition() function in the “caret” library to achieve the purpose. As for bootstrapping, I use a for loop to randomly extract data from the train set via sample() function.

3.3.5 Decision Tree & Regression Models

- **Decision Tree**
  For decision tree models, I build 5 models for each split method, Entropy and gini impurity. The figure 6 is the sample decision tree.

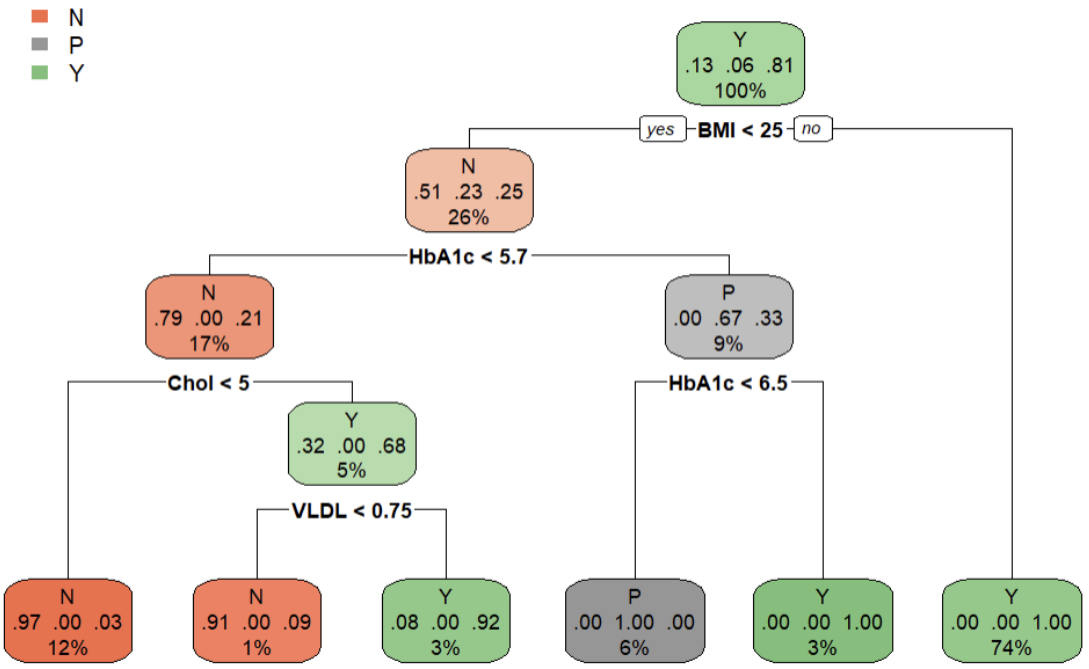


Figure 6. Sample Decision Tree (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

- **Regression Models**
  For regression models, since the target column in the dataset for this demonstration is categorical and there are more than two distinct values in this column, I use multinomial logistic regression to build 5 models for voting the final prediction.

3.3.6 Evaluation

To evaluate the prediction from both models, I used confusion matrix, accuracy and evaluation matrix to measure the outcome. The accuracy for decision tree models is 97.47%, whereas the accuracy for regression models is 92.93%. Figure 7-10 are the evaluations for

both models.

