WID3007 FUZZY LOGIC

Group Assignment

Diabetes Prediction Using Fuzzy Logic

Group Name: Fuzzing

Group Members:

Name	Matric No.	Email
Ting Wei Jing	17205876	wid190056@siswa.um.edu.my
Stanley Quek Chen Phin	17204457	wid190051@siswa.um.edu.my
Lee Cheng Hui	17204205	wid190024@siswa.um.edu.my
Soh Zhi Chen	17204929	wid190050@siswa.um.edu.my

Problem Statement

According to the CDC's National Diabetes Statistics Report (2022), there were an estimated 120 million or so adults that are living with diabetes or prediabetes within the United States. Malaysia does not fare much better either, as according to Akhtar et al. (2022), diabetes is expected to affect 7 million Malaysians by the year 2025. This is worrying as diabetes is a chronic condition, meaning that it is a long-term health problem that requires ongoing treatment. However, diabetes often goes undetected, and when left untreated, it results in serious health issues, such as heart failure, kidney damage, blindness, and so on. As such, our problem statement is: **Diabetes often goes undetected and untreated during its early stages, which can cause serious harm to the body and makes it difficult to treat down the line.**

Objectives

We developed a fuzzy logic system for predicting diabetic patients with these objectives in mind:

1. To identify individuals at high risk for developing diabetes, so that they can be targeted for preventive measures such as lifestyle interventions.
2. To diagnose diabetes at an early stage, so that people with the condition can receive treatment as soon as possible and minimise the risk of complications.
3. To assist medical professionals in the diagnosis of diabetes patients and to reduce the number of false negatives and false positive cases.

Dataset

The dataset (UCI Machine Learning, 2016) that has been used is taken from Kaggle, and it is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All collected sample data are from females who are at least 21 years old. The dataset consists of 768 patients. There are 268 samples identified as diabetic while 500 samples are identified as non-diabetics. However, there are missing values in the dataset and only 392 samples are retained after performing data cleaning.

Table 1: Description of the dataset columns

Column	Description
Number of pregnancies	Number of times the person gets pregnant
Plasma glucose concentration (Glucose)	Plasma glucose concentration in a person's body
Diastolic blood pressure	Diastolic blood pressure in (mm Hg)
Triceps skinfold thickness	Triceps skinfold thickness in (mm)
Serum insulin	2-Hour serum insulin (μU/ml)
Body mass index (BMI)	Body mass index (weight in kg/(height in m)^2)
Diabetes pedigree	History of diabetes associated with a particular person
Age	Age of a person
Class variable	Yes (diabetic) and No (non-diabetic)

System Input

Feature Selection

There are a total of 8 feature columns in the dataset. However, not all features would have a huge impact or correlation with the diabetes diagnosis. Hence, a Gradient Boosted decision tree implemented in XGBoost library was used to build a pseudo classifier to calculate the feature importance score of each feature. Based on the ranking result in Figure 1 below, Glucose, BMI and Age are the top 3 important features which will be used to build the knowledge base. By plotting the Pairplot of the top 3 features shown in Figure 2, Glucose feature has been further proven to be the most important feature as the scatter plot shows an obvious separable “Yes” and “No” clustering compared to the other features. Hence, the Glucose feature will be the fixed antecedent in the fuzzy system’s rules.

