WQD7001 PRINCIPLES OF DATA SCIENCE

GA2

Predictive Modeling for Early Intervention: Developing a Diabetes Risk Classification System

Name	Matrix No.	Roles
Tan Kai Qi	23051355	Group Leader
Alicia Chua Yang Qin	23060317	Oracle
Che Jia Min	23053571	Detective
Ng Chee Kang	23051336	Secretary
Sang Yong Han	22087278	Maker

Contents

1. Executive Summary 1
2. Data Modeling 1
   Packages, Dataset Splitting and Model Building 1
   Cross-Validation, Random Forest Model and Features Importance Analysis 2
3. Data Interpretation 5
   Data splitting 5
   Logistic Regression 5
   10-fold Cross-validation 6
   Random Forest 6
   Fine-tuning of Random Forest parameters 7
   Feature importance 8
4. Data Product 9
   Deployment 9
   End User Feedback 9
5. Plan For Reproducible Research 10
   Documentation of Research Process 10
   Code Availability 11
6. Insights And Conclusion 12
7. Appendixes 14
8. References 14

1. Executive Summary

This project with title “Predictive Modeling for Early Intervention: Developing a Diabetes Risk Classification System” is aimed to identify a robust classification model for predicting an individual's risk of developing diabetes based on a set of provided attributes, enable early intervention by identifying individuals at a high risk of diabetes and identify prevention efforts by providing actionable insights to individuals, healthcare professionals, and public health authorities. This project focuses on utilizing predictive modeling methodologies to categorize individuals into different categories of diabetes risks.

In the first part of the project, we conducted a complete Exploratory Data Analysis (EDA) that paved the way for the current phase of our project. Now, we will perform the modeling for our dataset. We will explain the steps of modeling and the rationale behind each step.

Upon importing the dataset from 'diabetes_data.csv', we focused on performing basic checks to confirm the structure and completeness of the data to avoid any negative impact on the model that will be built in the next step. In this part, we use str() and is.na() to confirm the completeness of the dataset.

The variables selected for our model, including gender, polyuria, polydipsia, sudden weight loss, polyphagia, irritability, and partial paresis, were chosen based on their relevance in the context of diabetes risk factors. These variables are indicated as significant indicators in our EDA. Therefore, we consider them to be foundational elements in our predictive modeling.

Our data is binary categorical data. At the same time, logistic regression is proficient in handling binary classification problems, making it an ideal tool for analyzing the relationships between our selected features and the probability of diabetes. Based on these conditions, logistic regression is the most suitable modeling technique to help us achieve the objective.

In addition, this report will delve into the development of the proposed model for diabetes prediction, the data product and the feedback of end users of the product.

2. Data Modeling

Packages, Dataset Splitting and Model Building

Apart from the basic packages that we commonly use in R, we do use some other packages in this modelling step, including “caret”, “MASS”, “pROC”, “randomForest”, and “lime”. These packages help us to perform the complete modelling step.

Before model development, we meticulously prepared our dataset. We divided it into an 80% training set and a 20% testing set through the “createDataPartition” function. To ensure consistent splitting, we set a random seed of 1. Moreover, we transformed the gender variable into a factor to suit our analysis's needs and make it more understandable.

Our first logistic regression model included all the selected features. This full feature model highlighted the significant impact of variables such as gender, polyuria, polydipsia, and irritability in predicting diabetes, and it also underlined their critical role in our analysis.

The performance of this initial model on the testing set was good. It achieved an accuracy of 90.38% and an Area Under the Curve (AUC) of 0.971, showcasing its predictive solid capability. Specifically, the model displayed a balanced accuracy of 89.74%, sensitivity of 87.18%, and specificity of 92.31%, showing its effectiveness in distinguishing between diabetic and non-diabetic cases.

We developed a second simplified model to build a better model with higher prediction accuracy by removing two variables with lower correlations to diabetes - polyphagia and partial paresis. We aimed to test the robustness of our model with fewer predictors. This model also provided impressive results, coming with high accuracy and AUC as the first model, with an accuracy of 92.31% and an AUC of 0.9718.

