Predict Diabetes From Medical Records – Neural Network model investigation

Deirdre Boland 28 Feb 2025

Dataset and Objective

Dataset

- Several medical features given with the output of whether patient is diabetic or not.

Objective

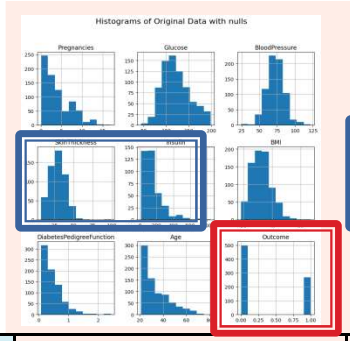
- Compare performance of an artificial neural network (ANN) to other supervised classification models to predict diabetic or not

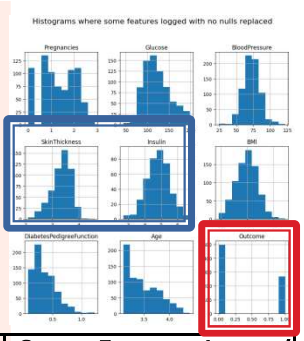
- Reference – Kaggle: https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records/input?select=diabetes.csv

Data exploration and cleaning

- 1 target **Outcome**:
  - ~2:1 ratio non-diabetic (0) to diabetic (1)
- 8 predictor features – no nulls
- Initial assessment:
  - Illogical 0 values for several features -> likely missing data -> 0 values made null
    - **Glucose:** Plasma glucose concentration after 2 hours in an oral glucose tolerance test
    - **Blood Pressure**: Diastolic blood pressure (mm Hg)
    - **Skin Thickness**: Triceps skin fold thickness (mm)
    - **Insulin:** 2-Hour serum insulin (mu U/ml)
    - **BMI**: Body mass index (weight in kg/(height in m)^2)
  - No changes made to values:
    - **Pregnancies** state number of times pregnant but can't distinguish 0 values into male or female or missing data
    - **Diabetes Pedigree function** – assess the genetic predisposition of an individual to diabetes based on their family history of disease.
    - **Age**
- Compared non-logged and logged of skewed features
- Replaced nulls with median/mean depending on skew post logging if applicable:
  - A lot of replacements affected correlation

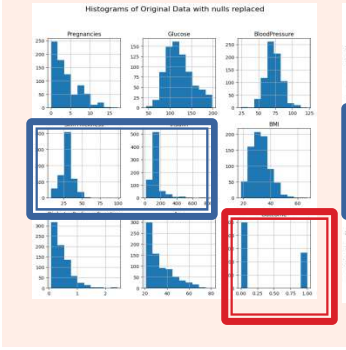
nulls

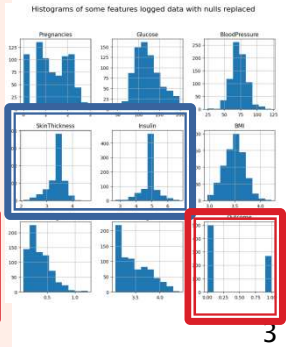




Correlation matrix	Original data Outcome		*Some Features Logged* Outcome	
	nulls	nulls filled	nulls	nulls filled
Pregnancies	*0.22*	*0.22*	*0.18*	*0.18*
Glucose	0.49	0.49	0.49	0.49
BloodPressure	0.17	0.17	0.17	0.17
SkinThickness	*0.26*	*0.21*	*0.26*	*0.22*
Insulin	*0.30*	*0.20*	*0.35*	*0.25*
BMI	*0.31*	*0.31*	*0.32*	*0.32*
Diabetes Pedigree Function	*0.17*	*0.17*	*0.18*	*0.18*
Age	0.24	0.24	0.27	0.27
Outcome	1.00	1.00	1.00	1.00