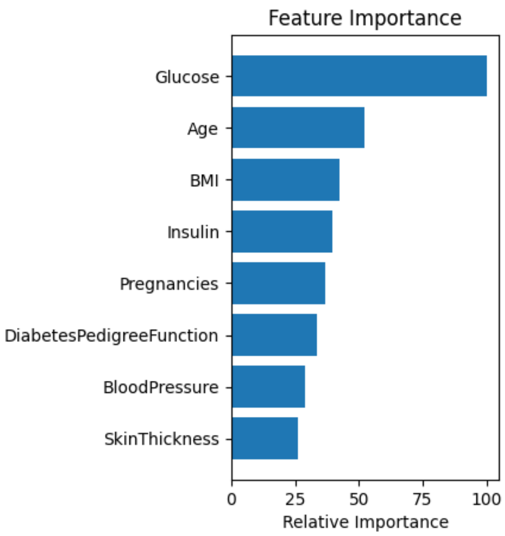


Figure 1: Feature importance ranking

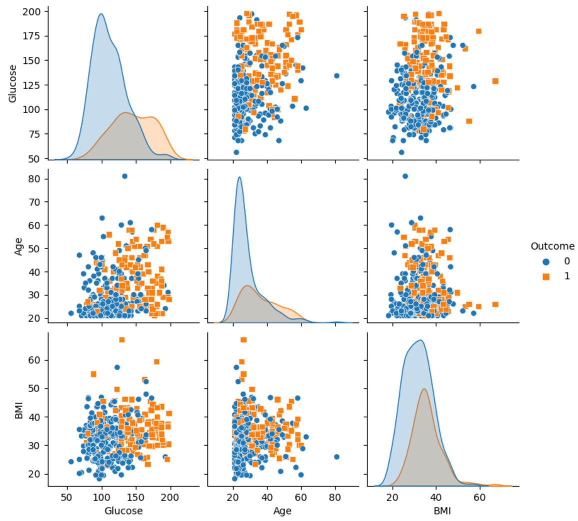


Figure 2: Pairplot of the top 3 features

System Output

The only output of this fuzzy logic system is the prediction of diabetes which is represented in the linguistic terms of "Yes" and "No".

Fuzzification (Linguistic Term)

Linguistic Term

Since glucose levels and BMI are related to the medical field and a lot of scientific research has been done on them, the linguistic terms and their corresponding values will not vary too much from person to person. For glucose level, we have taken the values from Freeth (2021), meanwhile the BMI values have been taken from Coulman and Toran (2020). Using these values, we extended each of their ranges by a small margin so that there is an overlap of membership functions. For the third feature, age, it is a little more controversial as its linguistic terms might be interpreted differently by different individuals. Hence, we have taken the range from Thompson et al. (2015), with the range being more promising and acceptable for all of our group members.

In Table 1, Table 2 and Table 3, the value range for each of the features and their corresponding linguistic terms are shown.

Table 2: Linguistic Terms Grouping for Glucose Level

Linguistic Term	Value Range
Very Low	53 - 70
Low	70 - 97.5
Medium	97.5 - 125
High	125 - 162.5
Very High	162.5 - 200

Table 3: Linguistic Terms Grouping for BMI

Linguistic Term	Value Range
Underweight	0 - 18.5
Normal weight	18.5 - 25
Overweight	25 - 30
Obesity 1	30 - 35
Obesity 2	35 - 40
Obesity 3	40 - 68

Table 4: Linguistic Terms Grouping for Age

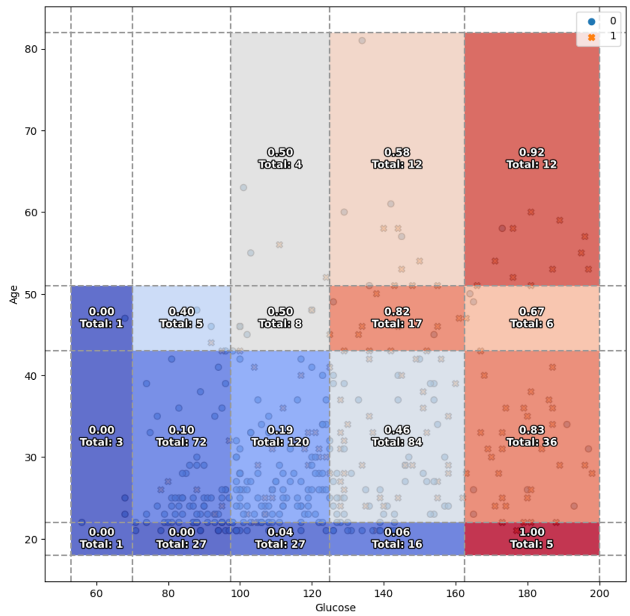
Linguistic Term	Value Range
Young	18 - 22
Middle Age	22 - 43
Old	43 - 51
Very Old	51 - 82

Knowledge Base Design

Rules Construction

The next step is to construct the appropriate rules for our fuzzy system. In order to determine the relationship between glucose level and age, the Grid Partition method was used.

