Interim Project Report_21034774D_CHAN Hou Ting Constant

by Chan Hou Ting Constant

General metrics

26,614	3,950	359	15 min 47 sec	30 min 23 sec
characters	words	sentences	reading time	speaking time

Score

89

This text scores better than 89% of all texts checked by Grammarly

Writing Issues

136	43	93
Issues left	Critical	Advanced

Plagiarism

This text hasn't been checked for plagiarism

Writing Issues

75	**Correctness**
4	Text inconsistencies
2	Conjunction use
1	Misplaced words or phrases
6	Confused words
3	Wrong or missing prepositions
2	Incomplete sentences
10	Incorrect phrasing
4	Determiner use (a/an/the/this, etc.)
15	Ungrammatical sentence
2	Incorrect verb forms
1	Pronoun use
5	Improper formatting
9	Misuse of semicolons, quotation marks, etc.
5	Punctuation in compound/complex sentences
3	Incorrect noun number
1	Faulty subject-verb agreement
1	Comma misuse within clauses
1	Misspelled words

49	**Clarity**
33	Passive voice misuse
7	Wordy sentences
9	Unclear sentences

10	**Delivery**	
9	Inappropriate colloquialisms	
1	Tone suggestions	
2	**Engagement**	
2	Word choice	

Unique Words

20%

Measures vocabulary diversity by calculating the percentage of words used only once in your document

unique words

Rare Words

35%

Measures depth of vocabulary by identifying words that are not among the 5,000 most common English words.

rare words

Word Length

5.2

Measures average word length

characters per word

Sentence Length

11

Measures average sentence length

words per sentence

Interim Project Report_21034774D_CHAN Hou Ting Constant

xviii

The Hong Kong Polytechnic University

Department of Electrical and Electronics Engineering

Project ID: [FYP_ lpchau_20240523150449]

Machine learning model to predict the risk of diabetes[1]

by

CHAN Hou Ting Constant

21034774D

Final Year Project - Interim Project Report 2024/2025 Sem 1

Bachelor of Science (Honours)

In

Internet and Multimedia Technologies

of

The Hong Kong Polytechnic University

Supervisor: Prof CHAU Lap Pui Date:27-December-2024

Abstract

Diabetes has become a noteworthy social issue in the world, where more and more people have been diagnosed with diabetes in recent years. To find out the problem, machine learning is one of the approaches used to predict diabetes. This project introduces two datasets, the Pima Indian Diabetes dataset and the NHANES dataset. In addition, a machine learning framework proposed by Tasin et al. [4] is the baseline model in this model. Moreover, polynomial regression and SMOTE are applied to predict missing values and class imbalance problems. Furthermore, Random Forest and XG Boost are used for the prediction of diabetes in this project, and Random Forest and XG Boost have the highest accuracy in the Pima Indian Diabetes dataset and the NHANES dataset, respectively. In the future, exploring other key features and preprocessing methods will be the major options to get better results for this project.

Contents

7

List of Figures

List of Tables

INTRODUCTION

Diabetes has been increasingly prevalent across the world for the past several years. It is estimated that currently, in 2022, around 8.3 million people [1] exhibit symptoms of diabetes, which comprises approximately 10.4% of the world's population, which makes ignoring diabetes impossible. Modern people have a fast life that we can hardly spare time to exercise, which leads to unhealthy living habits, such as obesity, sleep deprivation, etc. There are two main categories of common diabetes: Type 1 and Type 2. The former is congenital, hereditary, etc. [2], and the latter is its own acquired poor eating habits, lack of exercise, etc. [2]. Machine learning has been applied in diabetes prediction for years to predict if a patient is likely to develop diabetes. One of the advantages of machine learning is that it can generate the corresponding prediction based on the dataset's content. It allows people to make suitable decisions based on the predictions generated by machine learning algorithms. This project explores machine learning algorithms and applies them to diabetes predictions.

Overview

Background

One of the most significant issues in the world is Diabetes. As mentioned earlier, many people, whether adults, youth, or children, have had diabetes in recent years. Early for diabetes, people used to do blood glucose measurements to check their blood glucose levels to see if they had diabetes or not. In the past, a blood glucose meter was used. With the evolution of science and technology, another approach that is put forth[20] is the identification of diabetes[1] by using machine learning algorithms. People can view the predicted output generated by machine learning algorithms to check if they are at risk for diabetes[1]. It saves time for people who are allowed to[21] check their health conditions without viewing the indices from the body. Moreover, it decreases the probability of misjudgment due to human factors.