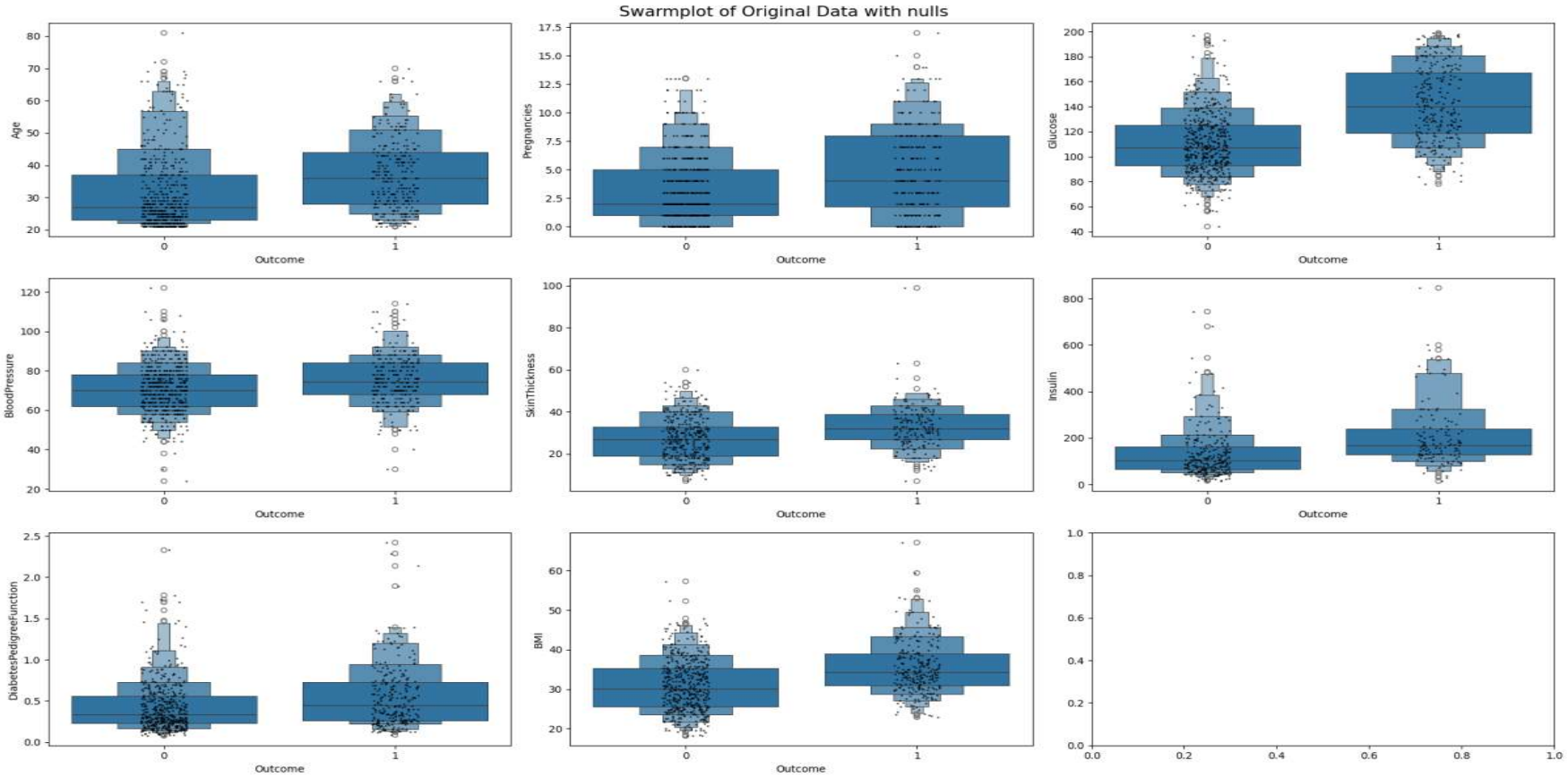
Nulls filled





Data visualisation

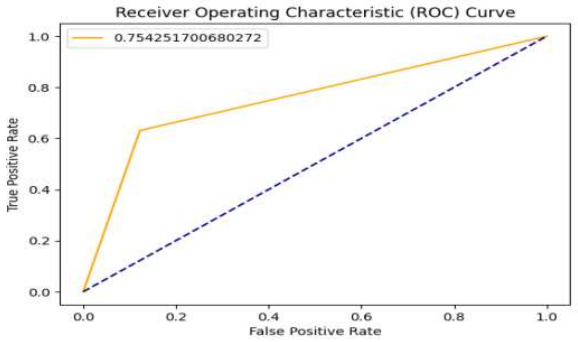
- Features by 0 non-diabetic and 1 diabetic:
  - Diabetic data tends to skew higher