In Figure 3 and 4, all the points were plotted in different ways (blue O for no diabetes, orange X for diabetes) and their coordinates were plotted based on the value of glucose and age/BMI. We then drew boundary lines for dividing the points into different regions. Next, we calculated the number of points with their class in each region and got a real number ratio indicating the positive class vs total number of points. Lastly, we have selected the regions with ratio either >= 0.8 or <= 0.2 to be our fuzzy system rules. In these regions, the values of glucose level and age/BMI were having more unique characteristics that could accurately represent each class. An example rule derived from Figure 3 is "*If the glucose level is high and the age is young, then the prediction for diabetes is no*".



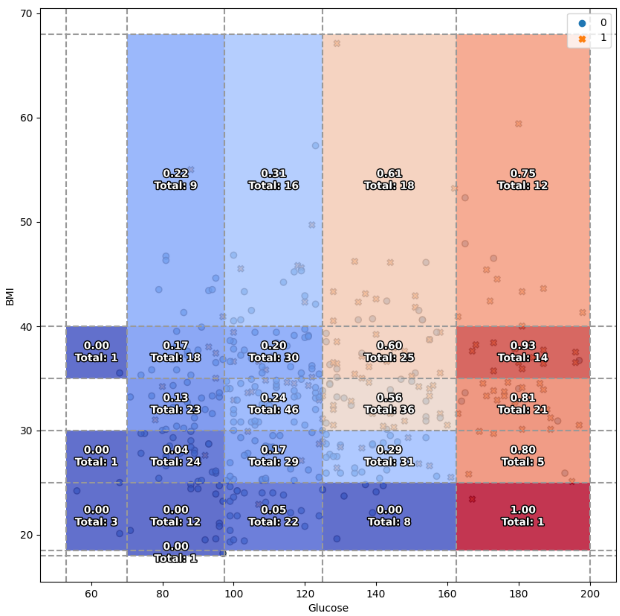


Figure 3: Partitioned Scatter Plot for Glucose vs Age

Figure 4: Partitioned Scatter Plot for Glucose vs BMI

Rules

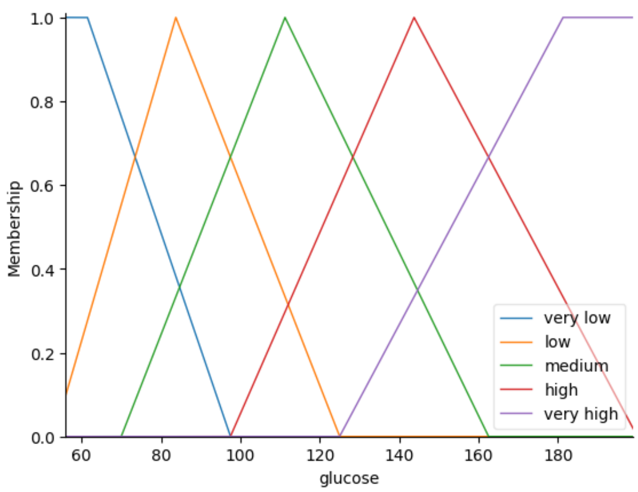
1. *If the glucose level is **very low** AND the age is **young** OR **middle age** OR **old**, then the prediction for diabetes is **no**.*
2. *If the glucose level is **low** AND the age is **young** OR **middle age**, then the prediction for diabetes is **no**.*
3. *If the glucose level is **medium** AND the age is **young** OR **middle age**, then the prediction for diabetes is **no**.*
4. *If the glucose level is **high** AND the age is **young**, then the prediction for diabetes is **no**.*
5. *If the glucose level is **high** AND the age is **old**, then the prediction for diabetes is **yes**.*
6. *If the glucose level is **very high** AND the age is **young** OR **middle age** OR **very old**, then the prediction for diabetes is **yes**.*
7. *If the glucose level is **very low** AND the BMI is **normal weight** OR **overweight** OR **obesity 2**, then the prediction for diabetes is **no**.*
8. *If the glucose level is **low** AND the BMI is **underweight** OR **normal weight** OR **overweight** OR **obesity 1** OR **obesity 2**, then the prediction for diabetes is **no**.*
9. *If the glucose level is **medium** AND the BMI is **normal weight** OR **overweight** OR **obesity 2**, then the prediction for diabetes is **no**.*
10. *If the glucose level is **high** AND the BMI is **normal weight**, then the prediction for diabetes is **no**.*
11. *If the glucose level is **very high** AND the BMI is **normal weight** OR **overweight** OR **obesity 1** OR **obesity 2**, the prediction for diabetes is **yes**.*

