Peer-Review
Lau, Aydan H C. 2025. "Machine Learning Classifiers for Non-Invasive Questionnaire-Based Type 2 Diabetes Detection." *Journal of High School Science* 9 (3): 595–611. https://doi.org/10.64336/001c.144808.

I found this to be an interesting topic, well presented. Congratulations. However, I have several concerns that need to be addressed.

1.Since your intention is to catch undiagnosed persons with diabetes, it seems illogical to remove all prediabetes cases from the dataset. There are 96 million persons in the US with prediabetes out of which 80% (77 million are undiagnosed). Out of these, approx. 10% will go on to develop diabetes each year = 8 million new cases of undiagnosed diabetes per year. Compare this to people who actually have full-blown diabetes but are undiagnosed = 8.7 million. The number of prediabetes cases who go on to develop full blown diabetes but are undiagnosed hence represents 100% of the currently undiagnosed full blown diabetec population. Your algorithm is hence significantly less useful if you focus solely on the binary classification task. Please explain and describe in the manuscript why you did not include 3 classes (diabetes, no-diabetes or pre-diabetes) in your work.

2.you mention that your testing set datasets were balanced 50:50 with diabetes and no-diabetes cases. What about the training set? If there was class imbalance in the total dataset, how was it addressed? If not addressed, why? Implications? please explain and describe in the manuscript.

3.How was BMI normalized (on a scale of 0 to 1)? Please describe in the manuscript.

4.Were the following points considered to minimize false negatives?: lowering the classfication threshold? , deliberately introducing class-imbalance (in favor of diabetic cases) by synthetic data to increase the number of diabetic cases?, appropriate hyperparameter tuning and or selecting those features that would skew the model toward a low false negative rate?, Ensemble methods (did you combine all three models together?). Explain each point and why it was not addressed in the manuscript instead of opting for an outside model generalized cost function.

5.If you are going to adopt an outside model generalized cost function anyway, does it really matter how your base model predicts? You can always incrase the cost associated with a false negative. Why then is it important to test several models for their relative efficacy in minimizing false negatives?

6.I am usure of this idea but I will put it to you anyway. Why don't you run two models for each person? Run the neural network to minimize false negatives (increase sensitivity), then SIMULTANEOUSLY use a cost function for the logistic regression to minimize false positives (increase specificity). That way, a physician or an individual can balance sensitivity with specificity. This is because, a person may not unnecessarily want to be subjected to blood tests when they are actually not in danger of developing diabetes. Please explain in the manuscript why SIMULTANEOUSLY running two models with conservative sensitivity and conservative specificity numbers (each derived from a different model) may not be better in terms of unnessasarily subjecting true non-diabetics to blood tests.

7.Please run a robustness test for your 3 algorithms. i.e similar to a k-fold cross-validation.

8.Please present confusion matrixes and AUC curves for the cost-unweighted and cost-weighted Neural network.

9.Please run all the numbers through a heat map showing co-linearity numbers among the 21 questions (for example, q 20 and q 21 are probably co-linear, as are questions 9 and 10…….). Present VIF values as appropriate. Is it possible for you to 'force' any of your ML models to drop (one out of X) correlated features? Will this result in a better model?

10.As another non-AI exercise, please perform a check with a simple excel worksheet and report the correlation between (the average of the normalized value) for the # of non colinear questions answered verus the predicted diabetes/non-diabetes value (0 or 1).

I found this to be an interesting topic, well presented. Congratulations. However, I have several concerns that need to be addressed.

