1.0 Introduction

This report presents a comprehensive review and replication study of the paper titled 'Machine Learning-Based Application for Predicting Risk of Type II Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study' [1]. The paper was selected for review and replication for the following reasons:

1) The publishing journal (IEEE Access) is *reputable*, and ranked in the top quartile by Scimago Journal Rank (SJR) [2];
2) It was published in 2020, ensuring *relevance*;
3) The data are provided, allowing for partial *reproducibility*;
4) Its application of both *classical statistical* and *machine learning* methods demonstrates diverse approaches, with differing features and explainability; and
5) The diabetes (and broader disease detection) field is highly relevant for machine learning, presenting significant *social opportunity*.

1.1 Background Information

The World Health Organisation (WHO) estimated The Kingdom of Saudi Arabia's (KSA) population as 30 million in 2016 [3], with 7 million diabetic and another 3 million pre-diabetic [4]. Globally, diabetes rates have escalated: from 1980 to 2014, the number of people with diabetes almost quadrupled from 108 to 422 million [5], whilst the global population increased only 65% in the same period [6]. The rise is more pronounced in developing nations but also affects developed countries.

The preventable nature of Type II Diabetes Mellitus (T2DM) is emphasised in the literature. Studies demonstrate the effectiveness of lifestyle modifications as prevention and treatment measures [7], including exercise, diet, and weight management. Type II Diabetes typically develops in middle-aged and older adults and is marked by a gradual onset [8]. Therefore, early detection and intervention are critical. Online risk assessment tools such as the Finnish Diabetes Risk Score (FINDRISC) [1] consider demographic, health, and lifestyle related factors, which inspired the study design.

2.0 Study Overview

2.1 Design

Cross-sectional surveys are observational in nature and measure population exposures at a given point in time. The study notes several reasons for favouring this design:

- The surveys can be conducted online, making them economical and fast to implement [26];
- They provide on-demand information on exposures and outcomes; and
- Studies can serve as a baseline for future longitudinal studies [25].

The survey was distributed online to students, teaching, and non-teaching staff at King Abdulaziz University (KAU) in the Western region of Saudi Arabia. The questions were chosen based on common attributes from recent diabetes prediction models to ensure consistency with established research. The questionnaire [1] is listed below, noting questions 1 through 9 were used as predictors, with smoking status also collected:

Variable/Question	Levels	Baseline
Region: Choose the region of your residence.	1: Abwa, 2: Jeddah, 3: Khulays, 4: Medina, 5: Masturah, 6: Mecca, 7: Rabigh, 8: Sabar, 9: Thual, 10: Yambu	1
Gender: What is your gender?	0: Female 1: Male	0
Age: How old are you?	0: < 40 years, 1: 40 - 49 years, 2: 50 - 59 years, 3: ≥ 60 years	0
BMI: What is your Body Mass Index (BMI)?	0: < 25 kg/m^2 1: 25 - 30 kg/m^2 2: > 30 kg/m^2	0
Waist Size: What is your waist size?	Male: 0: < 94 cm 1: 94 - 102 cm 2: > 102 cm Female: 0: < 80 cm 1: 80 - 88 cm 2: > 88 cm	2
Physical Activity: Do you engage in at least 30 minutes of physical activity daily?	0: Yes (≥ 30 minutes/day) 1: No	1
Diet: Do you eat fruits and vegetables daily?	0: Every day (fruits/vegetables) 1: Not every day	0
Blood Pressure: Have you ever taken hypertension (blood pressure) medication?	0: No BP medication, 1: Taking BP medication	0
Family History: Has anyone in your family been diagnosed with diabetes?	0: None 1: Grandparents, uncles, aunties, cousins 2: Parents, siblings	0
Smoking: Do you smoke?	0: Non-smoker, 1: Smoker	0
Response Variable (**Class**)	0: non-diabetic, 1: Diabetic Based on fasting plasma glucose ≥ 5.6 mmol/L	N/A