Membership Function

Both Trapezium and triangle membership function shapes are implemented for all linguistic terms. An accuracy comparison experiment of using different combinations of membership function shapes is conducted and the best combination of membership function shapes which achieves the highest accuracy is shown in Table 5.

Table 5: Best Membership Function Shape Combination

	Glucose	BMI	Age	Prediction	Accuracy
Shape	Triangle	Trapezium	Triangle	Triangle	80.36%



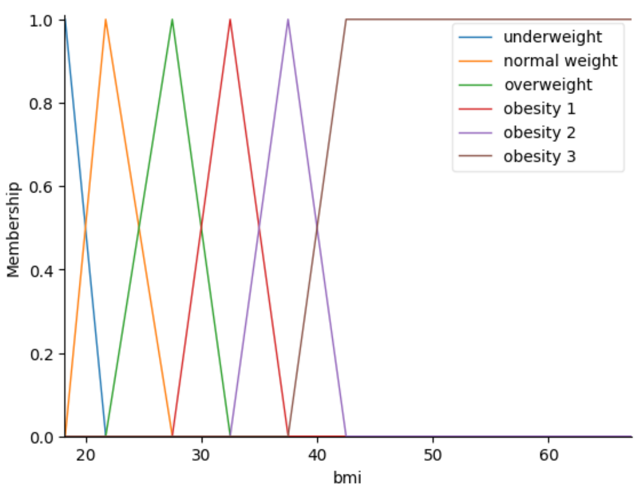


Figure 5: Final membership function for Glucose

Figure 6: Final membership function for BMI

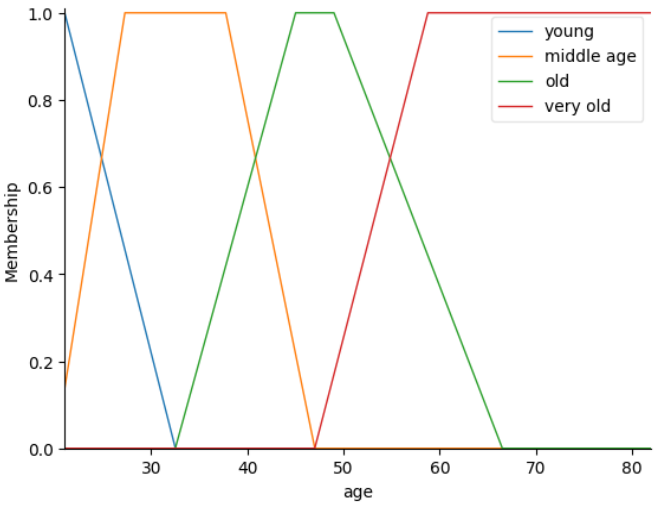


Figure 7: Final membership function for Age

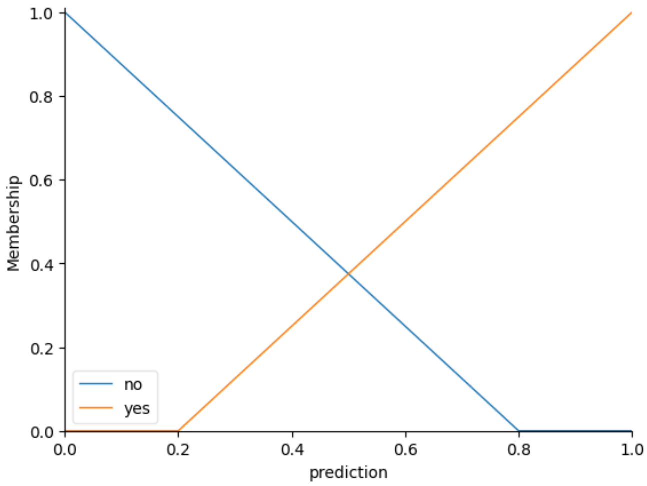


Figure 8: Final membership function for Diabetes Prediction

Fuzzy Inference System (FIS)

The Mamdani fuzzy model is used as the architecture of our FIS because it is easier to build compared to the Sugeno model which needs to design extra crisp functions for the consequent of the rules.

Defuzzification

All available defuzzification methods in *skfuzzy* have been attempted and the centroid method is used as it is the default method and has achieved the highest accuracy. A threshold of 0.5 is used to diagnose the patients as diabetic from the output crisp value. If the output crisp value is greater than 0.5, we predict the patient as diabetic else we predict them as non-diabetic.

Test Run

5 samples are randomly selected from the cleaned dataset and fed into the designed fuzzy inference system.

Table 6: Test Run Result

Run	Input			Output	
	Glucose	BMI	Age	Diabetes Prediction	Ground Truth
1	105	34.9	25.0	No	No
2	188	32.0	22.0	Yes	Yes
3	123	34.1	28.0	No	No
4	163	31.6	28.0	Yes	Yes
5	142	24.7	21.0	No	No

Conclusion