1. Since your intention is to catch undiagnosed persons with diabetes, it seems illogical to remove all prediabetes cases from the dataset. There are 96 million persons in the US with prediabetes out of which 80% (77 million are undiagnosed). Out of these, approx. 10% will go on to develop diabetes each year = 8 million new cases of undiagnosed diabetes per year. Compare this to people who actually have full-blown diabetes but are undiagnosed = 8.7 million. The number of prediabetes cases who go on to develop full blown diabetes but are undiagnosed hence represents 100% of the currently undiagnosed full blown diabetic population. Your algorithm is hence significantly less useful if you focus solely on the binary classification task. Please explain and describe in the manuscript why you did not include 3 classes (diabetes, no-diabetes or pre- diabetes) in your work.
2. you mention that your testing set datasets were balanced 50:50 with diabetes and no-diabetes cases. What about the training set? If there was class imbalance in the total dataset, how was it addressed? If not addressed, why? Implications? please explain and describe in the manuscript.
3. How was BMI normalized (on a scale of 0 to 1)? Please describe in the manuscript.
4. Were the following points considered to minimize false negatives?: lowering the classification threshold? , deliberately introducing class-imbalance (in favor of diabetic cases) by synthetic data to increase the number of diabetic cases?, appropriate hyperparameter tuning and or selecting those features that would skew the model toward a low false negative rate?, Ensemble methods (did you combine all three models together?). Explain each point and why it was not addressed in the manuscript instead of opting for an outside model generalized cost function.
5. If you are going to adopt an outside model generalized cost function anyway, does it really matter how your base model predicts? You can always increase the cost associated with a false negative. Why then is it important to test several models for their relative efficacy in minimizing false negatives?
6. I am usure of this idea but I will put it to you anyway. Why don't you run two models for each person? Run the neural network to minimize false negatives (increase sensitivity), then SIMULTANEOUSLY use a cost function for the logistic regression to minimize false positives (increase specificity). That way, a physician or an individual can balance sensitivity with specificity. This is because, a person may not unnecessarily want to be subjected to blood tests when they are actually not in danger of developing diabetes. Please explain in the manuscript why SIMULTANEOUSLY running two models with conservative sensitivity and conservative specificity numbers (each derived from a different model) may not be better in terms of unnecessarily subjecting true non-diabetics to blood tests.
7. Please run a robustness test for your 3 algorithms. i.e. similar to a k-fold cross- validation.
8. Please present confusion matrixes and AUC curves for the cost-unweighted and cost-weighted Neural network.
9. Please run all the numbers through a heat map showing co-linearity numbers among the 21 questions (for example, q 20 and q 21 are probably co-linear, as are questions 9 and 10…….). Present VIF values as appropriate. Is it possible for you to 'force' any of your ML models to drop (one out of X) correlated features? Will this result in a better model?
10. As another non-AI exercise, please perform a check with a simple excel worksheet and report the correlation between (the average of the normalized value) for the # of non colinear questions answered versus the predicted diabetes/non-diabetes value (0 or 1).

Dear Reviewer,

I'm grateful for your insightful comments. I have revised the manuscript based on your feedback. Please see my answers to your questions below.

1. The prediabetes cases were removed from the dataset because there was only 2.4% of prediabetes cases while there are 14.2% of diabetic cases and 83.4% of non-diabetic cases. I determined there weren't enough pre-diabetic cases to train the models properly. As the number of prediabetes cases was small, excluding that wouldn't affect the quality of training. However, I agree that from a clinical perspective, categorizing pre-diabetes with diabetes prioritizes the detection of at-risk individuals, aligning with the objective of early intervention. I have now included the pre-diabetic cases into diabetic cases. The above reasoning has been included in section 2.3 of the manuscript.

The pre-diabetic 2.4% is much lower than CDC forecast of 38% of adult population. The disparity between the dataset and real-world is probably because BFRSS survey, from which the dataset is derived, relies on self-reporting. From the data above, up to 94% (=1 - 2.4%/38%) of pre-diabetes may be unaware that they were pre-diabetic.

2. In the last manuscript, the diabetic cases were split into two sets: 80% training and 20% testing. For the training set, non-diabetic cases were randomly selected without replacement and added to the diabetic base to create a 50:50 class ratio. For the testing set, non-diabetic cases were randomly selected without replacement and added to the diabetic base to maintain the dataset's original prevalence ratio of 14.2%.

In the revised manuscript, the class ratio of the training set was kept balanced at 50:50 while the prevalence of the testing set was revised to 14.7%, which is consistent with real-world prevalence as reported by the CDC for a realistic assessment of model's generalization ability. The above has been included in the last paragraph of section 2.4.

3. BMI, along with other metrics that require normalization, were normalized using the MinMaxScaler in scikit-learn. MinMaxScaler works well to preserve the relationships between data points. Also, all the factors are naturally bounded to an extent, which helps as MinMaxScaler simply scales down the differences. The description has been included in the last paragraph of section 2.3.

4. As for your points to minimize false negatives:
    - I have revised the manuscript to include steps to tune probability threshold (Threshold Tuning under section 2.4). Results are included in section 5.1.
    - The training set was deliberately made balanced. However, I considered against using synthetic data because the dataset is already very large with over 236,000 data points and over 33,000 diabetic cases. Therefore, it seemed not worthwhile to risk creating unrealistic profiles from synthetic data generation.
    - I have now included feature selection (section 2.2).
    - I haven't considered ensemble methods as the performance of the three models was fairly close and the benefits of ensemble methods seemed limited.