Problem Statement

Although diabetes prediction with machine learning has been implemented[22] recently, much research on diabetes prediction with machine learning algorithms has indicated different results. For example, Mujumdar and Vaidehi [3] indicated that Logistic Regression and Adaboost perform well with high accuracy in diabetes prediction. Tasin et al. [4] commented that XGBoost with the ADASYN approach performs well in diabetes prediction. It is difficult for people to judge whether machine learning models are the most suitable for diabetes prediction. Furthermore, much research has been done on diabetes prediction using machine learning algorithms and different datasets (NHANES and Pima Indian Diabetes) and data preprocessing[6] methods. Making the same topic on diabetes prediction with machine learning algorithms has caused different results. Therefore, this project is a good research topic to work on[23]. The results will be determined by collecting data from different sources, implementing the methodology, and comparing different methodologies.

Dataset

Pima Indian Diabetes Dataset

Pima Indian Diabetes Dataset

Features

Pregnancies

Glucose (2 hours in an oral glucose tolerance test (mg/dL))

BloodPressure (Diastolic blood pressure (mm Hg))

SkinThickness

Insulin (2-Hour[24] Serum insulin (μh/ml))

BMI

DiabetesPedigreeFunction

Age

Target

Outcome

Table 1: Information on Pima[25] Indian Diabetes Dataset

The Pima Indian Diabetes Database provided information about the patients who have Diabetes[1] or not. The dataset source comes from the National Institute of Diabetes and Digestive and Kidney Diseases [5]. A total of 768 patients were recorded[26] in the Pima Indian Diabetes Database, which are Pima Indians that are at least 21 years old females.

A total of 9 variables were listed[27] in the dataset, which included eight features and one target variable. Here is the explanation of these variables:

Pregnancies: It means the number of times pregnant

Glucose (Blood Sugar): It is a group of carbohydrates [6] that provides energy[28] for the body, and mg/dL is the measuring unit of glucose. If[29] the glucose is lower than 140 mg/dL, it is considered normal[29] [7].

BloodPressure[30]: It means heart[30] beats and pumps blood into the arteries [8]. Lack of exercise and obesity would result in Higher blood pressure, and it would cause health risks such as headache and dizziness.

SkinThickness: It estimates the body fat on thighs and limbs.

Insulin: It helps regulate blood sugar levels and is important for energy production and storage.

BMI: It measures body fat based on Height and Weight. 18.5 to 23 is considered a healthy weight and a normal body level.

BMI=WeightHeight2

DiabetesPedigreeFunction: It is a function that scores the probability of Diabetes[1] based on Family history.

Age: The age of all patients is at least 21 years old.

Outcome: A variable that diagnosed Diabetes[1] or not.

2013-2014 NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) is a project that the National Center for Health Statistics implemented. This project aims to collect data from American adults and children through interviews and body checks. NHANES collected dietary intake, physical examinations, and laboratory tests. Also, this project uses population-based sampling that includes the entire American population. This dataset is available for open access and widely used for health research and public health initiatives. Here is the abstract of the dataset:

NHANES Dataset

Features

Demographic

SEQN (ID of interviewee)

RIAGENDR (Gender)

RIDAGEYR (Age)

Diet

DR1DAY (Intake day of the week)

DR1TKCAL (Energy (kcal) take[31] in 1 day)

Examination

BMXBMI (BMI)

BPXDI1 (Blood Pressure)

Labs

LBXGLT (Glucose)

LBXIN (Insulin)

Questionnaire

DIQ010 (Diabetes_Diagnosis)

ALQ120Q (alcoholic drinks taken per day/ months)

Table 2: Abstract of NHANES Dataset

NHANES Dataset is divided[32] into five parts, which are[33] demographic[33], diet[33,34], examination, labs and questionnaire[33].

Demographic: it[36] means[36] the characteristics of a population, which include gender, age and[36] marital status, etc[35].

Diet: it means[37] the dietary intake information collected from the interviewees. Nutrient information like[39] Energy[28,39] taken, Vitamins[39], fats and carbohydrates are recorded[38,39] in the database.

Examination: it[40,41] means[40] the physical examinations and medical tests conducted on the interviewees, such as BMI and blood pressure.

Labs: it[42] means the laboratory tests performed on biological samples collected from the interviewees, such as glucose levels and Insulin.

