# Project describtion document

For the classification dataset :

**General information on dataset:**

Name : diabetes\_prediction\_dataset

Number of classes : 9

Number of rows: 100K

Number of samples (training ): from 85 to 80 %

Number of samples (testing ): from 15 to 20%

**Implementation details:**



At feature extraction : all features been used (8) around 65\*8

Cross validation : 5 folds

Hyperparameters : criterion='gini',max\_depth=5,min\_samples\_leaf=1,min\_samples\_split=2

Gridsearch : used



New features extracted with encoding techniques

**Result details:**



**A screenshot of a computer code

Description automatically generated**



1. **Confusion Matrix**:
   * + True Positives (TP): 457
     + True Negatives (TN): 8508
     + False Positives (FP): 41
     + False Negatives (FN): 622

**Precision**:

Precision answers the question: “What proportion of positive identifications was actually correct?” {TP}{TP + FP}

**Recall**:

Recall answers the question: “What proportion of actual positives was identified correctly?” {TP}{TP + FN}

test Accuracy: 93.11383464894058

train Accuracy: 93.24180658406041