5. I think the performance of the base models still matters as ROC curve of model A may possibly completely dominate that of model B (i.e. ROC curve of model A is

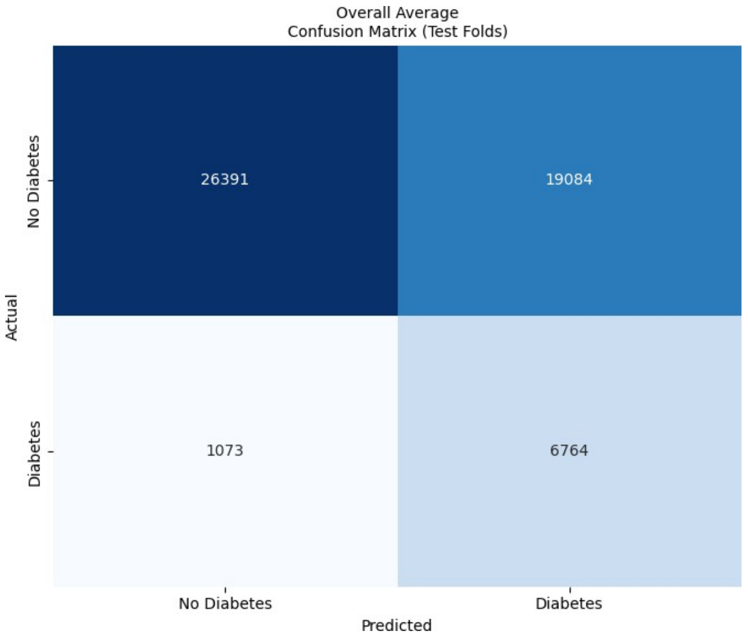
on the left side of that of model B with no overlap). In that case, model A will always outperform model B at all possible cost ratios. The outside model generalized cost function (CWA) can only help pick a better model among existing models but cannot make a model perform better.

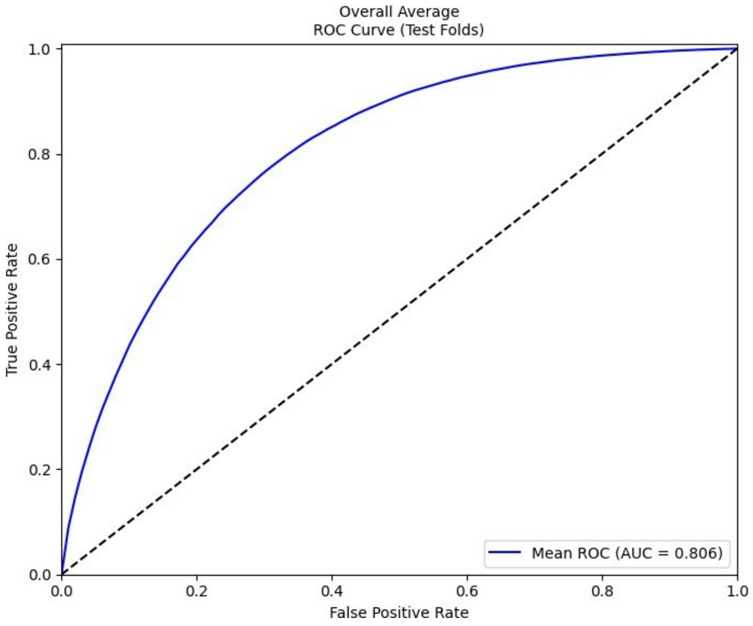
The Cost-weighted Accuracy (CWA) is a performance metric just like accuracy or F1-score. CWA can quantify the trade-off between false positives and false negatives in threshold tuning and model comparison but it cannot improve fundamental model performance. I have added section 4.6 (Comparison to Existing Metrics) to improve clarity.

6. This is a very interesting idea. But it seems the dual-model approach (one model with very high sensitivity and another model with very high specificity, say 90% for both) will likely result in two models giving opposite results most of the time. As the two models disagree in most cases, the final decision will be left to human judgment. To make an extreme example, two trivial models are used. One model always predicts true to minimize false negatives to zero and another model always predicts false to minimize false positives to zero. It will be up to the physician or the individual to pick which model to trust or which model they want to trust, rendering both models impractical.