Cross-Validation, Random Forest Model and Features Importance Analysis

Both logistic regression models showed excellent results in prediction of diabetes. To further validate the results obtained from our two logistic regression models, we applied a 10-fold cross-validation technique.

This method divides the dataset into ten equal parts, using each part as a testing set while the remaining nine parts serve as the training set. Such a comprehensive approach allows us to assess the models' performance across various subsets of data, ensuring that our results are not just a consequence of a particular data split. This step is essential for robust verification of our models' predictive accuracy. The average AUC values obtained from this process – 0.9623517 for Model One and 0.9604677 for Model Two – showed the stability of our logistic regression approach.

Our logistic regression models provided exceptional results. Both showcased high accuracy and Area Under the Curve (AUC) scores. However, our team decided to explore further. We understand the potential for even more refined models offering better predictive capabilities.

This pursuit of excellence led us to experiment with additional modelling techniques which extended our scope beyond logistic regression. The rationale behind this decision was to uncover any other model that might predict diabetes more effectively to ensure we had explored different aspects to achieve the best possible outcome in our predictive analysis.

Our exploration shifted towards random forest modelling. The reason for choosing this is that the characteristics of random forest match our objective, and it can also manage complex datasets and reduce overfitting. We built two versions of random forest models: model_rf_full with a complete set of features and model_rf_simplified with a reduced feature set. The initial configurations with 500 trees yielded an accuracy of 0.9231 for the full feature model and 0.9135 for the simplified model. Subsequently, we increased the number of trees to 600, which maintained the same level of accuracy for the full feature model at 0.9231; the accuracy for the simplified model also remained consistent at 0.9135.

To further improve the random forest models is to adjust the mtry parameter. We tried to manipulate the mtry to examine the model's results by applying mtry 1, 3, 5, 7 to the full variables random forest and 1, 3, 5 mtry for the simplified random forest model. In the manipulation of the mtry, we observed the following results: For the full-feature model with 500 trees, the accuracies for mtry settings of 1, 3, 5, and 7 were 90.38%, 92.31%, 94.23%, and 94.23%. When increased to 600 trees, these mtry settings showed accuracies of 89.42%, 92.31%, 94.23%, and 94.23%. For the simplified feature model, with 500 trees, the accuracies for mtry settings of 1, 3, and 5 were 91.35%. At 600 trees, these mtry values resulted in 89.42%, 91.35%, and 91.35% accuracy. These data showed the impact of different numbers of trees and mtry settings on the accuracy of the models. When using 500 trees as a baseline, the full-feature model achieved the highest accuracy with mtry values of 5 and 7. In contrast, in the simplified feature model, mtry values of 3 and 5 showed equivalent levels of efficacy.

At the end of the modelling part, we performed the feature importance analysis of our models. Feature importance analysis is essential for our logistic regression and random forest models. In logistic regression, we evaluated feature importance through the coefficient estimates in the model's results. This analysis clearly showed how each variable influenced the outcome, with significant coefficients for features like gender, polyuria, polydipsia, and irritability. For the random forest models, we used the "varImpPlot" function, which visually describes the importance of each variable and offers an additional layer of understanding about the factors most influential in predicting diabetes.

In this section of our report, we have explained the selection and evaluation of the logistic regression model. This model was validated using 10-fold cross-validation to examine its robustness and reliability in predicting diabetes. However, for better predictive performance, we explored the realm of random forest modelling. We adjusted various parameters and fine-tuned our random forest models to achieve greater precision in predictions. Besides, we also conducted feature importance analysis for both logistic regression and random forest models. This analysis provided valuable insights into which variables were most influential in predicting diabetes. It enhanced our understanding of diabetes. In the next part of our report, we will focus on discussing the results displayed by our models. This will include a detailed analysis of the predictive accuracy and the significance of different variables.

3. Data Interpretation

In the previous part of the project, we performed data modelling to predict diabetes based on our dataset. We utilised two different machine learning models, namely, Logistic Regression and Random Forest. In this section of the project, we will interpret the results obtained from the modelling phase and relate the model outcomes to our project objectives.