Questionnaire: it[44] means[43] the self-reported information collected from the interviewees through structured interviews and surveys. It covers the topics[46][45] that related[46][45] to health and lifestyle like[45] physical activity and health conditions. The details of the data processing would[47] be explained[48] in the following section.

CURRENT PROGRESS

For this project, machine learning is a suitable approach to diabetes prediction because diabetes prediction belongs to a classification task that determines whether or not the patients are diagnosed with diabetes[1]. The process includes data preprocessing[6], feature selection, training, testing, and performance evaluation. All the work is done[49] on Jupyter Notebook, a web-based application that provides an interactive computing notebook environment to describe the data analysis.

Tasin et al. [4] proposed a machine-learning framework that acquired 81% accuracy using XGBoost. The preprocessing[6] methods are extreme gradient boosting techniques for filling the missing value; SMOTE and ADASYN are applied[50] to address the class imbalance issue. In addition, it collected the samples from 203 people called RTML, which is used for filling "Insulin" and is the merged dataset. In this project, it will be the baseline model for the reference.

Data Preprocessing[6] on Pima Indian Diabetes dataset:

Figure 1: Pima Indian Diabetes Dataset (raw data)

The dataset has 768 rows (participants) with nine columns (features) before preprocessing. As shown in Figure 1, there are some values of zero in columns "SkinThickness" and "Insulin." There are no null values on the dataset, so filling in zero is unnecessary.

Figure 2: Check missing value on Pima Indian Diabetes Dataset

In Figure 2, many missing values existed in columns "SkinThickness" and "Insulin," which count for about 30% and 50% of the dataset. Other columns "Glucose", "BloodPressure" and "BMI" with value 0 will be filled by their mean as they are only a tiny minority of the whole dataset. To identify the rows where column "SkinThickness" is zero, variables "zero_SkinThickness_rows" and "non_zero_SkinThickness_rows" are defined to find the rows where it is zero and non-zero, respectively. To predict column "SkinThickness" become more reliable, columns "Glucose", "BloodPressure" and "BMI" and "Age" are used to assist the prediction of column "SkinThickness".

Polynomial Regression is applied to predict the missing value of column "SkinThickness". Aditya Shastry et al. [9] applied polynomial regression to predict the missing value in data preprocessing and it was helpful to improve the model performance. First, the degree of the polynomial features is set to 2 and a bias column is not included in the polynomial features. Next, fit_transform() is applied to find the metrics of overall statistical properties (mean, standard deviation). Then, linear regression will be applied to train and predict the column "SkinThickness."

Figure 3: Fill the[61] missing values on "SkinThickness"

In Figure 3, the predicted values of the column "SkinThickness" are based on the input features (the columns "Glucose," "Blood Pressure," "Age," and "BMI"). The reason for not using the column "Insulin" as the input feature for the prediction of the column "SkinThickness" is that "Insulin" has lots of missing values, and it affected the predicted result that some of the predictions would generate negative values. Therefore, the approach that uses the column without missing value as the input feature is appropriate for predicting the features with the missing values.

Figure 4: Fill the missing values on "Insulin"

Figure 5: Check if there have[62] missing[63] value (0) or not

Figure 4 uses columns "Glucose," "Blood Pressure," "Age," "BMI," and "SkinThickness" as the input features for the prediction of "Insulin," which are the missing value. As shown in Figure 4, polynomial regression[4] did not generate negative values after the input features were used[64]. After the polynomial regression[4], replace all the predicted values with all the missing values. As shown in Figure 5, there is no missing value in each column, which means all the predicted values are successfully replaced[65].

Figure 6: All the missing values are replaced[66] by the predicted values

Figure 7: Quantile-Quantile Plot of all features (Pima Indian Diabetes Dataset)

As shown in Figure 7, there are[67] the normal[67] distribution of all features which[67] are displayed[68] in Q-Q plot form. Spots in features "Glucose," "Blood Pressure," "SkinThickness," and "BMI" roughly follow the straight red line on the plot. Also, the spots on the features "Insulin",[69] "Age," and "DiabetesPedigreeFunction" are deviated from the red straight line. For example, most of the spots in the feature "Age" are concentrated on 20, meaning most participants are about 21 years old.