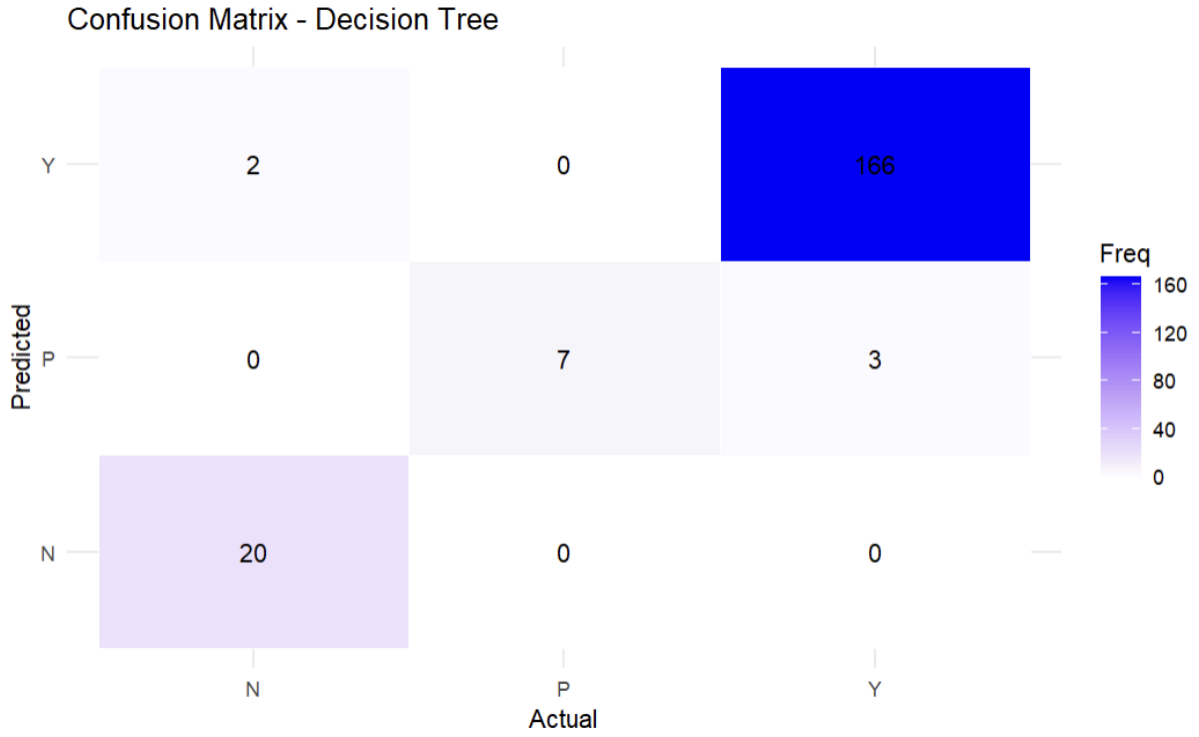


Figure 7. Confusion Matrix - Decision Tree (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

	Precision <dbl>	Recall <dbl>	F1 <dbl>
Class: N	0.9090909	1.0000000	0.9523810
Class: P	1.0000000	0.7000000	0.8235294
Class: Y	0.9822485	0.9880952	0.9851632

Figure 8. Evaluation Matrix - Decision Tree (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

