# **Diabetes Prediction Using Explainable AI Methods Mid Progress Report**

1. Abdelrahman Khaled 202201655
2. Yousef Ahmed Elrefaey 202201588
3. Omar Eldesoukey 202202155
4. Karim Wael 202202212

**Work report**

**1-Chawan, P., & Chaudhari, S. (2018). Logistic regression and SVM-based diabetes prediction system. *ResearchGate*.**<https://www.researchgate.net/profile/Pramila-Chawan/publication/326416823_LOGISTIC_REGRESSION_AND_SVM_BASED_DIABETES_PREDICTION_SYSTEM/links/5b4c80ddaca272c60947858a/LOGISTIC-REGRESSION-AND-SVM-BASED-DIABETES-PREDICTION-SYSTEM.pdf>, Abdalrahman Khaled 202201655

2-**Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm.**

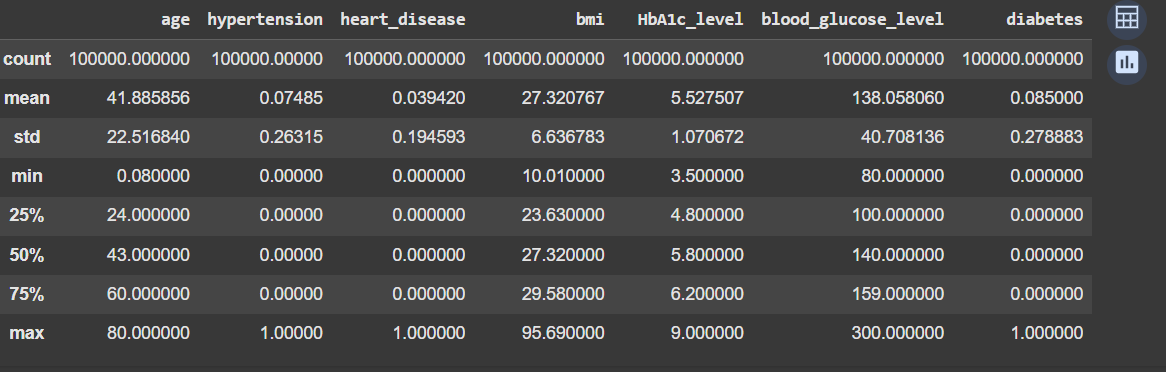
[**https://shmpublisher.com/index.php/joiser/article/view/245**](https://shmpublisher.com/index.php/joiser/article/view/245) **Yousef Ahmed**