Figure 8: Correlation matrix of all features and outcome[70] (Pima Indian Diabetes Dataset)

Figure 8 shows the correlation coefficient between all the features and the outcome. Based on the ranking in descending order, the relationship between the features and the outcome are "Glucose",[71] "Insulin",[72] "BMI",[73] "SkinThickness",[74] "Age",[75] "Pregnancies"[76] and "BloodPressure",[77] which the highest score and lowest score are "Glucose" and "BloodPressure" respectively. Cleveland Clinic [10] reported that glucose level[78] and diabetes[1] are strongly correlated. According to ranking, features "Glucose",[79] "Insulin",[79] "BMI"[79] and "SkinThickness" are selected as the key features for the prediction of diabetes.[1]

After defining the key features and "Outcome" as a class, the data will split into 80:20, 80% for training, and the rest for testing. Next, feature scaling is applied[80] to split data. Standardization is used in the feature scaling to normalize the data. In the standardization process,[81] the scaler will be defined and fit into the training data. Then, the testing data will be transformed[82] to finish the feature scaling.

Figure 9: Feature Importance (Pima Indian Diabetes dataset)

Figure 10: Balanced data (Pima Indian Diabetes Dataset)

As shown in Figure 9, "Glucose" has the highest rank in feature importance after standardization. Furthermore, the features rank the same as the correlation matrix. Then, SMOTE is applied[83] to process imbalance data, which can significantly improve model performance, especially for weak learners [11]. In Figure 10, all the data are balanced at 50:50 after applying SMOTE.

Data Preprocessing on NHANES dataset:

Since the NHANES dataset has five raw data sets, it is necessary to preprocess the data to ensure that it is readable and understandable to people because there are many features in each raw data set. Originally, medication (not mentioned in Table 1) was one of the raw data of the NHANES dataset, but no complete description can be found on the NCHS official website. In addition, diet does not have similar features compared to Pima Indian Diabetes dataset. Therefore, medication and diet will not be used in data preprocessing.

Figure 11: Information about demographic (NHANES dataset)

In Figure 11, each feature is complex to read for people as they are named in terms. To make the data more readable and understandable features similar to the Pima Indian Diabetes dataset are selected and relabeled. For example, "SEQN", "RIAGENDR" and "RIDAGEYR" are relabeled as "ID", "Gender" and "Age" respectively. After the relabeling, all the data are merged on the "ID" and uses a left join. Then, category mapping is created for gender and split into Male and Female, and categorical variables are converted into dummy variables.

Figure 12: Merged dataset (NHANES dataset)

Figure 13: Total number of each row with value 0

In Figure 12, the merged dataset collects the key features and is renamed. Despite the conversation, many null values are still on the merged dataset. Therefore, fill the null values with zero in the entire dataset and apply polynomial regression[4] to predict these values. In this situation, it is not suitable to use means or median to replace the null values as there are many null values in Figure 13, and this will negatively influence the prediction of diabetes[1] if applied.

Figure 14: Merged dataset after Polynomial Regression applied (NHANES dataset)

After the polynomial regression[4] applied[94], the predicted values replace the null values in the merged dataset and[94] it looks[95] understandable. In column "Outcome", there are five values (1 for Yes, 2 for No, 3 for Borderline, 7 for Refused, and 9 for Do not know). Values 1 and 2 are retained[96], and the rest are removed[97] because of their unclear result. Furthermore, the proportion of "Outcome" larger[98] than two in the dataset is 2%; consequently, removing these will not significantly impact the prediction of diabetes[1,99].

In addition, it is counterintuitive to use values 1 and 2 to represent having diabetes[1,100] and not having diabetes[1], especially since the value 2[101] will mistake people for invalid. Accordingly, it is necessary to use 1 and 0 to represent having diabetes[1] and no diabetes instead of 1 and 2.

Figure 15: Quantile-Quantile Plot of all features (NHANES dataset)

Figure 16: Correlation matrix of all features and outcome[102] (NHANES dataset)

In Figure 15, "Blood pressure," "Glucose," and "Insulin" are the normal distribution, and most of the spots on these features follow the red straight line. Additionally, "Age" looks like an inverse z shape. Most spots deviate from two ranges and can be considered a bimodal distribution.

In Figure 16, put the relationship between features and "Outcome" in descending order are[103] "Age",[103] "BMI",[103] "Insulin",[103] "BloodPressure",[103] "Glucose"[103] and "ID".[103] "ID" will not be used as the input feature because it is a negative for "Outcome".[104] In addition, "ID" is used to assign the numeric labels to each entry, which means it does not have a meaningful relationship with "Outcome".[105] Moreover, the result of the correlation matrix is different from the Pima Indian Diabetes dataset in that "Age" is the highest score in the NHANE dataset, but "Glucose" is the highest score in the Pima Indian Diabetes dataset.