Table 1: Categorical Variable Coding Scheme

2.2 Objectives

The study addresses several stated objectives related to Type II Diabetes Mellitus (T2DM) risk prediction in the Western region of Saudi Arabia:

1. Estimate the *prevalence* of diabetes in the Western region of Saudi Arabia.
2. Identify significant *risk factors* and *associations* between the exposures and the outcome.
3. Develop a real-time, data-driven *predictive application* for classifying diabetes risk.
4. In the future, conduct a *cohort study* to estimate the incidence and causes of diabetes in Saudi Arabia, using the current survey as a baseline.

2.3 Methods & Results

The study adopted a variety of statistical and machine learning techniques, conducted in phases. Firstly, the Pearson Chi-squared test of independence was applied to each combination of the categorical explanatory variable with the categorical response (10 in total). The test uses contingency tables to assess whether the association between the given explanatory variable and the response is significant, and not due to chance (such as sampling variability).

After applying the independence testing, the study used Cramer’s V to measure the strength of the association. All 10 variables were found to be statistically significant, with each null hypothesis (that no association exists) rejected. Smoking showed the strongest association with a Cramer V of 0.494, followed by diet (0.111), blood pressure (0.108) and physical activity (0.106).

A binary logistic regression model was further employed to describe the direction and strength of associations between the explanatory variables and the response. Odds ratios indicate the relative importance of explanatory variables, while confidence intervals quantify uncertainty in the estimates. The formula below represents the log-odds of the probability of the outcome (high-risk class), expressed as a linear combination of the explanatory variables.

$$log\ (\frac{P}{1-P}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Figure 1: Binary Logistic Regression

Forward stepwise regression was adopted to screen the explanatory variables for statistical significance, with the Hosmer-Lemeshow test used as a goodness-of-fit measure. The Wald Chi-Square Test was also applied to assess the presence and strength of associations. Thereafter, the dataset was pruned to retain only the significant variables based on the findings of the Chi-squared independence tests and binary logistic regression.

The study observes six significant risk factors found by the forward binary logistic regression, namely smoking, BMI, diet, region, blood pressure, and gender. Smoking has the highest odds ratio at 21.3, while BMI, diet and three region categories had odds ratios between 1 and 2. A full summary of the significant predictors' identified by logistic regression is pictured in Table 3.

The data was split into a stratified 80% training set and a 20% independent test set. To handle class imbalance in the training set (approximately 80% low-risk, 20% high-risk), the study applied synthetic minority oversampling technique (SMOTE) to equally balance the classes. A comprehensive flowchart [1] presented in the paper in Figure 1 provides an insightful overview of the study and its key phases. Although SMOTE is a widely known (and utilised) sampling technique, it has limitations, including potential for bias by artificially generating data, impacting the data integrity and overall model interpretability.

In the final analysis phase, *nine* different classifiers were applied to the balanced and unbalanced test sets using SPSS: Bayes Point Machine, Averaged Perceptron, Decision Forest, Locally Deep Support Vector Machine (SVM), Support Vector Machine, Decision Jungle, Logistic Regression, Boosted Decision Tree and Neural Network. The study favours the F1 score as its preferred model selection metric. F1 offers a balanced measure of both precision and recall, aiding in interpreting the associations and key risk factors.

The Random Forest (Decision Forest) classifier was the best performer on the balanced dataset, achieving an F1 score of 0.829. After hyperparameter tuning using 10-fold cross-validation, the model's F1 score improved to 0.8453 ± 0.0268. The final model was validated on reputable public datasets including NHANES and the Pima Indian Diabetes (PID) dataset, where it generalised well. This justified the Decision Forest's selection for the real-time diabetes classification application, which was later launched. A full summary of the classifier results is presented in Tables 4 and 5 of this report.