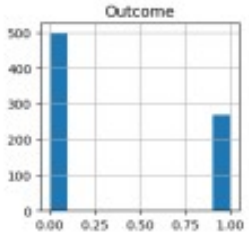


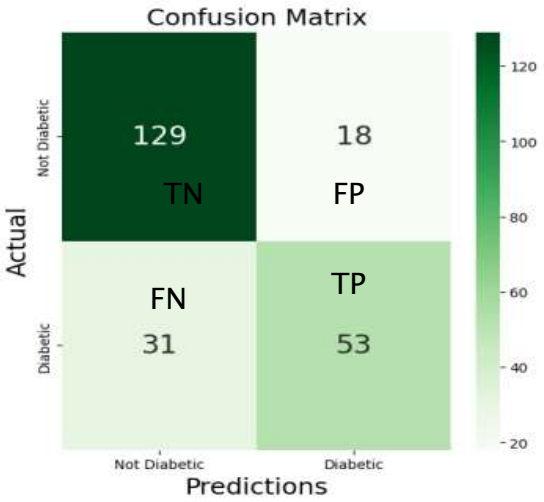
Model performance metrics

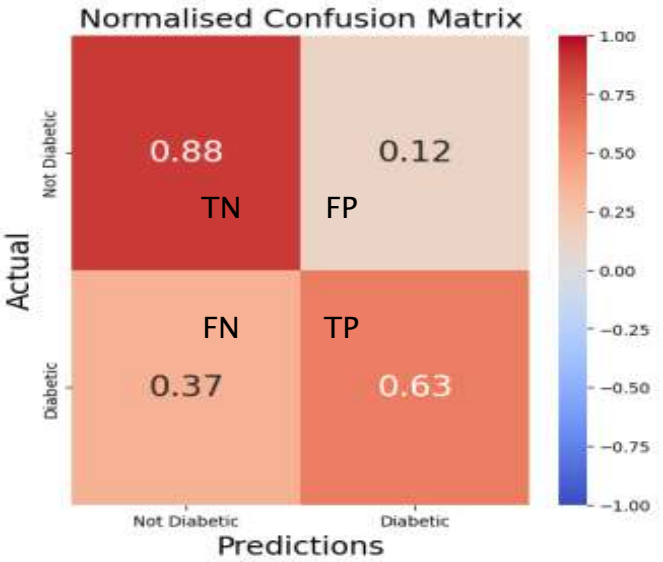
- Model performance comparison AUC score
  - Area under ROC curve (AUC) plots the **sensitivity** /recall (true positive = TP/(TP+FN)] and the **specificity** (false positive = TN/TN+FP).
  - 0 (poor) to 1 (perfect).
- Data Z- scaled before modelling
- 70/30 train/test split
- Supervised classification models:
  - **Logistic regression (LR) model**:
    - Null filled logged features (AUC: 0.754) better than null filled non-logged (AUC: 0.729)
  - LR better than null filled logged features **SVM** (AUC linear: 0.749 AUC, sigmoid: 0.699 AUC, rbf: 0.691)

Logistic regression model evaluation:









Do neural networks predict better?

- Logistic regression (LR) model AUC: 0.754
- Sequential ANN models:

most ->---------- > ---------- > least complex

3 hidden layers [8 – (50–20–4) – 1] 1 hidden layer [8 – (16) – 1] Binary preceptron

Layer (type)	Output Shape	Param #
dense_76 (Dense)	(None, 50)	450
dense_77 (Dense)	(None, 20)	1,020
dropout_8 (Dropout)	(None, 20)	0
dense_78 (Dense)	(None, 4)	84
dropout_9 (Dropout)	(None, 4)	0
dense_79 (Dense)	(None, 1)	5

Total params: 4,679 (18.28 KB)

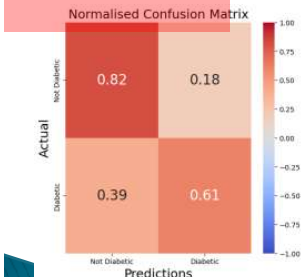
Trainable params: 1,559 (6.09 KB)

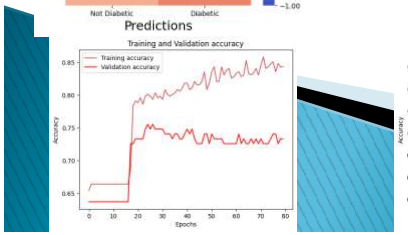
Non-trainable params: 0 (0.00 B)

Optimizer params: 3,120 (12.19 KB)