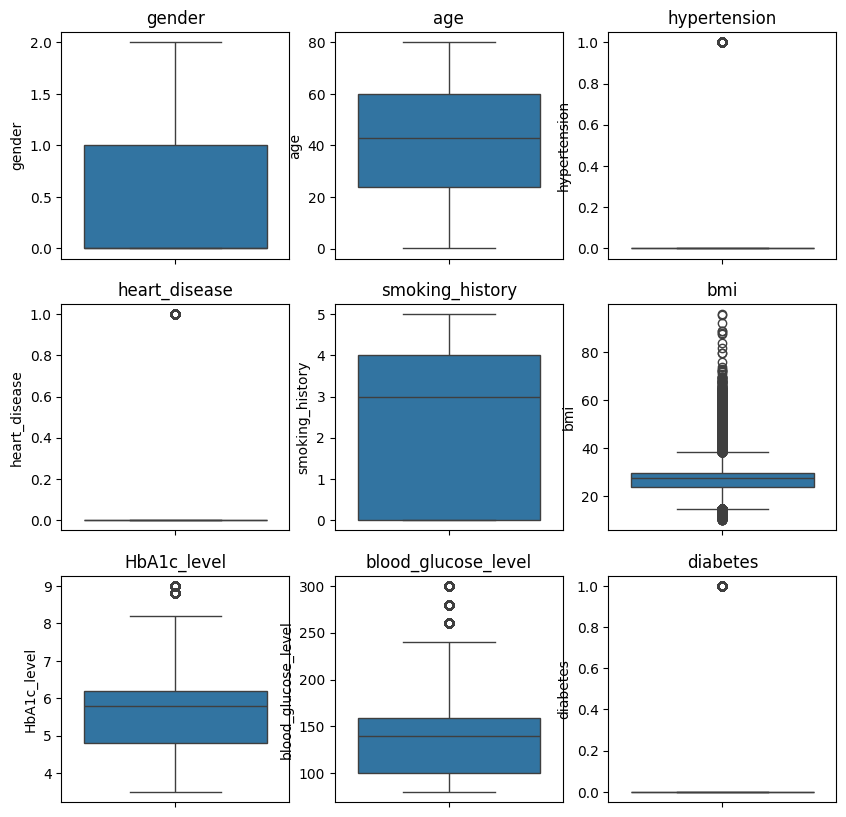
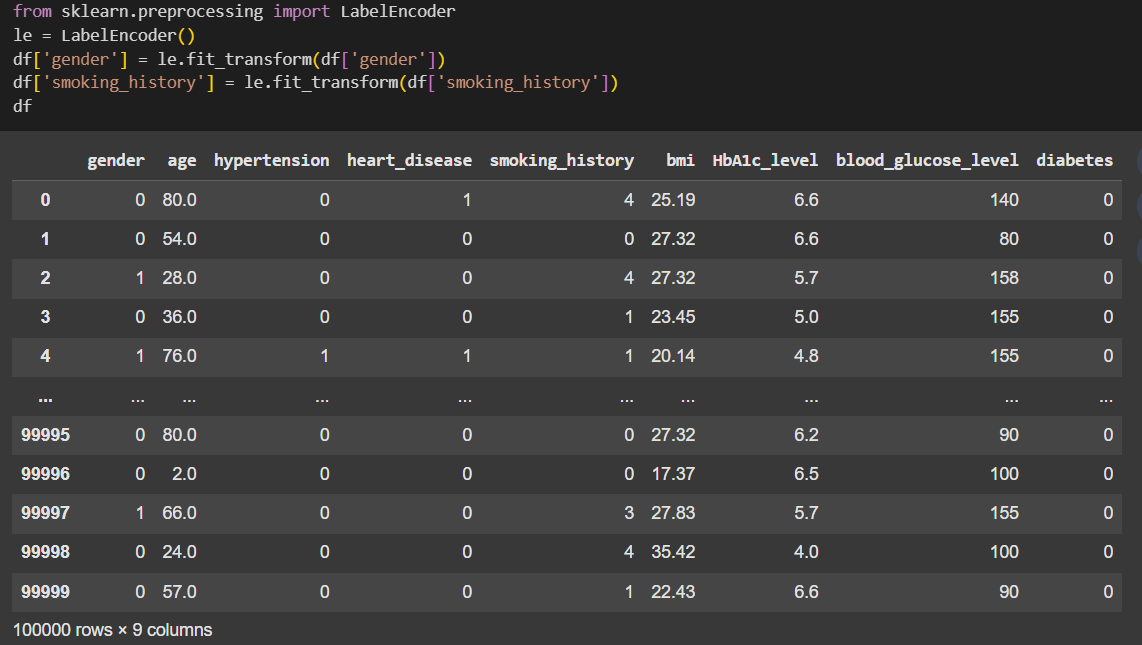
**3-An Improved Artificial Neural Network Model for Effective Diabetes Prediction**

[**https://philpapers.org/rec/EL\_DPU-5**](https://philpapers.org/rec/EL_DPU-5) **KARIM WAEL**