3.0 Proposed Methodology

The proposed methodology involves replicating and extending the results from the study (where possible), in the staged manner the study adopted. This will be conducted entirely using R. As the study utilised SPSS for the entire analysis with SMOTE as the class balancing technique, some deviations may be observed in the results due to differences in the capabilities of SPSS and R. This will be further discussed in the following sections. Firstly, the Chi-Squared independence testing will be undertaken and compared with the study's results. This will provide an initial overview of significant predictors, while Cramer's V will be used as a measure of association strength.

Subsequently the logistic regression analysis will be replicated to estimate the odds ratios of the significant risk factors. This involves examining how the different predictors and their underlying categories (e.g. Region, BMI level, etc.) influence the likelihood of being classified as a high-risk of developing diabetes relative to the baseline category. This step is critical to understanding the relative impact and significance of each risk factor in the model. Both the Chi-squared and logistic regression analyses will be undertaken using the raw unbalanced survey data. The study's presentation of the significant odds ratios is statistically unprofessional, as they are presented in tabular format in supplementary tables separate to the main document. The analysis conducted in R seeks to visualise the significant odds ratios in a more interpretable manner.

The study adopted the Hosmer-Lemeshow test as the goodness-of-fit measure for the forward logistic regression. The Hosmer-Lemeshow test is a popular fitness measure in logistic regression settings. However, it does have well documented limitations, and can be less reliable for complex models or large datasets [49]. Accordingly, the replication study aims to assess goodness-of-fit using both predictive performance and simulated residuals in addition to the Hosmer-Lemeshow test.

Finally, oversampling will be applied to balance the classes in the training set, with classifiers tested on both balanced and unbalanced test sets. Whilst the study provides balanced and unbalanced datasets [50] after applying SMOTE, it does not include separate training and testing sets. Hence to avoid data leakage, the raw unbalanced dataset will be used, with oversampling applied to synthetically balance the data. This will ensure the analysis replicates the study as closely as possible while avoiding data leakage.

The classifiers trained will be those with available R packages, namely logistic regression, Support Vector Machine (SVM), neural network, boosted decision tree (XGBoost), and a random (decision) forest. It is worthwhile noting the study incorrectly refers to the Bayes Point Machine classifier as Niave Bayes in several sections of the report. The analysis will evaluate the chosen classifiers' performance on both balanced and unbalanced test sets, aiming to replicate the classifier results obtained in the original study. The best performing classifier's hyperparameters will be tuned to (ideally) enhance performance. A comparison of all results will be presented and discussed throughout, with appropriate limitations and critiques provided where relevant.

Notable considerations in improving the logistic regression method include the addition of interaction terms, feature selection techniques, and/or the inclusion of region as a random effect rather than a fixed effect. Given the observational nature of the survey across defined regions of Saudi Arabia, region was treated as a fixed effect to gauge each region's impact. Ultimately this will address the key objectives of the original study by informing the real-time predictive application and estimating the overall prevalence. Interaction terms were modelled but later discounted due to preserving interpretability and avoiding overfitting through a simpler, parsimonious model. Ridge and LASSO regression are adopted for feature selection. Finally, the study's representation of baseline categories as the lowest prevalence of subjects with high risk diabetes is also misleading, as it does not consider the relative exposure of each grouping (waist size and physical activity were coded with conventionally higher risk groups as the baseline). An intuitive baseline is selected for each predictor to improve interpretability.

4.0 Results

4.1 Chi-squared Testing Results

Please refer to https://github.com/kyle-reardon/MXN442.git [14] for the RMarkdown script to reproduce the results. All chi-squared testing results match those reported in the paper. A summary of the tabulated results is shown below in Table 2. For binary variables, the study reports the chi-squared test statistic and associated p-value twice per pair, with one set belonging to the test with continuity correction (CC) and one without. This was replicated in R to ensure consistency in the processes. There are some very slight deviations to those reported in R based on suspected errors made in the paper. For instance, in section III (Results and Discussion), subsection C (Gender), the table reports the BMI tabulated results instead of the Gender results. Several other errors of this nature are made in this section of the paper as the study used figures from other variable sections.