Figure 17: Feature Importance (NHANES dataset)

In Figure 17, "Glucose" has the highest feature importance after standardization, the same rank as the Pima Indian Diabetes dataset. Also, it is worthy[106] to note[106] that "Insulin" is at the bottom of the feature importance. In the preprocessing[6], "Glucose",[107] "Age",[107] "BMI",[107] "BloodPressure"[107] and "Insulin" are selected as input features in order to make a similar environment to make the comparison.

Preliminary Result

Figure 18: Reproduce Result (XGB+ADASYN)

Figure 19: Paper Result (XGB+ADASYN)

Figure 20: Reproduce Result (AUC)

Figure 21: Paper Result (AUC)

Despite the same dataset and environment, it is found[108] that the results of the XG Boost Classifier are different, as shown in the figures above.

Figure 23: Preliminary Result (Random Forest) (Pima Indian Diabetes dataset)`

Figure 22: Preliminary Result (XG Boost) (Pima Indian Diabetes dataset)

Figure 24: Preliminary Result (XG Boost) (NHANES dataset)

Figure 25: Preliminary Result (Random Forest) (NHANES dataset)

Since the referenced paper proposed a machine learning framework that used XG Boost with ADASYN and other machine learning models like random forest for the comparison in the paper, XG Boost and Random Forest are used to predict diabetes in this project. For the Pima Indian Diabetes dataset, Random Forest obtains 77.9% accuracy, and XG Boost is the highest. It has 86.1% accuracy for the NHANES dataset. Adjustments have been implemented in the Pima Indian Diabetes dataset, such as the input features change to "Glucose", "Insulin", "BMI", "BloodPressure" and "Age", which are the same as the input features in the NHANES dataset. The reason for the different results in the two datasets can be the number of entries in the NHANES dataset, which is around ten thousand.

CHALLENGES

Data preprocessing for the NHANES dataset is challenging as it includes five different data sources. Also, each data has lots of features whether it is useful for diabetes prediction or not. Therefore, I need to read the description of the features to ensure that I[114] can remove the redundancy and[115,116] get the key features from each data and[117] merge them into a new data frame with filtered. Moreover, there are many missing values with 0[118] and null values on the Pima Indian Diabetes dataset and NHANES dataset[119]. For example, a number[120] of[121] rows with "LBXGLT" and "LBXIN" (value 0) have 7830 and 7082, respectively, which comprises approximately 77% and 70% of the whole NHANES dataset. The approach of filling value zero with the means is used[122] to predict the missing value but it performs poorly in the q-q plot as both are not[123] the[123] normal distribution. Therefore, polynomial[124] regression approach is used in this situation to predict the missing value and[124] it is better than filling value[124] zero with the means.

FUTURE WORKS

1. Try to use other preprocessing[6] methods

Standardization and SMOTE are applied as the data preprocessing[6] approaches in this stage. For feature scaling, methods like min-max scaling and mean normalization will be the options to transform the data to fit within a specific range or scale. For the class imbalance in datasets, techniques like ADASYN will be the possible choice to ensure that the model will not lead to poor performance in the minority class.

2. Try to select other key features for comparison

In this stage, "Glucose",[125] "Insulin",[125] "BMI",[125] "SkinThickness"[125] are the selected features in the Pima Indian Diabetes dataset. The criteria for selecting the input features are based[126] on the correlation matrix, which is lower than 0.2, to ensure that no[127] more redundancy features as the input features affect the model performance. For the NHANES dataset, "Glucose",[128] "Insulin",[128] "BMI",[128] "Age"[12] and "BloodPressure"[128] are the input features in[128] the dataset[128]. In the future, it is possible to[129] explore more features as input features and make comparisons to review which performs better under the same preprocessing[6] method.

References

World Health Organization, Diabetes. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed Dec. 09, 2024].

Smart Patient, Diabetes Mellitus. [Online]. Available: https://www.smartpatient.ha.org.hk/en/smart-patient-web/disease-management/disease-information/disease/DiabetesMellitus [Accessed Dec. 09, 2024].

A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in 2ND INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING ICRTAC -DISRUP - TIV INNOVATION ,[130] 2019, AMSTERDAM: Elsevier B.V, 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047 [Accessed Dec. 09, 2024]

I.[131] Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," Healthcare technology[132] letters, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039 [Accessed Dec. 09, 2024]

UCI Machine Learning and Kaggle Team, Pima Indians Diabetes Database, 2016. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Accessed Dec. 09, 2024].