Data splitting

Before building our machine learning models, we split our dataset into training and testing sets with a ratio of 80% to 20%. This split ratio is used due to our dataset being relatively small, with only 520 records. Therefore, setting aside a larger portion (80%) of the dataset as the training set can be advantageous as there is sufficient data for the model to learn from. The 80-20 split is considered a standard as it is widely employed by many.

Logistic Regression

We utilized Logistic Regression (LR) as one of the modelling techniques in our project. We developed two LR models, with the split ratio mentioned above. The first LR model contained all the selected features. The selected features include gender, polyuria, polydipsia, sudden weight loss, polyphagia, irritability, and partial paresis. These features were selected as our EDA showed that they have a strong correlation with diabetes. Performance metrics were used to assess and evaluate the performance of the model. The first LR model obtained a relatively good predictive performance on the testing set with an accuracy of 90.38%. This indicates that the model predicted most of the instances correctly. The model was also able to identify most positive instances and negative instances correctly as it displayed a relatively high sensitivity and specificity, which are 87.18% and 92.31%, respectively. There are 320 positive cases and 200 negative cases in our dataset. The balanced accuracy score provides a more realistic idea on the model performance. Our model achieved a balanced accuracy of 89.74% which indicates that the model is robust and not biased as it is able to make accurate predictions for both majority and minority classes. The high AUC of 0.971 further proves that the model is able to differentiate between the positive and negative classes (Erickson & Kitamura, 2021).

Feature importance for the LR model was evaluated by analysing the coefficient estimates and p-values of the features. Feature importance is crucial in identifying the features that are most significant in the prediction of diabetes. Our analysis showed that two features, namely, polyphagia and partial paresis have small coefficient estimates and large p-values, indicating that they might not be statistically significant, thus not being strong predictors of the target variable. This may be because these symptoms can also be related to other medical conditions or diseases

other than diabetes. Therefore, by removing these two features, a second simplified LR model was built to obtain a higher prediction accuracy. Our findings showed that better results were obtained for all performance metrics. The accuracy of the second simplified model showed an increase from 90.38% to 92.31%. This may be due to the removal of nonsignificant features, resulting in noise reduction in the model. The improvement of the second model in terms of different performance metrics also demonstrated the robustness of the model with fewer predictors. As a result, we are able to achieve our objective of identifying a robust classification model for diabetes prediction.

10-fold Cross-validation

Then, we further assessed and validated the performance of our LR models by using the 10-fold cross-validation technique. This technique plays an important role in determining the robustness as well as the generalisation ability of our models as it allows the evaluation of the model across various different data subsets. In our project, we chose to use 10-fold because a higher number of folds might be advantageous for smaller datasets as it allows for a more extensive usage of the available data, thus creating diverse subsets for the training and testing sets. This approach is also considered a common standard as it is often employed. We have obtained relatively high mean AUC values for both LR models, 0.9623517 and 0.9604677 for Model 1 and 2, respectively. This suggests that both models were able to perform well on different data subsets. In other words, it shows that the performance of the LR models is not strongly affected by a specific subset of the data and thus validating the robustness and generalisation ability of our models.

Random Forest

We then explored further by employing Random Forest (RF) as the second modelling technique in our project. Firstly, we built two RF models similarly to the LR models. The first RF model consists of all the selected features and the second simplified RF model that excludes polyphagia and partial paresis. Both models performed well as they achieved high accuracy levels with an initial configuration of 500 trees. Unlike our LR models, the second simplified RF model (91.35%) showed a slight decrease in accuracy compared to the first full-featured RF model (92.31%). As RF is effective in capturing the relationships between variables, removing some of the variables may cause information loss. The simplified LR model and full-featured RF model demonstrated the highest accuracy which is 92.31%. This may be due to the complexity of the LR and RF models whereby removing features in LR might simplify the model and improve

accuracy whereby in RF, it might lead to the inability to capture certain relationships and patterns, resulting in loss of information (Fox et al., 2017).

