Assignment 1 – Random Forest and AutoML

The dataset for this project is the “Pima Indians Diabetes Database”. This dataset is available here - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
(Note the text below is taken from the Kaggle website)

The Data
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The original paper to cite:
Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

Your challenge
Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

Constraints: I wish to see you attempt the challenge two ways:
Method 1: Using a Random Forest classifier
Method 2: Using an automated ML method via PyCaret

You will write up your work as a Jupyter notebook which you must contain clear commentary and code and should be executable by your examiner. I would like to see that you provide some commentary which gives the reader an idea of how AutoML works in general and how PyCaret works particularly (you can spend around 1000 words or 2 “A4 pages” or so doing that if you like – use your own words and thoughts here).
You will add a discussion section to this notebook in which you highlight the differences you see between your automated approach and your original hand engineered random forest attempt (500 word limit for this final discussion section – not a hard limit by the way). You will be providing a critical evaluation of the two approaches here.