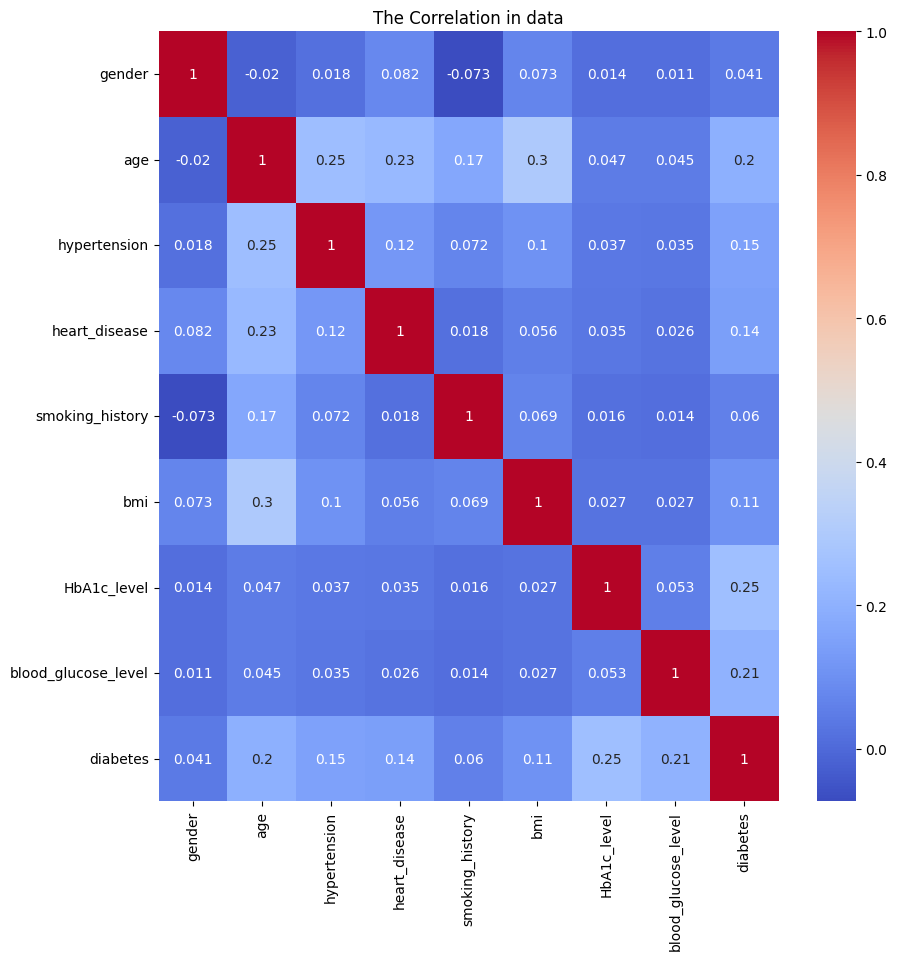
**Github repository link for my work:** <https://github.com/Hendawi1001/Abdalrahman-Khaled202201655-DSAI305-team13-code>

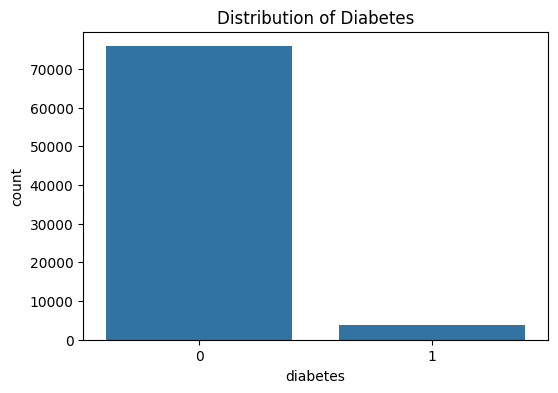
**Notebook link :** [Copy of Team13\_midprogress.ipynb](https://colab.research.google.com/drive/1WYq8KWRJRBheOo2zR2jwriDdt1U_jVhT?usp=sharing)