80 epochs

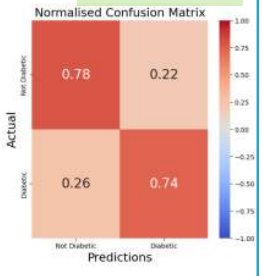
AUC: 0.712

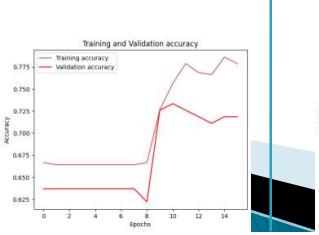




Early stop16 epochs

AUC: 0.760





Layer (type)	Output Shape	Param #
dense_94 (Dense)	(None, 16)	144
dense_95 (Dense)	(None, 1)	17

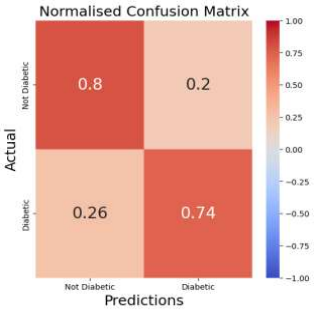
Total params: 485 (1.90 KB)

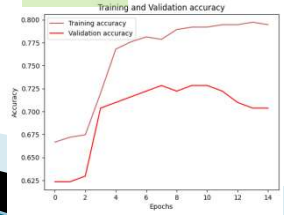
Trainable params: 161 (644.00 B)

Non-trainable params: 0 (0.00 B)

Optimizer params: 324 (1.27 KB)

AUC	Last epoch
0.773	13
0.770	14
0.764	25
0.767	13
0.761	35
0.767	15





Model re-run 6 times with early stop

Layer (type)	Output Shape	Param #
dense_101 (Dense)	(None, 1)	9

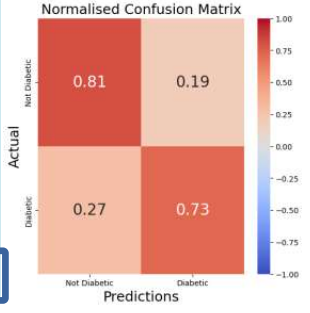
Total params: 29 (120.00 B)

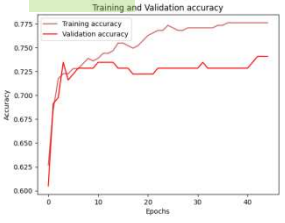
Trainable params: 9 (36.00 B)

Non-trainable params: 0 (0.00 B)

Optimizer params: 20 (84.00 B)