Healthline ,[133] Everything You[134] Need to Know About Glucose, 2024. [Online]. Available: https://www.healthline.com/health/glucose [Accessed Dec. 09, 2024].

E. Eyth, H. Basit and C.J. Swift, "Glucose Tolerance Test, "in StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2024. Available:

https://www.ncbi.nlm.nih.gov/books/NBK532915/#__NBK532915_dtls__
[Accessed Dec. 09, 2024]

Cleveland Clinic, Blood Pressure. [Online]. Available:
https://my.clevelandclinic.org/health/diagnostics/17649-blood-pressure
[Accessed Dec. 09, 2024].

K. Aditya Shastry et al., "Regression Based[135] Data Pre-processing[6] Technique for Predicting Missing Values," in Emerging Research in Computing, Information, Communication and Applications, Singapore: Springer Singapore Pte. Limited, 2021, pp. 95–102. doi: 10.1007/978-981-16-1338-8_9

Cleveland Clinic, Blood Glucose (Sugar) Test. [Online]. Available:
https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test
[Accessed Dec. 11, 2024].

Train In Data, SMOTE in Python: A guide to balanced datasets. [Online]. Available: https://www.blog.trainindata.com/smote-in-python-a-guide-to-balanced-datasets/ [Accessed Dec. 11, 2024].

Appendices

Appendix 1: Pima Indian Diabetes dataset[136]

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Appendix 2: 2013-2014 NHANES dataset

https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey/data?select=diet.csv

1.	*diabetes; Diabetes*	Text inconsistencies	Correctness
2.	~~where~~ → and	Conjunction use	Correctness
3.	*To find out the problem*	Misplaced words or phrases	Correctness
4.	*regression; Regression*	Text inconsistencies	Correctness
5.	*are used*	Passive voice misuse	Clarity
6.	*preprocessing; Preprocessing; preprocess; Pre-processing*	Text inconsistencies	Correctness
7.	~~dataset~~ → Dataset	Confused words	Correctness
8.	~~dataset~~ → Dataset	Confused words	Correctness
9.	~~dataset~~ → Dataset	Confused words	Correctness
10.	in the	Wrong or missing prepositions	Correctness
11.	in the	Wrong or missing prepositions	Correctness
12.	is have	Incomplete sentences	Correctness
13.	~~have missing~~ → a missing	Incorrect phrasing	Correctness
14.	*are replaced*	Passive voice misuse	Clarity
15.	*etc*	Inappropriate colloquialisms	Delivery
16.	*etc.*	Inappropriate colloquialisms	Delivery
17.	*etc.*	Inappropriate colloquialisms	Delivery
18.	~~or not~~	Wordy sentences	Clarity
19.	*was used*	Passive voice misuse	Clarity

20.	*is put forth*	Passive voice misuse	Clarity
21.	~~are allowed to~~ → can	Wordy sentences	Clarity
22.	*been implemented*	Passive voice misuse	Clarity
23.	*on*	Inappropriate colloquialisms	Delivery
24.	~~2-Hour~~ → 2-hour	Confused words	Correctness
25.	the Pima	Determiner use (a/an/the/this, etc.)	Correctness
26.	*were recorded*	Passive voice misuse	Clarity
27.	*were listed*	Passive voice misuse	Clarity
28.	*energy; Energy*	Text inconsistencies	Correctness
29.	*If the glucose is lower than 140 mg/dL, it is considered normal [7].*	Unclear sentences	Clarity
30.	*BloodPressure: It means heart beats and pumps blood into the arteries [8].*	Ungrammatical sentence	Correctness
31.	~~take~~ → taken	Incorrect verb forms	Correctness
32.	*is divided*	Passive voice misuse	Clarity
33.	*NHANES Dataset is divided into five parts, which are demographic, diet, examination, labs and questionnaire.*	Incorrect phrasing	Correctness
34.	*NHANES Dataset is divided into five parts, which are demographic, diet, examination, labs and questionnaire.*	Unclear sentences	Clarity
35.	*etc*	Inappropriate colloquialisms	Delivery