Variable	**# Deg. Freedom**	**Chi-Squared**		**p-value**		**Interpretation**	**Cramer's V**
		With CC	Without CC	With CC	Without CC		
Region	9	33.93		0.000092		Reject H0	0.083
Age	3	16.16		0.001047		Reject H0	0.057
Gender	1	10.88	11.13	0.000973	0.000849	Reject H0	0.048
BMI	2	7.53		0.023159		Reject H0	0.039
Waist Size	2	14.4		0.000746		Reject H0	0.054
Physical Activity	1	54.16	54.69	1.85E-13	1.41E-13	Reject H0	0.106
Diet	1	60.17	60.75	8.71E-15	6.50E-15	Reject H0	0.111
Blood Pressure	1	56.03	56.59	7.12E-14	5.36E-14	Reject H0	0.108
Family History	2	7.33		2.60E-02		Reject H0	0.039
Smoking	1	1192.88	1195.39	2.15E-261	6.12E-262	Reject H0	0.494

Table 2: Chi-squared Testing Results

4.2 Logistic Regression Results

The binary logistic regression model trained on the raw unbalanced dataset with all features matched the study's results rather closely. This generally holds true for the coefficient estimates to approximately 1 decimal place. A full summary of the paper's results in addition to my results obtained in R is pictured below in table 3.

Throughout the paper they incorrectly refer to the significant regions by referring to the wrong region number. It is suspected this may be to do with how the baseline coding scheme was set up in SPSS, as region 3 is referred to as region 2, region 6 is referred to as region 5, etc. This can be verified by cross-checking their supplementary document of results where they list the signicant region numbers and names together, which differ from those defined in the report section II (Methodology).

Explanatory Variable	Sub-category	Paper Results (SPSS)				My Results (R)			
		Coefficient	Odds Ratio (OR)	95% Odds Ratio CI		Coefficient	Odds Ratio (OR)	95% Odds Ratio CI	
				Lower	Upper			Lower	Upper
Region	Region3 (Khulays)	0.618	1.855	1.13	3.05	0.621	1.861	1.131	3.06
	Region6 (Mecca)	0.541	1.718	1.15	2.57	0.537	1.711	1.147	2.572
	Region7 (Rabigh)	0.66	1.934	1.31	2.85	0.663	1.94	1.319	2.878
Gender	Male	-0.26	0.771	0.629	0.946	-0.259	0.772	0.629	0.947
BMI	BMI (25 - 30 kg/m^2)	0.477	1.611	1.153	2.251	0.471	1.602	1.152	2.25
	BMI (> 30 kg/m^2)	0.56	1.751	1.254	2.444	0.558	1.746	1.256	2.451
Physical Activity	Physical Activity (Yes)	0.397	1.487	1.244	1.778	-0.398	0.672	0.561	0.803
Healthy Diet	Healthy diet (Not daily)	0.606	1.833	1.519	2.213	0.604	1.829	1.516	2.21
BP.	BP (Yes)	0.625	1.869	1.547	2.257	0.622	1.863	1.544	2.253
Family History	Family History (Parent, siblings)	-0.307	0.735	0.554	0.977	-0.306	0.737	0.555	0.98
Smoking	Smoking (Yes)	3.059	21.31	17.11	26.55	3.056	21.234	17.116	26.578

Table 3: Binary Logistic Regression Results

As the coefficients are exponentiated to obtain the odds ratios (in addition to the odds ratio confidence intervals), the odds ratios and associated confidence intervals veer slightly from those found by the paper. However, broadly the raw results obtained by the binary logistic regression model are consistent with each other. It should also be noted the baseline was reversed for Physical Activity.

The paper further applied the Wald test on the fitted logistic regression model. This was also conducted in R, with the results almost perfectly matching those found by the paper. There are some very slight deviations in the statistic estimates, however, broadly the results are consistent. Please refer to the full .rmd script and the paper (section III, subsection K, Table 12) for a comparison of these results.