AUC	Last epoch
0.762	44
0.762	44
0.759	43
0.759	44
0.759	42
0.768	45



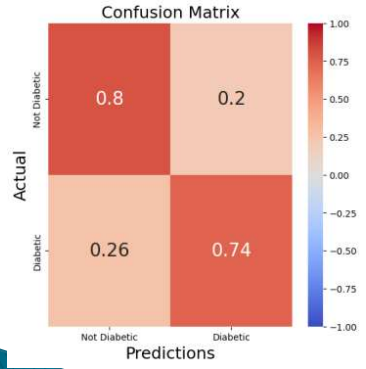


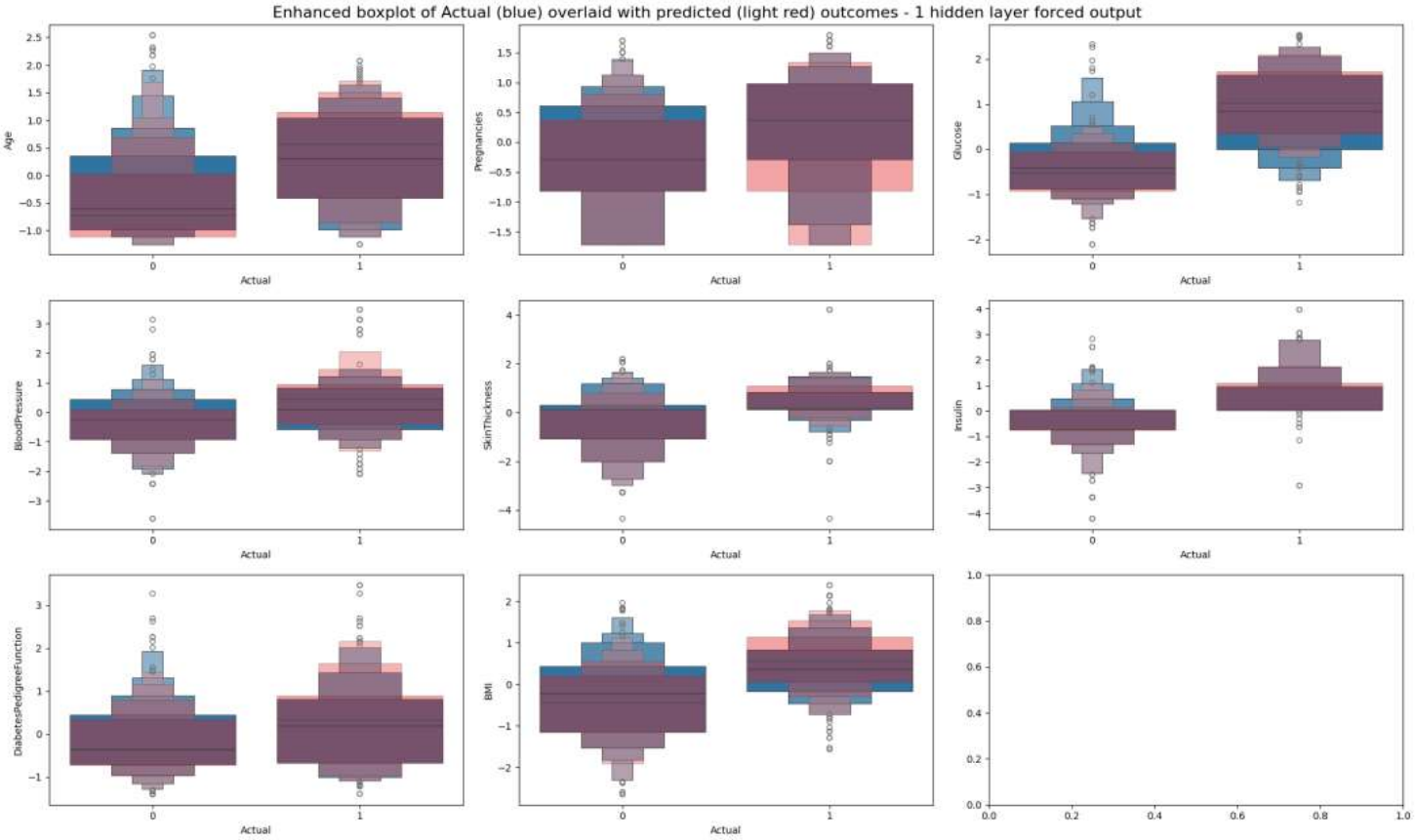
Model re-run 6 times with early stop

Exploring an ANN prediction

- 1 hidden layer [8 – (16) – 1] Epoch 13 AUC: 0.767
- Forced model to give consistent output with random seed

- Actual (blue), predicted (light red), overlap (purple)
- Predicted tends to skew low for non-diabetic (0 – red below/low purple overlap) and high for diabetic (1 – red above/high purple overlap) except diabetic (1) pregnancies