36.	*Demographic: it means the characteristics of a population, which include gender, age and marital status, etc.*	Incorrect phrasing	Correctness
37.	~~it means~~	Incorrect phrasing	Correctness
38.	*are recorded*	Passive voice misuse	Clarity
39.	*Nutrient information like Energy taken, Vitamins, fats and carbohydrates are recorded in the database.*	Incorrect phrasing	Correctness
40.	*Examination: it means the physical examinations and medical tests conducted on the interviewees, such as BMI and blood pressure.*	Incorrect phrasing	Correctness
41.	*Examination: it means the physical examinations and medical tests conducted on the interviewees, such as BMI and blood pressure.*	Unclear sentences	Clarity
42.	*Labs: it means the laboratory tests performed on biological samples collected from the interviewees, such as glucose levels and Insulin.*	Unclear sentences	Clarity
43.	~~it means~~ → This means	Pronoun use	Correctness
44.	*Questionnaire: it means the self-reported information collected from the interviewees through structured interviews and surveys.*	Unclear sentences	Clarity
45.	*It covers the topics that related to health and lifestyle like physical activity and health conditions.*	Ungrammatical sentence	Correctness
46.	*It covers the topics that related to health and lifestyle like physical activity and health conditions.*	Unclear sentences	Clarity

47.	~~would~~ → will	Incorrect verb forms	Correctness
48.	*be explained*	Passive voice misuse	Clarity
49.	*is done*	Passive voice misuse	Clarity
50.	*are applied*	Passive voice misuse	Clarity
51.	*There are no null values on the dataset, so filling in zero is unnecessary.*	Unclear sentences	Clarity
52.	the Pima	Determiner use (a/an/the/this, etc.)	Correctness
53.	*Other columns "Glucose", "BloodPressure" and "BMI" with value 0 will be filled by their mean as they are only a tiny minority of the whole dataset.*	Ungrammatical sentence	Correctness
54.	~~0~~ → zero	Improper formatting	Correctness
55.	*To predict column "SkinThickness" become more reliable, columns "Glucose", "BloodPressure" and "BMI" and "Age" are used to assist the prediction of column "SkinThickness".*	Ungrammatical sentence	Correctness
56.	~~".~~ → ."	Misuse of semicolons, quotation marks, etc.	Correctness
57.	*Aditya Shastry et al. [9] applied polynomial regression to predict the missing value in data preprocessing and it was helpful to improve the model performance.*	Incorrect phrasing	Correctness
58.	*is set*	Passive voice misuse	Clarity
59.	, and	Punctuation in compound/complex sentences	Correctness

60.	*is not included*	Passive voice misuse	Clarity
61.	in the	Wrong or missing prepositions	Correctness
62.	is have	Incomplete sentences	Correctness
63.	~~have missing~~ → a missing	Incorrect phrasing	Correctness
64.	*were used*	Passive voice misuse	Clarity
65.	*are successfully replaced*	Passive voice misuse	Clarity
66.	*are replaced*	Passive voice misuse	Clarity
67.	*As shown in Figure 7, there are the normal distribution of all features which are displayed in Q-Q plot form.*	Ungrammatical sentence	Correctness
68.	*are displayed*	Passive voice misuse	Clarity
69.	~~";~~ → ,"	Misuse of semicolons, quotation marks, etc.	Correctness
70.	~~outcome~~ → outcomes	Incorrect noun number	Correctness
71.	~~";~~ → ,"	Misuse of semicolons, quotation marks, etc.	Correctness
72.	~~";~~ → ,"	Misuse of semicolons, quotation marks, etc.	Correctness
73.	~~";~~ → ,"	Misuse of semicolons, quotation marks, etc.	Correctness
74.	~~";~~ → ,"	Misuse of semicolons, quotation marks, etc.	Correctness
75.	~~";~~ → ,"	Misuse of semicolons, quotation marks, etc.	Correctness

76.	Pregnancies,	Punctuation in compound/complex sentences	Correctness
77.	",	Punctuation in compound/complex sentences	Correctness
78.	~~level~~ → levels	Incorrect noun number	Correctness
79.	*According to ranking, features "Glucose", "Insulin", "BMI" and "SkinThickness" are selected as the key features for the prediction of diabetes.*	Ungrammatical sentence	Correctness
80.	*is applied*	Passive voice misuse	Clarity
81.	*In the standardization process, the scaler will be defined and fit into the training data.*	Unclear sentences	Clarity
82.	*be transformed*	Passive voice misuse	Clarity
83.	*is applied*	Passive voice misuse	Clarity
84.	*are balanced*	Passive voice misuse	Clarity
85.	*be found*	Passive voice misuse	Clarity
86.	the Pima	Determiner use (a/an/the/this, etc.)	Correctness
87.	*be used*	Passive voice misuse	Clarity
88.	*are named*	Passive voice misuse	Clarity
89.	*To make the data more readable and understandable features similar to the Pima Indian Diabetes dataset are selected and relabeled.*	Incorrect phrasing	Correctness

