DIABOT

PROTECTING YOUR HEALTH THROUGH EARLY DETECTION

By Group Morningstar

CONTENT

ABOUT DIABOT

THE TOPIC & GOALS

The DiaBot machine learning model aims to assist medical staff in detecting not only the presence of Diabetes in at-risk patients, but also identify patients which are at risk of developing it.

THE DATASET

To achieve the goals above, this first instance of the DiaBot model scoured through a total of **230.000 patients**, learning how to best identify at-risk patients.

HANDLING THE DATA

DATASET

In order to achieve our results, we processed the data of 230.000 patients, including:

- Elimination of ambiguous data
- Scaling of numerical data

Our data is composed of 22 entries per patient, ranging from lifestyle habits, to major health indicators, such as, but not limited to, drug comsumption and health stats.

VARIABLES

blood pressure, cholesterol, BMI, smoking habits, stroke history, heart disease, physical activity, eating habits, drug use, healthcare access, mental and physical care, gender, income and education

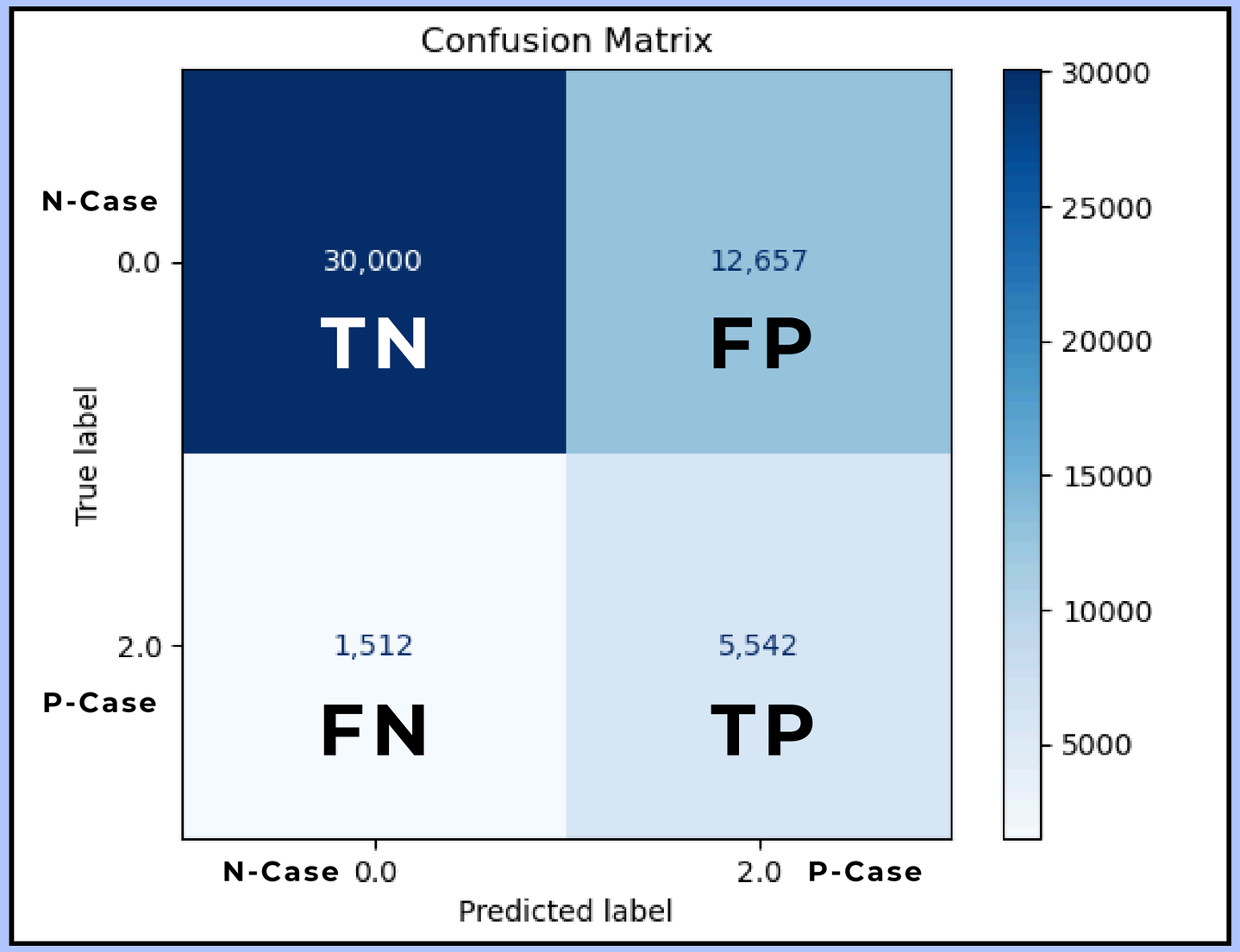
FEATURE ENGINEERING

PRIMARY TECHNIQUES

To ensure the best prediction quality for our DiaBot model, we have applied exhaustive feature selection, such as:

- Undersampling (Equalizing Group Share)
- Aggressive Hyperparameter Optimization
- Comparison between multiple Model types
  - K-Nearest Neighbors
  - Random Forest (Classifier)
  - Bagging Ensemble

CONFUSION MATRIX

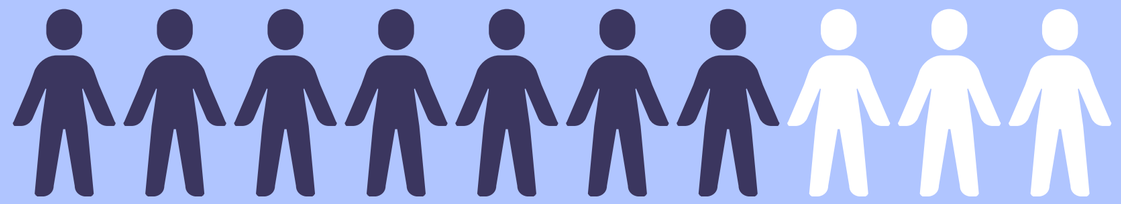


- **TN: 30,000** – Correctly predicted "N" when true label was "N"
- **FP: 12,657** – Predicted "P" when true label was "N"
- **FN: 1,512** – Predicted "N" when true label was "P"
- **TP: 5,542** – Correctly predicted "P" when true label was "N"

STATISTICS

72%

ACCURACY - ALL



STATISTICS - P CASES

72%

ACCURACY - ALL

79%

RECALL - CASE P

31%

