Diabetes Prediction

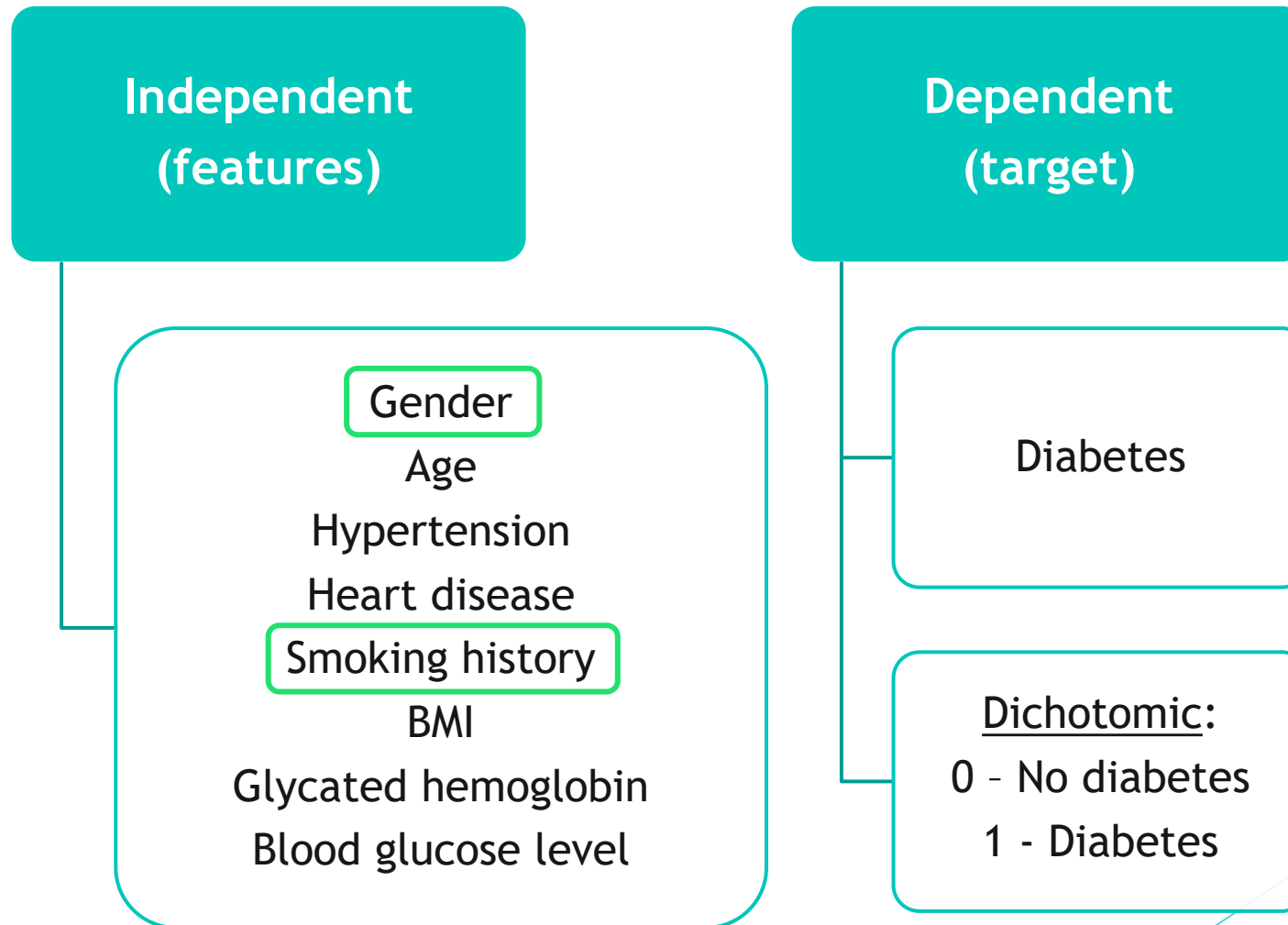## Mini Project - Week 7

Lara Castanheira

First Steps

- Understand the data set

- Define the dependent variable (**target**) and the independent variables (**features**)

- Data cleaning
  - Check for null values and drop
  - Check for duplicates and drop

- Convert categorical variables into dummy variables

Variables

Independent (features)

- Gender
- Age
- Hypertension
- Heart disease
- Smoking history
- BMI
- Glycated hemoglobin
- Blood glucose level