The study finally conducted forward stepwise logistic regression using the Hosmer-Lemeshow as a goodness-of-fit model selection measure throughout. The order in which the forward stepwise regression is conducted is not reported in the study. Therefore, the results do not reflect those in the study. More clarity in the paper may have allowed these results to be replicated. In any case, the Hosmer-Lemeshow has limitations as earlier highlighted. Residuals were simulated as a model fitness measure which showed an improvement in fit when removing the four insignificant variables found by the paper (waist size, age, physical activity, and family history).

4.3 Classifier Results

The final phase of the replication study involved the classification section. Firstly, the data was divided into 80/20 train/test sets using a stratified split. Both the unbalanced train and test sets were then randomly oversampled to achieve a 50/50 class balance. R's sampling packages differ from those available in SPSS. As SMOTE is not well suited to categorical data (as it uses k nearest neighbors to calculate Euclidean distances), random oversampling was applied as a substitute. Given the training and testing data will differ from those used in the study, it is anticipated the classification results may also deviate somewhat.

Classifier	Balanced Test Results									
	Accuracy		Precision		Recall		F1		AUC	
	Paper	Mine	Paper	Mine	Paper	Mine	Paper	Mine	Paper	Mine
SVM	0.795	0.828	0.744	0.792	0.882	0.891	0.807	0.839	0.816	0.847
Logistic Regression	0.798	0.810	0.747	0.770	0.887	0.883	0.811	0.823	0.847	0.845
Logistic Reg. (Ridge)	-	0.818	-	0.773	-	0.900	-	0.831	-	0.845
Logistic Reg. (LASSO)	-	0.814	-	0.772	-	0.891	-	0.827	-	0.845
Neural Network	0.815	0.810	0.78	0.785	0.884	0.855	0.827	0.819	0.853	0.836
Random Forest	0.821	0.827	0.776	0.789	0.89	0.891	0.829	0.837	0.867	0.838
XGBoost	0.808	0.827	0.77	0.790	0.875	0.891	0.819	0.838	0.847	0.839

Table 4: Balanced Classifier Results (green = my best F1; yellow = paper's best F1)

Classifier	Unbalanced Test Results									
	Accuracy		Precision		Recall		F1		AUC	
	Paper	Mine	Paper	Mine	Paper	Mine	Paper	Mine	Paper	Mine
SVM	0.815	0.793	0.564	0.493	0.318	0.899	0.407	0.637	0.837	0.847
Logistic Regression	0.808	0.768	0.527	0.462	0.354	0.894	0.423	0.609	0.846	0.850
Logistic Reg. (Ridge)	-	0.769	-	0.464	-	0.904	-	0.613	-	0.850
Logistic Reg. (LASSO)	-	0.769	-	0.464	-	0.899	-	0.612	-	0.850
Neural Network	0.808	0.787	0.59	0.485	0.118	0.869	0.197	0.622	0.836	0.844
Random Forest	0.789	0.790	0.464	0.489	0.4	0.899	0.43	0.633	0.822	0.841
XGBoost	0.793	0.791	0.476	0.490	0.405	0.899	0.438	0.635	0.832	0.845

Table 5: Unbalanced Classifier Results

The results tables show some deviation from the paper's results, as anticipated. The balanced test set results are broadly similar. Most metrics are within approximately 0.01 to 0.02 of each other in the balanced table. Notably, the top performing classifier was different, as SVM outperformed the paper's best untuned classifier (XGBoost) F1 score by 0.01. The random forest and XGBoost also achieved greater F1 scores than all untuned classifiers used by the paper. Ridge and LASSO regression were adopted in the binary logistic regression model which showed an improvement in all performance metrics.