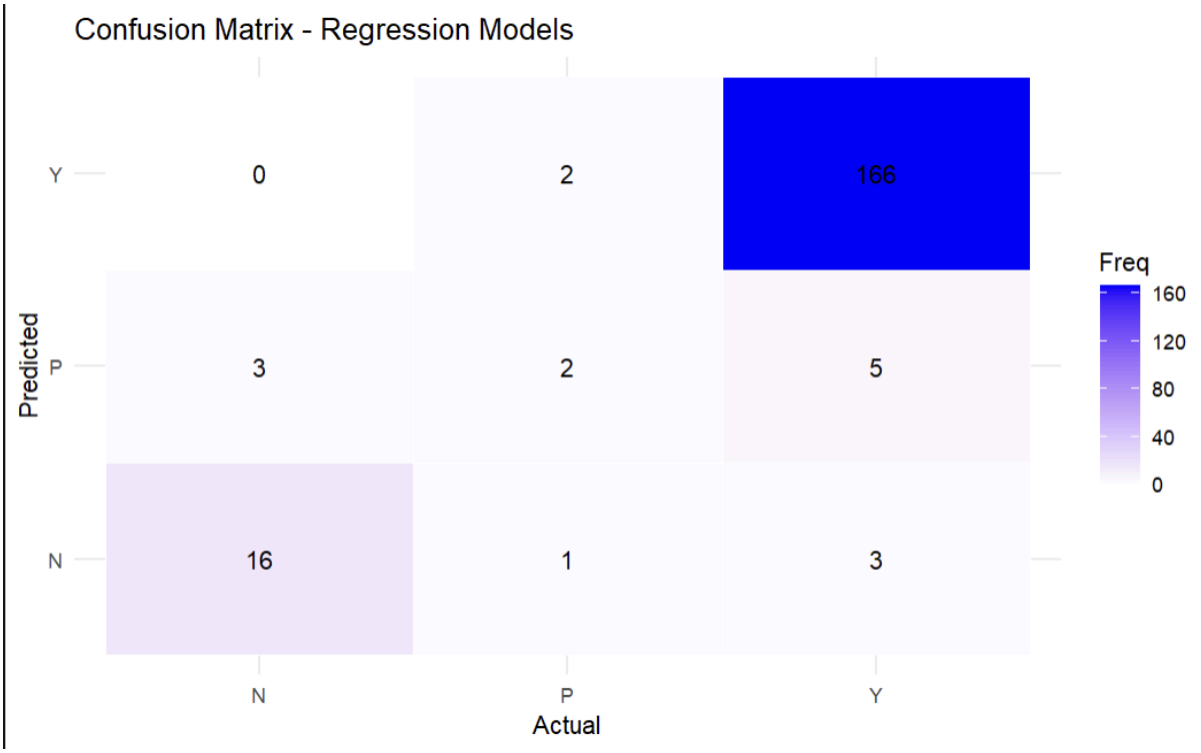


Figure 9. Confusion Matrix - Regression Models (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

	Precision <dbl>	Recall <dbl>	F1 <dbl>
Class: N	0.8421053	0.8000000	0.8205128
Class: P	0.4000000	0.2000000	0.2666667
Class: Y	0.9540230	0.9880952	0.9707602

F*igure 10.* Evaluation Matrix - Regression Models (2024). [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

3.3.7 Conclusion

Moghaddam et al. (2024) state that if the accuracy of the model is around 85-90%, then this model is considered a reliable model for clinical use. Therefore, the demonstration above is feasible since both accuracy of models are above 90%.

4. Standard for Data Science Process, Data Governance and Management

4.1 Standard for Data Science Process

This type of data science project can follow CRISP-DM (Cross-Industry Standard Process for Data Mining), introduced by Hotz (2024). This framework is the most typical methodology for data science project and it includes the following stages:

1. **Business Understanding:** This stage focuses on identifying the goals and requirements of the project. In this project, using machine learning models to support healthcare providers enhancing the accuracy of diabetes diagnosis is one of the goals of this project. Requirements of this project can be the minimum accuracy of models. For instance, the accuracy of models produced from this project should be above 90%.
2. **Data Understanding:** This stage focuses on identifying, collecting and analysing the feasibility of datasets. Collecting and familiarising the diabetes dataset from Kaggle is included in this stage.
3. **Data Preparation:** This stage is the most significant part of the project since 80% of the project is completed in this stage. This stage focuses on preparing data from training models. This stage includes data cleaning and data wrangling.
4. **Modelling:** Training and testing models are included in this stage. In this project, this stage begins from the splitting of train and test sets and the bootstrapping of samples to building models.
5. **Evaluation:** This stage focuses on evaluating models. In this project, the use of confusion and evaluation matrices to evaluate models is part of this stage.
6. **Deployment:** After the completion of models, models should be deployed to existing systems for practical use. In this project, the integration of models and existing medical systems is included in this stage.

4.2 Data Governance and Management

Data governance and management is about ensuring the security, confidentiality, accuracy, accessibility, availability and usability of data. Data governance and management is the establishment of multiple internal data policies to guide how data is collected, stored, processed and disposed of (What is Data Governance?, n.d.). I provide a feasible practice of data governance and management for this project, including potential ethical concerns with the use of the data..

- **Accessibility**
  - **Role-Based Access Control (RBAC):** It is able to reduce the danger of illegal access by letting administrators assign the roles to users based on individual responsibilities (Maulina & Rasjid, 2024).
- **Security**
  - **Encryption:** Encryption prevents breaches of data. It prevents unauthorised users from accessing the information in the database (Google Cloud, n.d.).
- **Confidentiality**
  - **Anonymisation:** Due to the data for this project being highly sensitive, the anonymisation of data can assure there is no connection between the information in the database and any single person (Rupp & Grafenstein, 2024).
- **Ethical Concerns**
  - **Informed Consent:** Personal data can only be used after the agreement from each individual patient.
  - **Data Bias:** The performance of models might be affected by bias in the data. Hence, bias data should be removed from the training data to enhance the accuracy of models.

