KING'S OWN INSTITUTE*

Success in Higher Education

INDIVIDUAL ASSIGNMENT COVERSHEET

Family Name: SABABA Given Name: TASNIM

Student Number: 20028086 Lecturer's/ Tutor`s Name: Dr. Firoz Anwar

Subject Code & Name: ICT370 Data Analytics

Assignment Title: Individual Progress Report and Reflection-Submission Week 10

Declaration

(This declaration must be completed by the student or the assignment will not be marked.)

I certify the following:

- I have read and understood the *Student Academic Misconduct Policy*.
- This assignment is my own work based on my personal study and or research.
- I have acknowledged all material and sources used in the preparation of this assignment including any material generated in the course of my employment.
- The assignment has not previously been submitted for assessment.
- I have not copied in part or in whole or otherwise plagiarised the work of other students.
- I have read and I understand the criteria used for assessment.
- The assignment is within the word and page limits specified in the unit outline.
- The use of any material in this assignment does not infringe the intellectual property / copyright of a third party.
- I understand that this assignment may undergo electronic detection for plagiarism, and an anonymous copy of the assignment may be retained on the database and used to make comparisons with other assignments in future.
- By completing this coversheet in full and submitting this assignment electronically, I am bound by the conditions of the KOI's *Student Academic Misconduct Policy* and the declaration on this coversheet.

sababa
Signature

03 / 09 / 2025
Date

Assignment Receipt

Family Name: SABABA Given Name: TASNIM

Student Number: 20028086 Lecturer's/ Tutor`s Name: Dr. Firoz Anwar

Subject Code & Name: ICT370 Data Analytics

Assignment Title: Individual Progress Report and Reflection-Submission Week 10

Signature

03 / 09 / 2025
Date

Introduction:

This biweekly assignment focuses on the predictive analysis of the diabetes prediction data set. The purpose was to apply classification methods to ascertain the probability of diabetes given the population and clinical data. Implementation followed evaluation and comparison of the Logistic Regression and Random Forest models. Evidence of the work done, along with discussion of the findings, comparison with scholarly literature, suggestions for improvement, and reflection on the concepts learned, are contained in the visualizations and metrics.

The following two classification models stem from this analysis:

- **Logistic Regression**: It is a very efficient approach and it is the most widely used model in clinical research.
- **Random Forest:** Recognized for non-linear patterns and interdependencies.

The coefficients in Logistic Regression are easier to interpret, and odds ratios are clinically meaningful. Since Random Forest is an ensemble of decision trees, he tends to achieve higher values of recall, ROC and ROC-AUC. The models conveniently contrast interpretability to predictive performance.

model	accuracy	precision	recall	f1	roc_auc	avg_precision
Logistic	0.889	0.427	0.894	0.577	0.963	0.821
Random Forest	0.969	0.93	0.693	0.794	0.963	0.857

Figure 1: Model performance metrics

Here Random Forest outperformed Logistic Regression, as it also had the higher recall and ROC-AUC values and thus was better at classifying actual positives. Regression had more explanatory power, but also lower sensitivity in comparison to the other measures.

Logistic Regression:

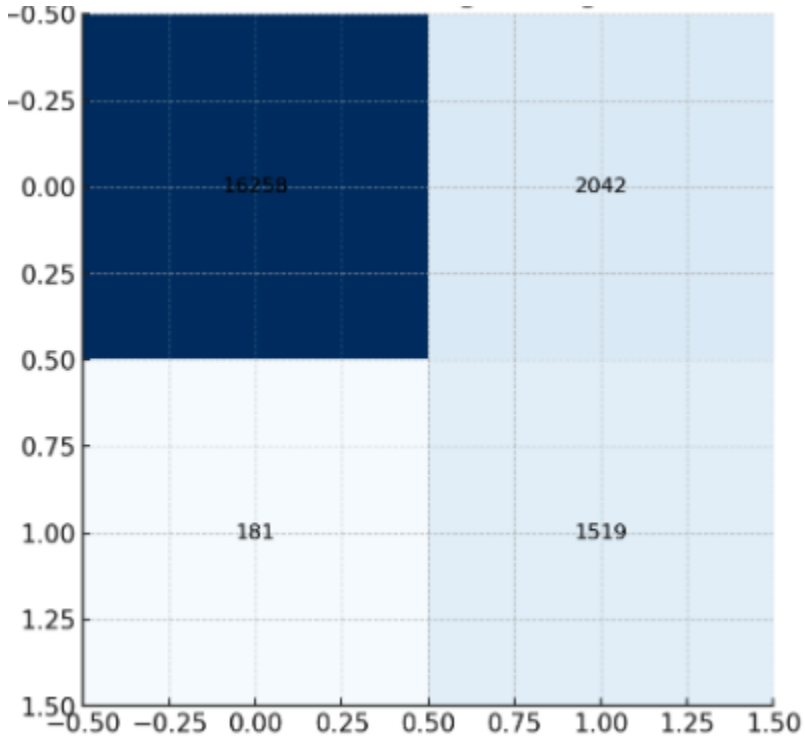

Figure 2: Confusion Matrix- Logistic Regression

Random Forest:

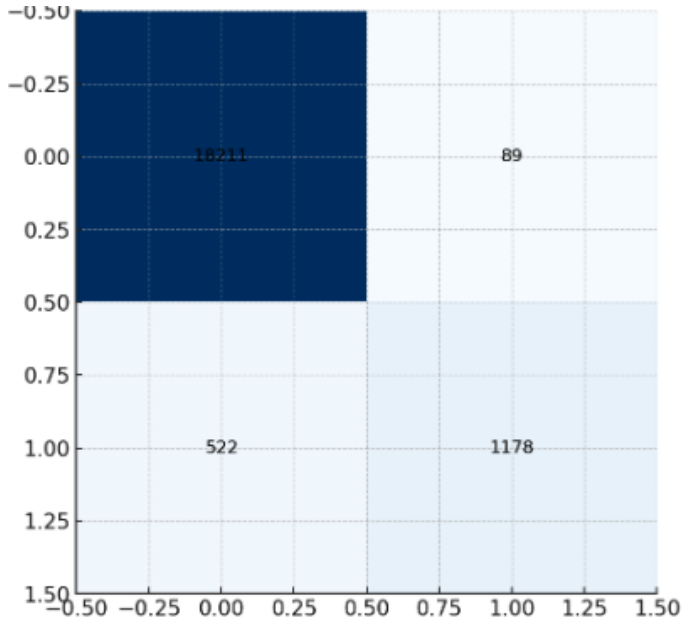

Figure 3: Confusion Matrix- Random Forest

This below ROC curve compares the discriminative ability of both - logistic and random forest models:

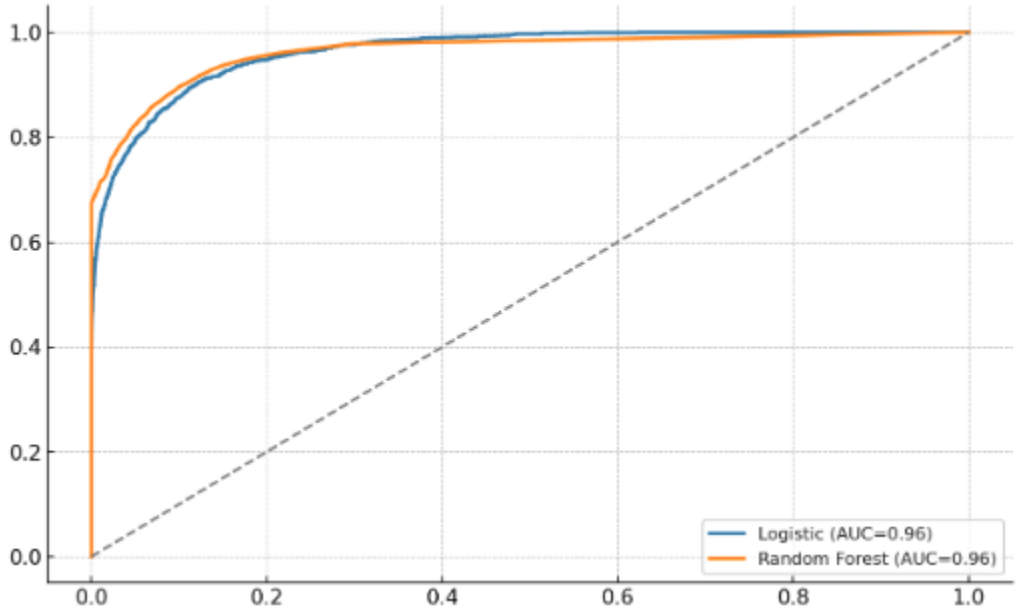

Figure 4: ROC Curve

The Precision-Recall curve illustrates model performance under class imbalance:

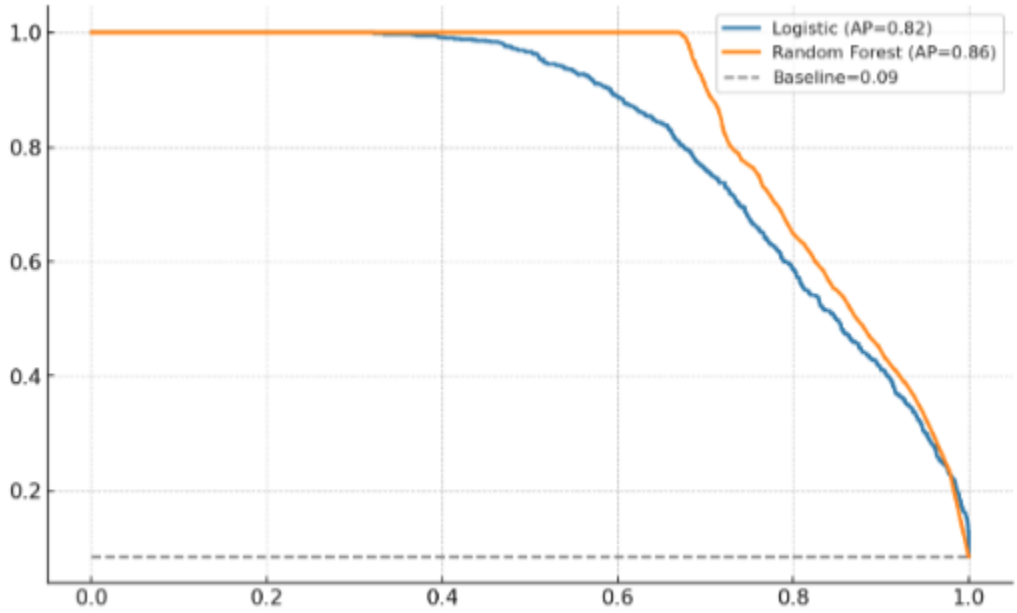

Figure 5: Precision-Recall Curve

Novel Findings:

This analysis verified that HbA1c and blood glucose levels are predictor factors of diabetes and confirmed them with clinical practice. Age and BMI were labeled tertiary factors. For recall and AUC, random forest showed better results than logistic regression which makes it better for the detection of diabetes positive cases. The class imbalance analysis of the Precision-Recall curve was very informative as it illustrated the need to monitor the balance between sensitivity and specificity of the classifier.

Comparison with Scholarly Literature:

My biweekly results are in accordance with the literature. Kivimäki et al. (2022) pointed out the advantage of using Logistic Regression for diabetes risk assessment because its results are easier to explain. Random Forests are used to predict the progression of HbA1c (Yamaguchi et al. 2021) and are also used to improve the predictive power of models. Zheng et al. (2019) used Logistic Regression, neural networks, and decision trees in his work and noted that the non-linear models are useful as it correlates with our results of the ROC and PR curves. More recently, a study finds that HbA1c and glucose level are consistently the top most features across various ML models(Al-Douri et al. 2024) , while my analysis showed that they had the strongest dominating features for the model.

Suggestions for Improvement:

A few actions could improve the results:

1. For added confidence, use k-fold cross validation
2. Adjust probabilities to match the expected frequencies.
3. Shift the cutoffs to optimize for recall (screening) and precision (diagnostic).
4. Add elements such as lifestyle and health history.
5. Employ explainability methods such as SHAP for better understanding of Random Forest outcomes.

Conclusion:

Through the course of these biweekly exercises, I learned skills in participant-driven and predictive analytics. First, I quantified center and dispersion, studied distributions with histograms and box diagrams, and obtained data sets which had anomalies like outliers and skewed distributions. On top of this, I assimilated several classification models to predict a patient's diabetes. I learned to data wrangle categorical and numeric attributes, assess models based on a matrix of accuracy, recall, F1, ROC-AUC and PR-AUC, and confusion matrix evaluation. I also studied the significance of features and thought about how the insights I have developed aligned with literature. This particular project helped me to understand the integration of descriptive analytics with predictive modeling and the systematic presentation of the findings in an academic format.

References:

- Kivimäki, M., et al. (2022) 'Prediction of type 2 diabetes mellitus onset using logistic regression', eLife 'Accessed' 12 9 2025 'at' https://elifesciences.org/articles/71862 .
- Yamaguchi, T., et al. (2021) 'Random forest approach for predicting future HbA1c from health-check data', BMJ Nutrition, Prevention & Health. 'Accessed' 12 9 2025 'at' https://nutrition.bmj.com/content/early/2021/03/09/bmjnph-2020-000200
- Zheng, T., et al. (2019) 'Application of three statistical models for predicting the risk of diabetes', BMC Endocrine Disorders. 'Accessed' 12 9 2025 'at' https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0456-2
- Al-Douri, Y., et al. (2024) 'Feature importance and model performance for (pre)diabetes prediction across ML models', Computers in Biology and Medicine. 'Accessed' 12 9 2025 'at' https://www.sciencedirect.com/science/article/pii/S1018364724004956