The overall accuracy for this fuzzy inference system based on 392 samples from the diabetes dataset is 80.35%. To further improve the accuracy, knowledge from the domain expert is needed and a huge amount of data from different demographics are required for further analysis and more ideal fuzzification rules. Nevertheless, the output result of the diabetes prediction should only be used as a diagnosis reference or guideline, regardless of the accuracy of the prediction. This is because there exist more reliable medical diagnosis testings such as A1c Test and Fasting Plasma Glucose Test.

References

Akhtar, S., Nasir, J. A., Ali, A., Asghar, M., Majeed, R., & Sarwar, A. (2022). Prevalence of type-2 diabetes and Prediabetes in Malaysia: A systematic review and meta-analysis. *PLOS ONE*, *17*(1). https://doi.org/10.1371/journal.pone.0263139

Centers for Disease Control and Prevention. (2022, June 29). *National Diabetes Statistics Report*. Centers for Disease Control and Prevention. Retrieved January 14, 2023, from https://www.cdc.gov/diabetes/data/statistics-report/index.html

Coulman, K., & Toran, S. S. (2020, July 29). *Body mass index may not be the best indicator of our health – how can we improve it?* The Conversation. https://theconversation.com/body-mass-index-may-not-be-the-best-indicator-of-our-health-how-can-we-improve-it-143155

Freeth, A. (2021, September 22). *All About Diabetes: Causes, Myths, Treatment, and Home Remedies*. eMediHealth. https://www.emedihealth.com/glands-hormones/diabetes/manage-diabetes

Thompson, E. C., White-Schwoch, T., Tierney, A., & Kraus, N. (2015). Beat Synchronization across the Lifespan: Intersection of Development and Musical Experience. *PloS one*, *10*(6), e0128839. https://doi.org/10.1371/journal.pone.0128839

UCI MACHINE LEARNING. (2022, December). Pima Indians Diabetes Database. Retrieved October 7, 2016, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database