The unbalanced results differ substantially. It is suspected that the paper trained the models on the raw unbalanced set when testing on the unbalanced test set, as indicated by the recall values of approx. 0.3 to 0.4 for most classifiers. This approach is questionable. To replicate the real-world setting in which the classification application is deployed, the model trained on balanced data should be tested on unbalanced data to mimic samples being received. This allows a more meaningful comparison of the testing performance and generalisability.

Future work in the domain could benefit from a variety of considerations. The sample may be representative of Western Saudi Arabia, however, is quite a localised region. Inferences cannot be made about diabetes prevalence in other locations. Similarly, a developed model may not generalise to other populations. The synthetic class balancing introduces bias to the model and can lead to overfitting, despite improving test performance. Fine-tuning could have been applied to all classifiers to gauge which model may generalise best. Lastly, whilst the study considered notable risk factors used by other diabetes researchers, it may have overlooked potentially important factors.

Reference List

- [1] A. H. Syed and T. Khan, “Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study,” IEEE Access, vol. 8, pp. 199539-199561, 2020.
- [2] Scimago. “Scimago Journal & Country Rank”. Accessed 1 September 2024. [Online]. Available: https://www.scimagojr.com/journalsearch.php?q=21100374601&tip=sid&clean=0
- [3] World Health Organisation. “Saudi Arabia”. Accessed 14 October 2024. [Online]. Available: https://data.who.int/countries/682#:~:text=Overall%20Population%20over%20time&text=In%20Saudi%20Arabia%2C%20the%20current,43%25%20to%2047%2C693%2C910%20by%202050.
- [4] M. A. Al Dawish et al. “Diabetes Mellitus in Saudi Arabia: A Review of the Recent Literature,” Curr. Diabetes Rev., vol. 12, no. 4, pp. 359-368, 2016, doi: 10.2174/1573399811666150724095130.
- [5] World Health Organisation. “Diabetes”. Accessed 12 September 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
- [6] World Bank Group. “Population, total”. Accessed 20 October 2024. [Online]. Available: https://data.worldbank.org/indicator/SP.POP.TOTL
- [7] M. A. Al Mansour, “The Prevalence and Risk Factors of Type 2 Diabetes Mellitus (DMT2) in a Semi-Urban Saudi Population,” Int. J. Environ. Res. Public Health, vol. 17, no. 1, pp. 7, Dec. 2019, doi: 10.3390/ijerph17010007.
- [8] National Library of Medicine. “Overview: Type 2 diabetes”. Accessed 21 October 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279509/#:~:text=Type%202%20diabetes%20can%20develop,Frequent%20urination
- [9] E. Virtanen et al. “Feel4Diabetes healthy diet score: development and evaluation of clinical validity,” BMC Endocr. Disord., vol. 20, suppl. 2, pp. 46, May 2020, doi: 10.1186/s12902-020-0521-x.
- [10] P. R. Regmi, E. Waithaka, A. Paudyal, P. Simkhada, and E. Van Teijlingen, “Guide to the design and application of online questionnaire surveys,” Nepal J. Epidemiology, vol. 6, no. 4, pp. 640-644, May 2017.
- [11] M. S. Setia, “Methodology series module 3: Cross-sectional studies,” Indian J. Dermatol., vol. 61, no. 3, pp. 261-264, 2016.
- [12] N. Surjanovic and T. M. Loughin, “Improving the Hosmer-Lemeshow goodness-of-fit test in large models with replicated Bernoulli trials,” J. Appl. Stat., vol. 51, no. 7, pp. 1399-1411, 2023, doi: 10.1080/02664763.2023.2272223.
- [13] Github. “Type-2-Diabetes-Mellitus-T2DM-Predictor”. Accessed 14 September 2024. [Online]. Available: https://github.com/SAH-ML/T2DM-Risk-Predictor/blob/master/README.md
- [14] Github. “MXN442 Replication Study of Type 2 Diabetes Risk Prediction”. Accessed 27 October 2024. Available: https://github.com/kyle-reardon/MXN442.git