Dependent (target)

Diabetes

Dichotomic:
0 – No diabetes
1 - Diabetes

Model – Logistic Regression

▶ Division of data - 80% for training and 20% for testing

▶ Data normalisation

▶ Train the model with normalised data

▶ Make predictions on the test set

▶ Evaluate the accuracy of the model (0.959) – When the model is used to make predictions, approximately 95.9% of these predictions are correct

▶ Check if the model is overfitting (accuracy – 0.958 = 0.001)

Model – Logistic Regression

▶ **ROC Curve:** is a graphical tool used to evaluate the performance of a binary classification model
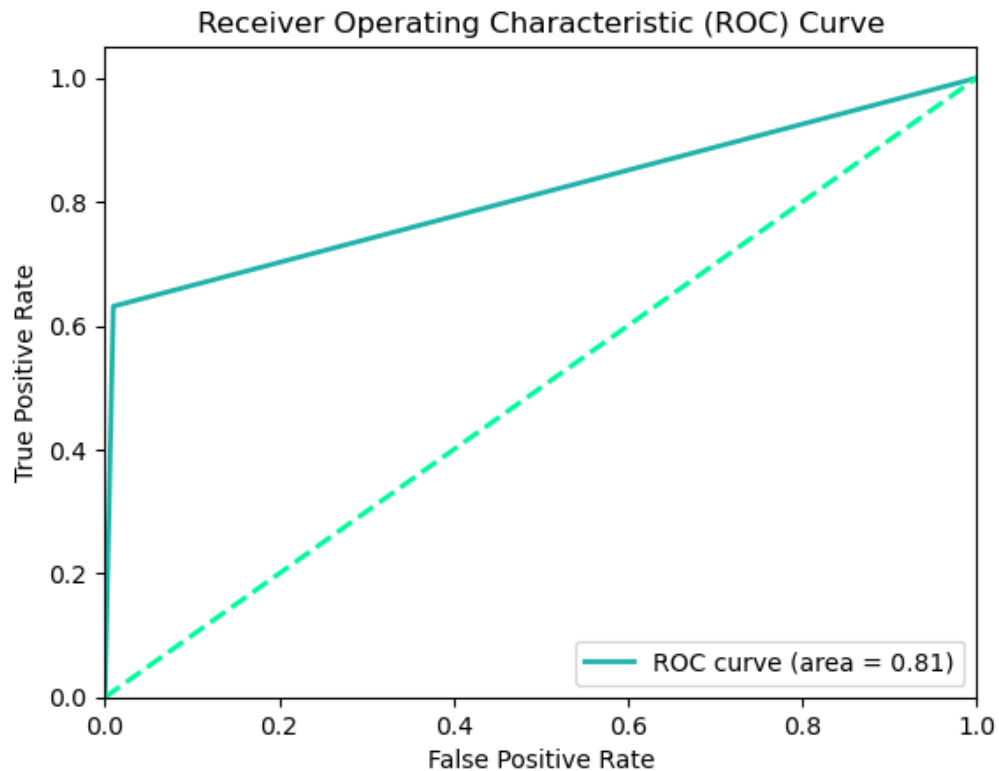
AUC (0.81)- between 0.8 and 0.9

↓

Very good

↓

solid ability to distinguish between positive and negative classes

Model – Logistic Regression

▶ Confusion Matrix

	Predicted	
Observed	**0**	**1**
0	True Negative - 1053	False Postive - 169
1	False Negative - 615	True Positive - 17393

▶ The model performs well overall, but it needs improvement, especially in its ability to correctly identify **positive cases**

Model – Logistic Regression - Conclusions

▶ Precision: ≈0.86 ⟶ indicates a high proportion of cases correctly identified as positive

▶ Recall: ≈0.63 ⟶ suggests that the model may be missing some positive instances

▶ F1-score: ≈0.72 ⟶ shows a good balance between precision and recall

▶ Error rate: ≈0.04 ⟶ requires more investigation to identify and correct possible flaws in the model

Next Steps

- Detailed descriptive statistics

- Bivariate analysis

- Hosmer & Lemeshow test

- Wald test
  - Odds ratio

$$Odds = \frac{Probability\ of\ event}{Probability\ of\ non\text{-}event}$$

		Outcome	
		Yes	No
Predictor	Yes	A	B
	No	C	D

$$OR = \frac{(A*D)}{(B*C)}$$

Thank you!