5. Reference

Akil, A. A. S., Yassin, E., Al-Maraghi, A., Aliyev, E., Al-Malki, K., & Fakhro, K. A. (2021). *Diagnosis and treatment of type 1 diabetes at the dawn of the personalized medicine era.* Journal of translational medicine, 19(1), 137. https://link.springer.com/article/10.1186/s12967-021-02778-6

Andreev, V. P., Liu, G., Zee, J., Henn, L., Flores, G. E., & Merion, R. M. (2019). *Clustering of the structures by using "snakes-&-dragons" approach, or correlation matrix as a signal.* PloS One, 14(10), e0223267–e0223267. https://doi.org/10.1371/journal.pone.0223267

Bittmann, F. (2021). *Bootstrapping : an integrated approach with Python and Stata*. De Gruyter. https://doi.org/10.1515/9783110693348

Chia, A. (2024, May 30). *What is a Data Engineer?* Splunk. https://www.splunk.com/en_us/blog/learn/data-engineer-role-responsibilities.html#:~:text=A%20data%20engineer%20develops%2C%20builds,systems%20like%20AWS%20or%20Azure.

Decision Trees in Machine Learning Using R. (2023 June 1). DataCamp. https://www.datacamp.com/tutorial/decision-trees-R

Fortinet. (n.d.). *What Is Encryption?* https://www.fortinet.com/resources/cyberglossary/encryption#:~:text=The%20Benefits%20of%20Encryption&text=Encryption%20ensures%20no%20one%20can,intercepting%20and%20accessing%20sensitive%20data.

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). *Machine learning and deep learning predictive models for type 2 diabetes: a systematic review.* Diabetology & metabolic syndrome, 13(1), 148. https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9#Sec6

Google Cloud. *What is Data Governance?* (n.d.). https://cloud.google.com/learn/what-is-data-governance

Hotz, N. (2024). *What is CRISP DM?* Data Science Process Alliance https://www.datascience-pm.com/crisp-dm-2/

IBM. (2024). *Multinomial Logistic Regression.* https://www.ibm.com/docs/en/spss-statistics/29.0.0?topic=regression-multinomial-logistic

Jupyter Notebooks. (n.d.). https://jupyter.org/

Kaggle. (n.d.). https://www.kaggle.com/

Kaggle. (n.d.). *Diabetes Dataset*. https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset

Maulina, A., & Rasjid, Z. E. (2024). *Unified Access Management for Digital Evidence Storage: Integrating Attribute-based and Role-based Access Control with XACML.* International Journal of Advanced Computer Science & Applications, 15(3). https://doi.org/10.14569/IJACSA.2024.01503131

Modak, S. K. S. (2024). *Machine and deep learning techniques for the prediction of diabetics: a review.* Multimed Tools. https://doi.org/10.1007/s11042-024-19766-9

MongoDB. (n.d.). https://www.mongodb.com/

Python. (n.d.). https://www.python.org/

R. (n.d.). https://www.r-project.org/

R Studio. (n.d.). https://posit.co/download/rstudio-desktop/

Rupp, V., & von Grafenstein, M. (2024). *Clarifying "personal data" and the role of anonymisation in data protection law: Including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection. The Computer Law and Security Report, 52,* 105932-. https://doi.org/10.1016/j.clsr.2023.105932

Saeedi, P., Salpea, P., Karuranga, S., Petersohn, I., Malanda, B., Gregg, E. W., Gregg, E. W., Unwin, N., Wild, S. H. & Willians, R. (2020). *Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas.* Diabetes research

and clinical practice, 162, 108086. https://www.sciencedirect.com/science/article/pii/S016882272030139X

Shiff, L. (2022, March 2). *What Does a System Architect Do?* Bmc. https://www.bmc.com/blogs/system-architect/

Simplilearn, (2024, August 26). *Data Scientist Job Description: Role, Responsibilities & Skills*. https://www.simplilearn.com/data-scientist-job-description-article#:~:text=A%20Data%20Scientist's%20roles%20and,manner%2C%20and%20propose%20solutions%20and

SQL Developer. (n.d.). Oracle. https://www.oracle.com/au/database/sqldeveloper/

Tableau. (n.d.). https://www.tableau.com/

Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z. et al. (2024). *Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm.* BMC Med Res Methodol 24, 220. https://doi.org/10.1186/s12874-024-02341-z

The Four V's of Big Data. (2023 January 31). https://opensistemas.com/en/the-four-vs-of-big-data/

Yadav, S. (2022 August 18). *The Complete Guide to the 4 V's of Big Data.* Baseline. https://www.baselinemag.com/analytics-big-data/the-complete-guide-to-the-4-vs-of-big-data/