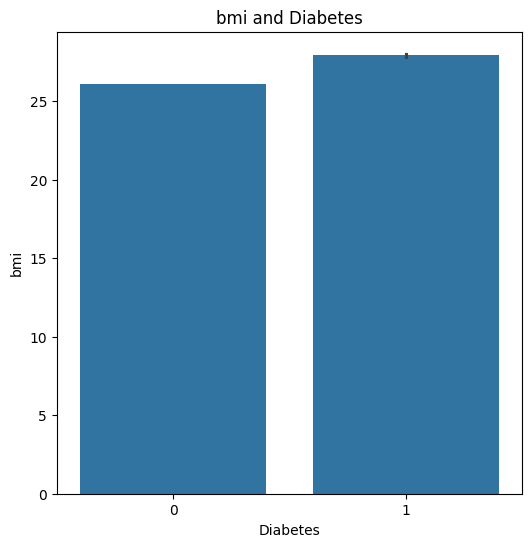
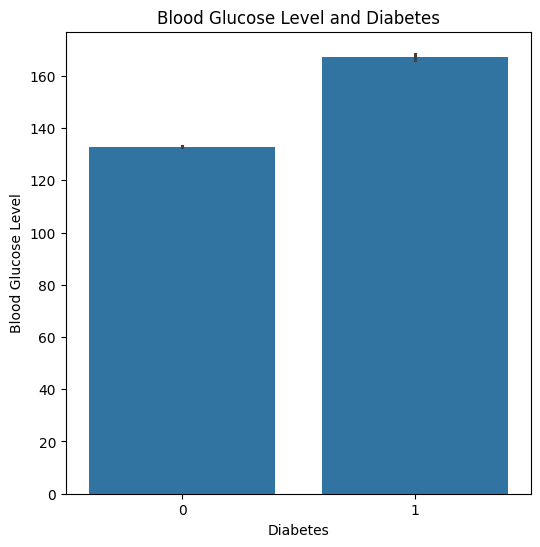
**1-preprocessing:**  
Data explanation, handle categorical variables, missing values, and scaling:

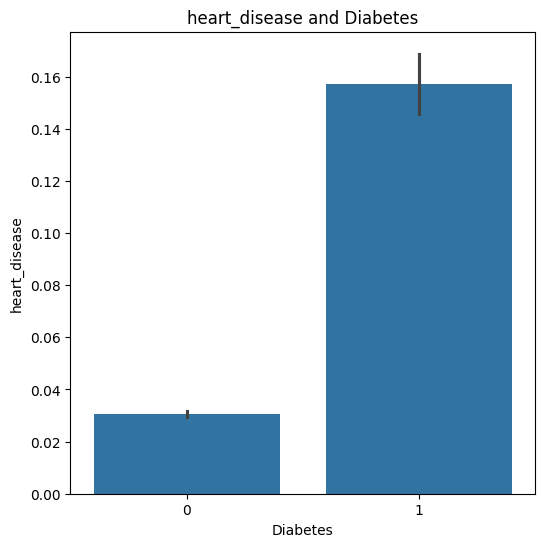


**2-EDA implementation :**

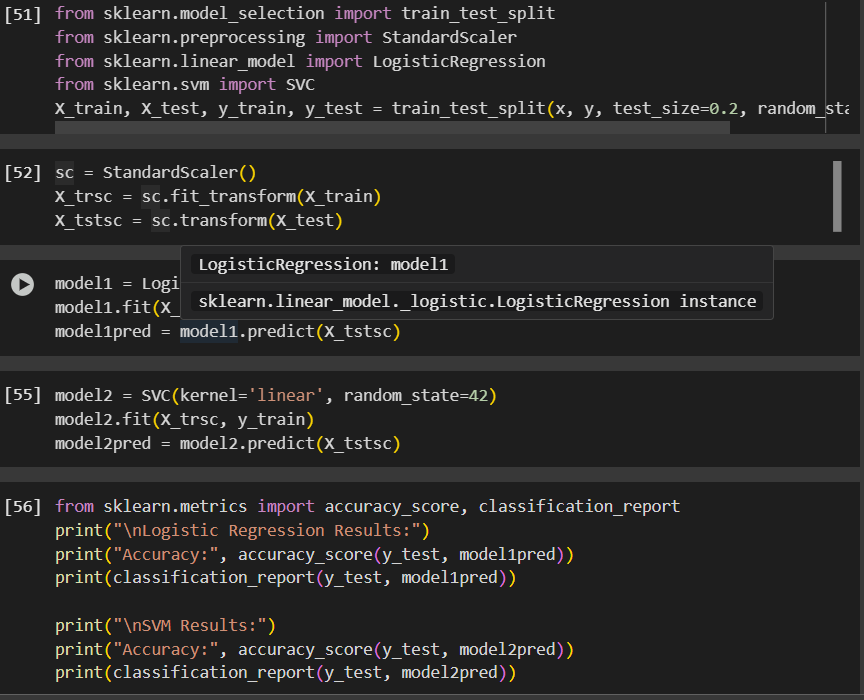
Analyze feature distributions and relationships with the target variable:







so, there is a strong relation between diabetes and hypertension heart\_disease , smoking\_history,and HbA1c\_level

**3-feature selection and machine learning models building:**

Abdelrahman Khaled:

The paper uses Logistic Regression and Support Vector Machine (SVM) models :

**The accuracy report of the two models:**

**Logistic Regression Results:**

**Accuracy: 0.9650424394844389**

**precision recall f1-score support**

**0 0.97 0.99 0.98 15181**

**1 0.76 0.34 0.47 724**

**accuracy 0.97 15905**

**macro avg 0.86 0.67 0.73 15905**

**weighted avg 0.96 0.97 0.96 15905**

**SVM Results:**

**Accuracy: 0.9643508330713612**

**precision recall f1-score support**

**0 0.96 1.00 0.98 15181**

**1 0.99 0.22 0.36 724**

**accuracy 0.96 15905**

**macro avg 0.98 0.61 0.67 15905**

**weighted avg 0.97 0.96 0.95 15905**

**Interpretation of the logistic regression model :**

**Optimization terminated successfully.**

**Current function value: 0.104533**

**Iterations 10**

**Logit Regression Results**

**==============================================================================**

**Dep. Variable: diabetes No. Observations: 63620**

**Model: Logit Df Residuals: 63611**

**Method: MLE Df Model: 8**

**Date: Thu, 17 Apr 2025 Pseudo R-squ.: 0.4469**

**Time: 00:39:32 Log-Likelihood: -6650.4**

**converged: True LL-Null: -12025.**

**Covariance Type: nonrobust LLR p-value: 0.000**

**==============================================================================**

**coef std err z P>|z| [0.025 0.975]**

**------------------------------------------------------------------------------**

**const -5.5967 0.063 -89.301 0.000 -5.720 -5.474**

**x1 0.1492 0.023 6.378 0.000 0.103 0.195**

**x2 0.9808 0.031 31.378 0.000 0.920 1.042**

**x3 0.1991 0.015 13.062 0.000 0.169 0.229**

**x4 0.1446 0.014 10.039 0.000 0.116 0.173**

**x5 0.1781 0.025 7.095 0.000 0.129 0.227**

**x6 0.4052 0.027 14.739 0.000 0.351 0.459**

**x7 2.3013 0.050 45.957 0.000 2.203 2.399**

**x8 0.9870 0.025 38.939 0.000 0.937 1.037**

**==============================================================================**

Final interpretation :High absolute z-score=89.301 → Strong predictor. p-value < 0.05=0 → Statistically significant feature.

**Conclusion of my work**

I have finished testing and reviewing the claim in the paper and it was showing their right claim about their models. I have faced some challenges with the outliers ,trying to fix them also in the scaling of the data , but the work finished successfully.