Fine-tuning of Random Forest parameters

Fine-tuning of the RF models was then conducted by adjusting the number of trees and the mtry parameter. Firstly, the number of trees was increased to 600 to evaluate the performance and stability of the RF approach. The results showed that the full-featured and simplified models with 600 trees had the same accuracy levels as those with 500 trees. This suggests that increasing the number of trees does not affect the accuracy of the models. This may be because the performance has reached a plateau, as the improvement in AUC by increasing the number of trees is minimal (Probst & Boulesteix, 2018). The accuracy of the LR and RF models are shown in Table 1.

Table 1: Accuracy of Logistic Regression and Random Forest models

Models	**Accuracy**
Full-featured LR model	90.38%
Simplified LR model	92.31%
Full-featured RF model (500 trees)	92.31%
Simplified RF model (500 trees)	91.35%
Full-featured RF model (600 trees)	92.31%
Simplified RF model (600 trees)	91.35%

Then, different mtry values were used to analyze their influence on the performance of the model. As the accuracy of RF models with 500 and 600 trees showed no difference, we will discuss the effect of mtry values on the accuracy of models with 500 trees. Our study found that mtry values of 5 and 7 on the full-featured model obtained the highest accuracy (94.23%). On the other hand, mtry values 3 and 5 on the simplified model demonstrated equivalent performance in terms of accuracy (91.35%).

The Out-of-Bag (OOB) error is most commonly used as a metric to determine the best mtry value, with the optimal mtry value having the smallest OOB error. A low OOB error rate indicates that the RF model has a good prediction performance. It also indicates that the model

is able to perform well on new, unseen data, reflecting a good generalization ability (Janitza & Hornung, 2018). Although mtry values of 5 and 7 obtained the highest accuracy for the full-featured RF model, mtry value of 7 is considered more optimal because it has a lower OOB error. As for the simplified RF model, mtry value of 3 is considered better than mtry value of 5 because it has a lower OOB error. By determining the optimal mtry value with the lowest OOB error, overfitting can be reduced.

Feature importance

As mentioned above, feature importance analysis was conducted for both LR and RF models. For our LR model, we identified a few features that are significant predictors of diabetes according to their coefficient estimates and p-values which are gender, polyuria, polydipsia and irritability. Polyuria and polydipsia are significant indicators of diabetes because the kidney is highly affected by this disease. High sugar levels in the blood put a lot of pressure on the kidneys in which the kidneys try to eliminate by producing more urine. This in turn leads to dehydration, resulting in thirst. Unstable blood sugar levels can also affect our mood and feelings resulting in symptoms like irritability. The relationship between gender and diabetes may be linked to the amount of visceral fat in the body (Oladimeji et al., 2021). As for our RF models, we created a variable importance plot to visualize the importance of each feature in the RF model. The plot is created based on the mean decrease in Gini impurity for each feature. A higher value suggests that the feature is important and contributes significantly to the predictive performance of the model.

The findings for the RF model are quite similar with the LR model with the exception whereby the feature, “irritability”, has a lower significance compared to features like partial paresis and sudden weight loss. The importance of each feature may be different depending on the type of machine learning model employed. This is because different models have their specific methods of determining the relationships between the features and the response variable. Random Forest has the ability to capture more complex relationships compared to logistic regression, thus assigning different importance ranking levels. In addition, the different metrics used by each model for feature importance analysis may lead to differences in the ranking of features. Features like polydipsia, polyuria and gender are the most crucial in predicting the response variable as seen in the LR model as well**.** By identifying these features as significant risk factors of diabetes, early healthcare interventions for diabetes can be implemented, thus achieving our second objective. As a result, better health outcomes could be achieved. The identification of these risk factors also provides actionable insights for healthcare practitioners.

The future direction of our project may involve exploring different machine learning models as well as utilizing different datasets.