PRECISION - CASE P

Accuracy: Out of all my picks, how many were the right color (both red and not red)?

Recall: Of all the red balls in the box, how many did I find?

Precision: Of the balls I picked, how many are actually red?

STATISTICS - N CASES

72%

ACCURACY - ALL

70%

RECALL - CASE P

95%

PRECISION - CASE P

Accuracy: Out of all my picks, how many were the right color (both red and not red)?

Recall: Of all the red balls in the box, how many did I find?

Precision: Of the balls I picked, how many are actually red?

MODEL PERFORMACE

*"The **DiaBot** model reliably rules out non-diabetes cases (95% precision for negatives) but needs additional tests to confirm diabetes cases due to a high rate of false positives (30% precision for positives), with an overall accuracy of 71%."*

P-CASES

POSITIVE PATIENTS

High recall (79%) ensures the **model catches most true diabetes cases**, making it valuable for screening and identifying at-risk patients.

However, with **low precision (31%)**, many people are falsely flagged as having diabetes, requiring additional testing for confirmation.

N-CASES

NEGATIVE PATIENTS

High precision (95%) means the model is very reliable at identifying non-diabetes cases, minimizing false positives.

However, with a **recall of 70%**, it may miss some true negative cases, potentially leaving a few non-diabetes patients flagged for further review.

CONCLUSION

While the first interation of the DiaBot shows very promising results, **more time** would have allowed for even **more extensive tuning**, as the necessary **computing power is already available**.

Future versions of the DiaBot would be even **more reliable handling positive cases**, and assist our doctors to enact preventive care or interventions early, **in addition** to its ability to reliabily **put true negative patients at ease.**

THANK YOU!