90.	*For example, "SEQN", "RIAGENDR" and "RIDAGEYR" are relabeled as "ID", "Gender" and "Age" respectively.*	Ungrammatical sentence	Correctness
91.	~~uses~~ → use	Faulty subject-verb agreement	Correctness
92.	*are converted*	Passive voice misuse	Clarity
93.	*is renamed*	Passive voice misuse	Clarity
94.	*After the polynomial regression applied, the predicted values replace the null values in the merged dataset and it looks understandable.*	Ungrammatical sentence	Correctness
95.	*After the polynomial regression applied, the predicted values replace the null values in the merged dataset and it looks understandable.*	Unclear sentences	Clarity
96.	*are retained*	Passive voice misuse	Clarity
97.	*are removed*	Passive voice misuse	Clarity
98.	~~larger~~ → more significant, more extensive, more prominent	Word choice	Engagement
99.	diabetes prediction	Wordy sentences	Clarity
100.	~~diabetes~~	Wordy sentences	Clarity
101.	~~2~~ → two	Improper formatting	Correctness
102.	~~outcome~~ → outcomes	Incorrect noun number	Correctness
103.	*In Figure 16, put the relationship between features and "Outcome" in descending order are "Age", "BMI", "Insulin", "BloodPressure", "Glucose" and "ID".*	Ungrammatical sentence	Correctness

104.	~~".~~ → ."	Misuse of semicolons, quotation marks, etc.	Correctness
105.	~~".~~ → ."	Misuse of semicolons, quotation marks, etc.	Correctness
106.	*Also, it is worthy to note that "Insulin" is at the bottom of the feature importance.*	Ungrammatical sentence	Correctness
107.	*In the preprocessing, "Glucose", "Age", "BMI", "BloodPressure" and "Insulin" are selected as input features in order to make a similar environment to make the comparison.*	Ungrammatical sentence	Correctness
108.	*is found*	Passive voice misuse	Clarity
109.	*been implemented*	Passive voice misuse	Clarity
110.	*Adjustments have been implemented in the Pima Indian Diabetes dataset, such as the input features change to "Glucose", "Insulin", "BMI", "BloodPressure" and "Age", which are the same as the input features in the NHANES dataset.*	Ungrammatical sentence	Correctness
111.	, whether	Punctuation in compound/complex sentences	Correctness
112.	~~useful~~ → helpful	Word choice	Engagement
113.	*I*	Inappropriate colloquialisms	Delivery
114.	*I*	Inappropriate colloquialisms	Delivery
115.	~~and~~	Conjunction use	Correctness
116.	and,	Punctuation in compound/complex sentences	Correctness

117.	, and	Comma misuse within clauses	Correctness
118.	~~0~~ → zero	Improper formatting	Correctness
119.	and NHANES datasets	Wordy sentences	Clarity
120.	~~a number~~ → the number	Determiner use (a/an/the/this, etc.)	Correctness
121.	~~a number of~~ → several, some, many	Wordy sentences	Clarity
122.	*is used*	Passive voice misuse	Clarity
123.	*The approach of filling value zero with the means is used to predict the missing value but it performs poorly in the q-q plot as both are not the normal distribution.*	Incorrect phrasing	Correctness
124.	*Therefore, polynomial regression approach is used in this situation to predict the missing value and it is better than filling value zero with the means.*	Ungrammatical sentence	Correctness
125.	*In this stage, "Glucose", "Insulin", "BMI", "SkinThickness" are the selected features in the Pima Indian Diabetes dataset.*	Ungrammatical sentence	Correctness
126.	*are based*	Passive voice misuse	Clarity
127.	~~that~~ no	Wordy sentences	Clarity
128.	*For the NHANES dataset, "Glucose", "Insulin", "BMI", "Age" and "BloodPressure" are the input features in the dataset.*	Ungrammatical sentence	Correctness
129.		Tone suggestions	Delivery
130.	INNOVATION ,	Improper formatting	Correctness

131.	*I*	Inappropriate colloquialisms	Delivery
132.	~~technology~~ → Technology	Confused words	Correctness
133.	Healthline ,	Improper formatting	Correctness
134.	*You*	Inappropriate colloquialisms	Delivery
135.	Regression-Based	Misspelled words	Correctness
136.	~~dataset~~ → Dataset	Confused words	Correctness