4. Data Product

Deployment

In recent years, integrating data science and healthcare has become more common to improve patient outcomes and healthcare decision-making. From the RStudio, the Shiny app was used for deployment purposes. Shiny is an R package that allows us to build interactive applications directly from R. It is typically used for data visualization, analysis and sharing insights with others. The Shiny framework facilitates the development of user-friendly, data-driven applications without requiring extensive web development expertise.

For our Diabetes Predictor, a robust dataset containing relevant features such as gender, polyuria, polydipsia, sudden weight loss, polyphagia, irritability and partial paresis are essential. Choosing an appropriate machine learning algorithm is an important step in developing an effective predictive model. In the context of our Shiny app, we select a random forest and train it using a labeled dataset, dividing the data into training and testing sets to evaluate the model's performance.

To allow end users access to the diabetes predictor, the deployment is hosted on shinyapps.io. The server allows a wider audience to access it without installing any software or files. Refer to the link https://kaiqi96.shinyapps.io/WQD7001_GA2_G8/ for exploration.

End User Feedback

The data product is presented to 3 end users where one is from the healthcare sector and the other two are from the community. Refer to below for the feedback:

- **User 1** (healthcare sector): A user-friendly system to predict the risk of diabetes. Great tool for quick health checking and providing awareness on diabetes.
- **User 2**: This diabetes predictor is straightforward. However, it would be even better if there were more details on how certain features are weighted in the prediction.
- **User 3**: The diabetes predictor is incredibly helpful! It provided me with a quick assessment of my risk for diabetes. The interface is user-friendly and motivates me to adopt a healthier lifestyle to reduce my risk.

5. Plan For Reproducible Research

Reproducibility is critical in this research to validate scientific findings as well as ensure the integrity and reliability of the research outcome. To reproduce our research results, here are the keynotes to be followed.

Documentation of Research Process

Following with the detailed documentation of our research by using OSEMN methodology as shown in Table 1.

Table 2: Steps to reproduce our research results.

Steps	Elaboration
Obtain - Data Collection	• The diabetes dataset is obtained from this link: Early Classification of Diabetes (kaggle.com) • The dataset contains the records of 520 individuals with 17 variables.
Scrub - Data Cleaning	• Dplyr package from R is used to clean and manipulate the data. • Identify missing values by using is.na() function across all variables to check for any missing value. • Perform validity checks to ensure data integrity by ensuring all values fell within the expected range of 0 and 1.
Explore – Summarizing the data, visualizing relationship and conducting statistical analysis	• Perform Exploratory Data Analysis by looking into the distribution of age, gender, and diabetes diagnosis within the dataset. • Examine the differences in the prevalence of symptoms associated with diabetes. • Using Pearson correlation coefficient to quantify the relationship between two variables, ranging from -1 for a perfect negative correlation to +1 for a perfect positive correlation.

	• Using Chi-squared test results to provide statistical solid support to the observed relationship between various symptoms and the diagnosis of diabetes.
Model – Model selection	• This includes Logistic Regression model and Random Forest model. • Divided dataset into training and testing based on 80:20 ratio, where 80% is for training and 20% is for testing through “createDataPartition” function. • Using 10-fold-cross-validation to examine its robustness and reliability in predicting modelling.
Interpret - Data interpretation	• The outcome of this model is reviewed to enable a tougher understanding of how various symptoms could correlate to an individual’s diabetes risk. • Document conclusions drawn from the analysis and their implications for future research or applications.

Code Availability

All the code used in this project can be found in this link (kaiqi96/WQD7001_PDS_G8 (github.com)) which covered all the data analysis, modelling, and other computational aspects such as scripts and algorithms specifically for this study. This link captured all changes made over the project time which allow users to trace the evolution of the project. By following these steps and maintaining detailed documentation, clear code and providing access to the necessary resources, it will increase the likelihood that others can replicate this study by verifying the findings and build upon the work more effectively.

6. Insights And Conclusion

This project aimed to develop a machine learning-powered classification model that can reliably assess an individual's risk of developing diabetes by using available diagnostic and health data such as based on simple diagnostic attributes that can be obtained through standard screening programs. With the standard OSEMN framework, the key project steps included obtaining a diabetes dataset, exploring, and visualizing feature distributions and correlations, trying machine learning algorithms like logistic regression and random forest, selecting the best-performing model, and interpreting the significant risk factors.