**Yousef Ahmed:**

**The paper uses Random Forest model:**

**Random Forest Results:**

**Accuracy: 0.972508020798761**

**precision recall f1-score support**

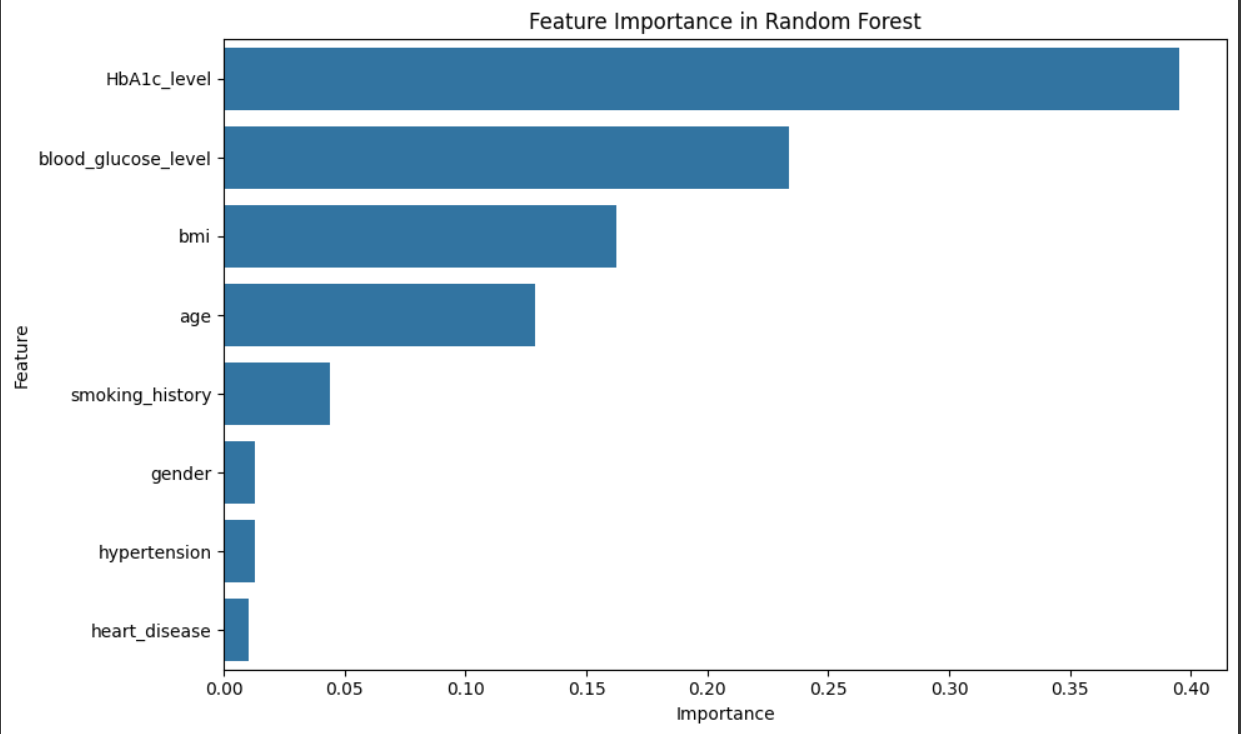
**0 0.97 1.00 0.99 17177**

**1 0.90 0.50 0.65 901**

**accuracy 0.97 18078**

**macro avg 0.94 0.75 0.82 18078**

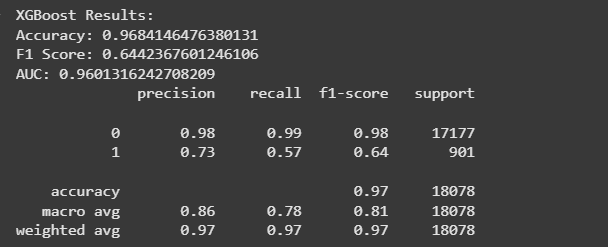
**weighted avg 0.97 0.97 0.97 18078**

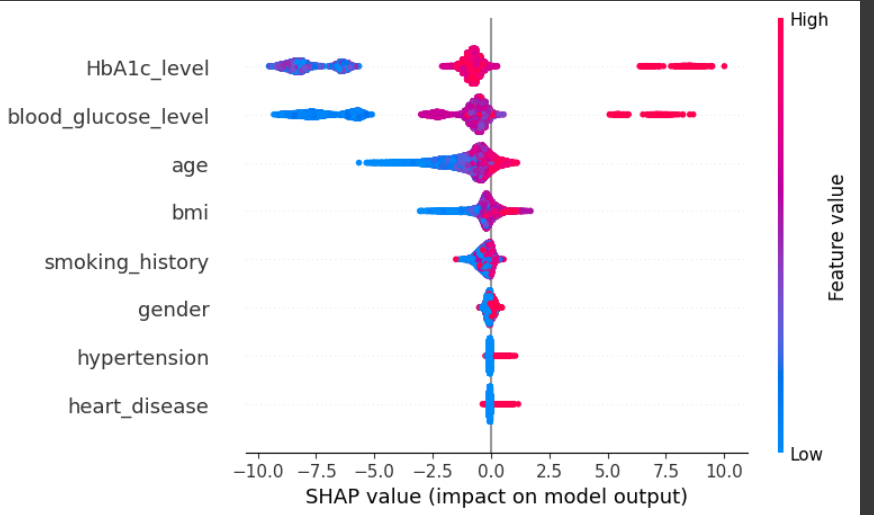
****

**According to the Random Forest model, the most important feature in predicting diabetes is HbA1c\_level, as it has the highest feature importance score. This indicates that this feature contributes the most to the model's decision-making process when classifying individuals as diabetic or non-diabetic.**

**Omar Eldeosukey-202202155**

* **My Results (XGBoost): 96.84% accuracy, 0.97 F1 score (weighted avg), 0.96 AUC.**