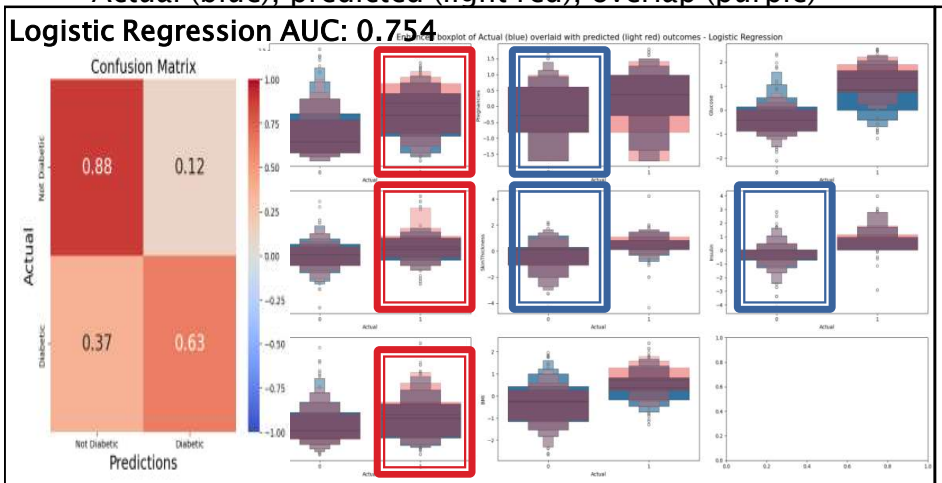




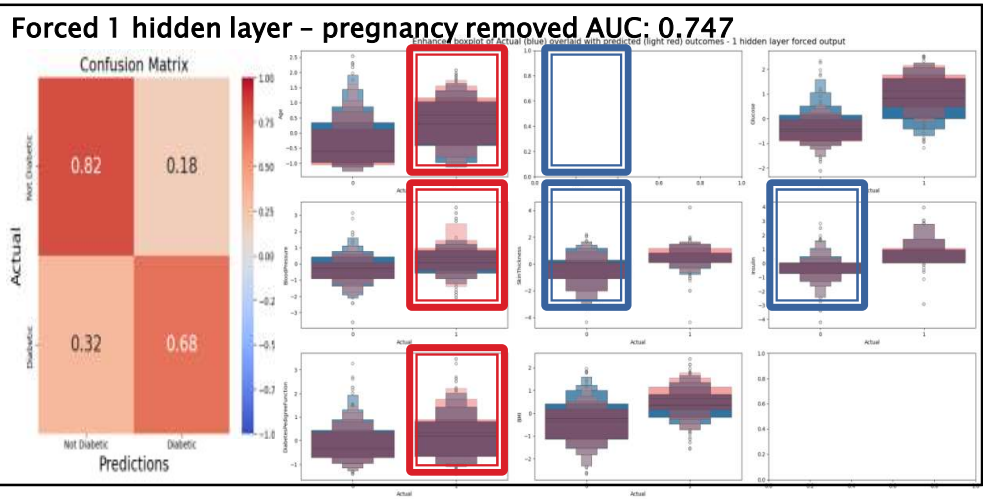
Comparing model predictions

- Actual (blue), predicted (light red), overlap (purple)

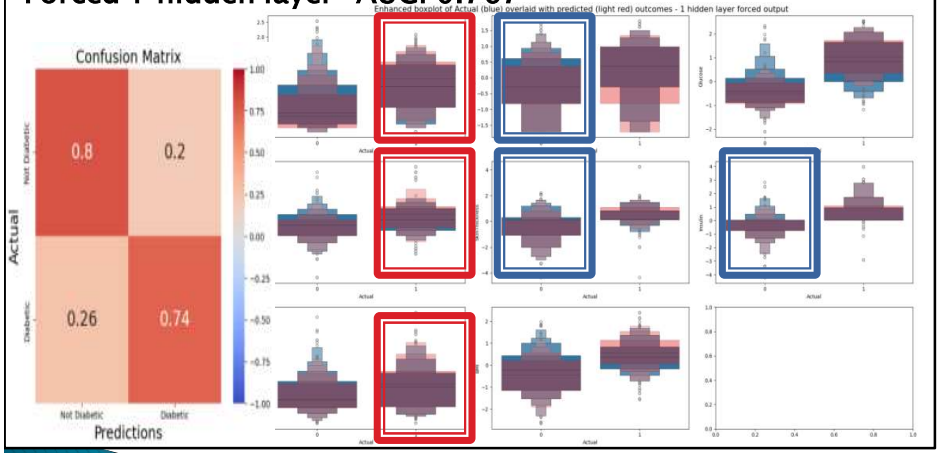
Logistic Regression AUC: 0.754



Forced 1 hidden layer – pregnancy removed AUC: 0.747



Forced 1 hidden layer –AUC: 0.767



- Pregnancy impacts performance for models shown:
  - without pregnancy ANN more similar to logistic regression (2 models above)
- Predicting non-diabetic (0 outcome) 2 models above better :
  - almost perfect layover predicted and actual for some features vs. model on left, see more blue
- Predicting diabetic (1) model to left best:
  - Less red predicted above the overlap (purple)on the 1 outcome (right) vs. 2 models above

Conclusion

- Conclusion
  - Simple neural network models with early stopping generally better prediction than linear regression/SVM
    - Takes longer time to compute for not too major an improvement
  - Prediction performance could be further improved:
    - Best models will give 20-25% false positive or false negative
- Future work
  - Better way to handle illogical 0 data e.g. Glucose, insulin?
  - Pregnancy data appeared to improve one ANN model:
    - More runs to confirm
    - More data on gender as there are differences in BMI and hormone regulation
  - Tweak parameters of ANN models to improve accuracy

Q&A

- Thanks for your attention!