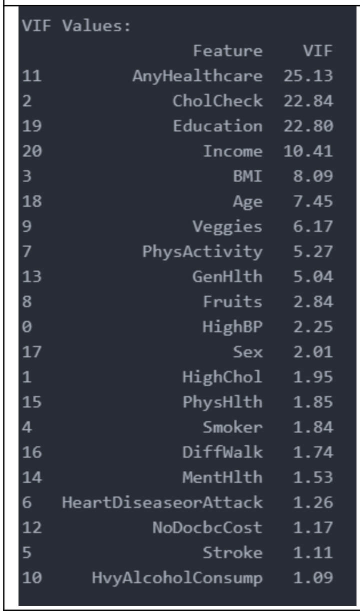
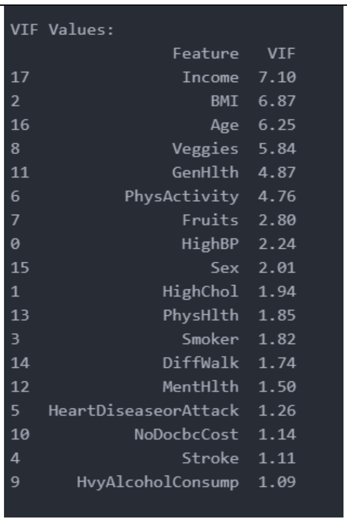
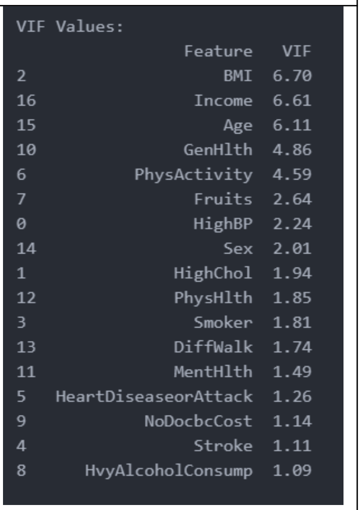
7. I have now run 5-fold cross-validation for the three models. Results were presented in section 5.2.

8. Confusion matrix and AUC curve, both averaged over five folds, for the neural network were included below. Both can be considered cost-unweighted as neither cost ratio or WAC has come into the picture.


Overall Average
Confusion Matrix (Test Folds)
Actual
No Diabetes
Diabetes
26391
19084
1073
6764
No Diabetes
Diabetes
Predicted


Overall Average
ROC Curve (Test Folds)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC (AUC = 0.806)

9. I have included the heat map and the VIF analysis in the manuscript (section 2.2 Feature Selection). I have managed to drop 4 features out of 21 features. After running a VIF analysis for 21 features, I dropped 3 features with VIF > 20. I ran a VIF again on the remaining 18 features and found 4 features with VIF > 5: Income (7.1), BMI (6.9), Age (6.2), and Veggies (5.8). Given that Income, BMI and Age are well-known clinically significant risk factors for diabetes, they were retained. Feature 'Veggies' was dropped and another VIF analysis was performed on the remaining 17 features. The 3 sets of VIF values are presented below.

21 features	**18 features**	**17 features**
		

```
VIF Values:
                 Feature    VIF
11          AnyHealthcare  25.13
2               CholCheck  22.84
19              Education  22.80
20                 Income  10.41
3                     BMI   8.09
18                    Age   7.45
9                 Veggies   6.17
7             PhysActivity   5.27
13                GenHlth   5.04
8                  Fruits   2.84
0                  HighBP   2.25
17                    Sex   2.01
1                HighChol   1.95
15               PhysHlth   1.85
4                  Smoker   1.84
16               DiffWalk   1.74
14               MentHlth   1.53
6    HeartDiseaseorAttack   1.26
12            NoDocbcCost   1.17
5                  Stroke   1.11
10      HvyAlcoholConsump   1.09
```

```
VIF Values:
                 Feature   VIF
17                Income  7.10
2                    BMI  6.87
16                   Age  6.25
8                Veggies  5.84
11               GenHlth  4.87
6           PhysActivity  4.76
7                 Fruits  2.80
0                 HighBP  2.24
15                   Sex  2.01
1               HighChol  1.94
13              PhysHlth  1.85
3                 Smoker  1.82
14              DiffWalk  1.74
12              MentHlth  1.50
5   HeartDiseaseorAttack  1.26
10           NoDocbcCost  1.14
4                 Stroke  1.11
9      HvyAlcoholConsump  1.09
```

```
VIF Values:
                 Feature   VIF
2                    BMI  6.70
16                Income  6.61
15                   Age  6.11
10               GenHlth  4.86
6           PhysActivity  4.59
7                 Fruits  2.64
0                 HighBP  2.24
14                   Sex  2.01
1               HighChol  1.94
12              PhysHlth  1.85
3                 Smoker  1.81
13              DiffWalk  1.74
11              MentHlth  1.49
5   HeartDiseaseorAttack  1.26
9            NoDocbcCost  1.14
4                 Stroke  1.11
8      HvyAlcoholConsump  1.09
```

10. I ran the correlation analysis of the average of the 17 features normalized with MinMaxScaler versus the diabetes/non-diabetes for the entire dataset (over 236,000 data points). The correlation coefficient was 0.2757.

Thank you for addressing my comments. Accepted. HOWEVER, you will need to provide a word doc of your manuscript formatted per the Journal's guidelines for the staff to begin copyediting. Please do so ASAP at the discussion board or upload.