The data obtained contained information on various tabular features like gender, polyuria, polydipsia, sudden weight loss, polyphagia, irritability, and partial paresis., etc. Exploratory data analysis revealed insights like correlations between diabetes and the prevalence of symptoms. Classification models like logistic regression and random forest were trained, and their performance was compared to select the best model.

Through this predictive modeling approach, the goal was to create an accurate risk screening tool that provides early diagnosis of vulnerability thereby enabling timely preventive interventions before onset through lifestyle modifications. The final simplified logistics regression model and full-featured random forest models (500 and 600 trees) demonstrated the highest accuracy of 92.31%, validating the usefulness of such data-driven screening approaches for facilitating preventive healthcare and associated cost savings.

By identifying the most influential risk factors for diabetes onset, the model also provides data-driven insights on designing scalable preventive health policies, campaigns, and targeted mitigation strategies for high-risk groups across demographic profiles. The key insights and conclusions are summarized as follows highlighting the usefulness of the developed methodology in facilitating preventive diabetes care through early diagnosis and risk-specific interventions.

The key insights from model performance, significant feature analysis, and evaluation of different modeling algorithms are summarized as follows: -

1. Polyuria and polydipsia are significant indicators of diabetes because the kidney is highly affected by this disease. That causes high sugar levels in the blood and puts a lot of pressure on the kidneys which the kidneys try to eliminate by producing more urine. This in turn leads to dehydration, resulting in thirst. Unstable blood sugar levels can also affect our mood and feelings resulting in symptoms like irritability.

2. Body composition factors like BMI and waist circumference were also major drivers, highlighting weight management through diet and exercise as key prevention strategies.
3. The machine learning model demonstrates the feasibility of accurately predicting an individual's diabetes risk using easily available attributes like demographics, vital stats, and basic lab tests.

In conclusion, the random forest classification model developed in this project can serve as a decision-making tool for policymakers and health workers to predict diabetes risk among people. The performance results validate its potential to be used as an early screening tool that can trigger suitable preventive interventions like counseling for weight loss, diet changes and exercise for high-risk individuals identified by the tool before disease onset. Such interventions can subsequently improve health outcomes and lower medical costs. As more training data gets aggregated over time, the model is expected to become even more robust and generalizable across populations or re-trained periodically to make it even more accurate and representative of diverse populations and age groups.

Overall, this project demonstrates the capabilities of AI and the promising accuracy obtained validates the usefulness of machine learning for predictive healthcare. As more data gets aggregated, the models are expected to become even more accurate. Specific groups can potentially be targeted better for interventions customized to their demographics and risk profiles.

7. Appendixes

kaiqi96. (2024, January 12). *kaiqi96/WQD7001_PDS_G8*. GitHub. https://github.com/kaiqi96/WQD7001_PDS_G8

Diabetes Predictor. (n.d.). Kaiqi96.Shinyapps.io. Retrieved January 12, 2024, from https://kaiqi96.shinyapps.io/WQD7001_GA2_G8/

8. References

Erickson, B. J., & Kitamura, F. (2021). Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiology: Artificial Intelligence, 3(3), e200126. https://doi.org/10.1148/ryai.2021200126

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental Monitoring and Assessment, 189(7), 316. https://doi.org/10.1007/s10661-017-6025-0

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. PLOS ONE, 13(8), e0201904. https://doi.org/10.1371/journal.pone.0201904

Oladimeji, O. O., Oladimeji, A., & Oladimeji, O. (2021). Classification models for likelihood prediction of diabetes at early stage using feature selection. Applied Computing and Informatics. https://doi.org/10.1108/ACI-01-2021-0022

Probst, P., & Boulesteix, A.-L. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. Journal of Machine Learning Research 18, 1–18. https://doi.org/10.48550/arXiv.1705.05654