* **Paper Results (XGBoost with ADASYN): 81% accuracy, 0.81 F1 score, 0.84 AUC.**

**System outperforms the paper’s XGBoost model by 15.84% in accuracy, 0.16 in F1 score, and 0.12 in AUC. The paper’s best non-XGBoost model (Bagging with ADASYN) achieved 79% accuracy and 0.79 F1 score, which my results surpass. The paper’s domain adaptation on the RTML dataset reached 96% accuracy and 0.95 F1 score, slightly below my 96.84% accuracy and 0.97 F1 score.**

****

**Karim wael   
The result of ANN:**

**The model achieved an accuracy of accuracy: 0.9610**

loss: 0.1142

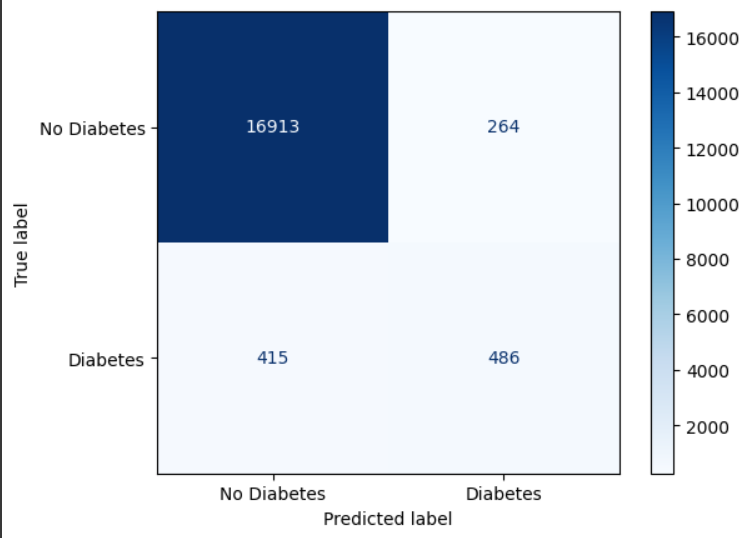
val\_accuracy: 0.9634 - val\_loss: 0.1115

**Paper’s result:**

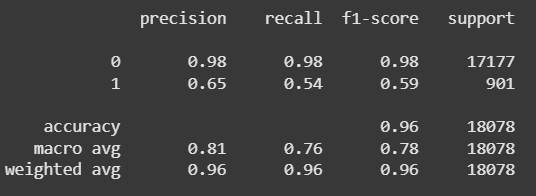
**Training Accuracy: 94.44%**

**Validation Accuracy: 92.708%**

**Visualizing the confusion matrix:**

****

**The classification report:**

****