RMIT UNIVERSITY

Assignment 1

Prof: Dr. Nguyen Thien Bao

Name: Loc Vu

Student Number: S3891483

1. Project overview

According to WHO [1], “diabetes is a disease that occurs either when the pancreas does not produce enough insulin (a hormone that regulates blood glucose) or when the body cannot effectively use the insulin it produces. Hyperglycaemia or raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body’s systems, especially the nerves and blood vessels.”
The goal of this project is to use machine learning methods to predict the onset of diabetes based on various characteristics. We will need to analyze, develop, and assess machine learning models that can accurately predict whether a patient will develop diabetes. We will be working with a dataset that includes various health metrics and demographic information, with the target variable being the presence or absence of diabetes. Our main tasks are as follows:

2. Strategy

2.1. Exploratory data analysis (EDA)

2.1.1. We will conduct a comprehensive exploratory data analysis to understand the distribution of features, identify correlations, and visualize patterns in the dataset.

2.2. Data Processing

2.2.1. We will handle missing values, outliers, and carry out necessary data preprocessing steps.

2.3. Model Development

2.3.1. We will propose three different machine learning models, including logistic regression, decision tree, and random forest, for predicting the onset of diabetes in patients.

2.4. Model Implementation (Hyper parameter tuning)

2.4.1. We will fine-tune hyperparameters using cross-validation techniques to optimize model performance, grid search tuning, and post-pruning with cost complexity pruning.

3. Evaluation

F1 Score: The F1 score is the harmonic mean of precision and recall while, precision is the ratio of correct prediction to the positive observation to the total predicted positives, and

recall is the ratio of correct prediction to the positive observation to the total observations in the actual class. We used the below formula for our calculation [2]

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2 \times tp}{2 \times tp + fp + fn}$$

Reasoning: Due to the nature of our project of predicting diabetes patients, false positive predictions (false predictions) should be as low as possible. Therefore, F1 score could handle our outcomes with its representation of both average of precision and recall.

Accuracy: "A metric that measures how often a machine learning model correctly predicts the outcome." A more simplified definition of accuracy is the proportion of total predictions that are correct. [4]

$$Accuracy = \frac{Correct\ Prediction}{Total\ Prediction}$$

Reasoning: Accuracy metric is a good measure when all the classes are balanced, however it can be misleading if the classes are not. Therefore, when taking accuracy into account, we will need to balance out all the classes.

Precision: A metric that shows how precise a model predicts the true positive. It is the ratio of true positives to total of true positives and false positives. [4]

$$Precision: \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Reasoning: The precision metric is a NOT good measure when the cost of false positives is high, like our project, the cost of falsely predicting a diabetes patient could cause many harms to them and may lead to late treatment.

Recall: Recall metric is a ratio of the actual positive cases, or how often it could correctly predict positive cases where the objective of the project is to reduce false negative cases. [4]

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Reasoning: Recall is a good measure when the cost of False Negatives is high, for instance with our project as mentioned above, the cost of falsely predicting an unhealthy patient could potentially harm them.

4. Data Overview

Training Data (data_train.csv): The data is stored in data_train.csv, is the primary dataset for our diabetes prediction task. Each row in this dataset represents an individual, with various features that include health metrics and demographic information. The status column in the dataset is the target variable, which indicates whether the individual has diabetes or not.
Test Data (data_test.csv): The test data, stored in data_test.csv, is used to evaluate the performance of our trained models. Like the training data, each row represents an individual and includes the same set of features. However, this dataset does not include the target variable (diabetes status), our task is to use our trained models to predict the diabetes status for each individual in this dataset.

5. Model Development

Upon initial analysis in the EDA section, both the test and training datasets appear to be evenly distributed. However, there are several skewed distributions and outliers present in the data. As a result, before we proceed with model development, we need to perform data scaling and outlier removal. These preprocessing steps are crucial to improving the performance of our machine learning models.

Logistic Regression with Linear Features: [5] "The model that estimates relationship between one independent variable and one dependent variable or target variable using a straight line."
Base model result (training performance):

- F1 Score: 0.89 - The model identifies positive samples correctly 90% of the time, while also minimizing false positives.
- Accuracy Score: 0.89 - The model makes correct predictions for 89% of all cases
- Recall Score: 0.89 - The model correctly classifies 89% of all actual positive cases, but misclassified 11% of them as negative.
- Precision Score : 0.89 - Among all cases that the model predicts as positive, 89% are indeed positive.

GridSearchCV tuned (training performance):

- F1 Score: 0.89 - The model identifies positive samples correctly 90% of the time, while also minimizing false positives.
- Accuracy Score: 0.89 - The model makes correct predictions for 89% of all cases

- Recall Score: 0.89 - The model correctly classifies 89% of all actual positive cases, but misclassified 11% of them as negative.
- Precision Score : 0.89 - Among all cases that the model predicts as positive, 89% are indeed positive.

Logistic Regression with Polynomial degrees of 2: with polynomial degrees of 2, the logistic regression model is more flexible and capable of fitting more complex patterns in the data.

Base model result (training performance):

- F1 Score: 0.90 - The model identifies positive samples correctly 90% of the time, while also minimizing false positives.
- Accuracy Score: 0.90 - The model makes correct predictions for 90% of all cases
- Recall Score: 0.90 - The model correctly classifies 90% of all actual positive cases, but misclassified 10% of them as negative.
- Precision Score : 0.90 - Among all cases that the model predicts as positive, 90% are indeed positive.

GridSearchCV tuned (training performance):

- F1 Score: 0.90 - The model identifies positive samples correctly 90% of the time, while also minimizing false positives.
- Accuracy Score: 0.90 - The model makes correct predictions for 90% of all cases
- Recall Score: 0.90 - The model correctly classifies 90% of all actual positive cases, but misclassified 10% of them as negative.
- Precision Score : 0.90 - Among all cases that the model predicts as positive, 90% are indeed positive.

Decision Tree: [6] “a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks.”

Base model result (training performance):

- F1 Score: 0.87 - The model identifies positive samples correctly 87% of the time, while also minimizing false positives.
- Accuracy Score: 0.88 - The model makes correct predictions for 88% of all cases
- Recall Score: 0.88 - The model correctly classifies 88% of all actual positive cases, but misclassified 12% of them as negative.

- Precision Score : 0.89 - Among all cases that the model predicts as positive, 89% are indeed positive.

GridSearchCV tuned (training performance):

- F1 Score: 0.98 - The model identifies positive samples correctly 98% of the time, while also minimizing false positives.
- Accuracy Score: 0.98 - The model makes correct predictions for 98% of all cases
- Recall Score: 0.98 - The model correctly classifies 98% of all actual positive cases, but misclassified 2% of them as negative.
- Precision Score : 0.98 - Among all cases that the model predicts as positive, 98% are indeed positive.

Random Forest: [7] "combines the output of multiple decision trees to reach a single result."

Base model result (training performance):

- F1 Score: 0.90 - The model identifies positive samples correctly 90% of the time, while also minimizing false positives.
- Accuracy Score: 0.90 - The model makes correct predictions for 90% of all cases
- Recall Score: 0.90 - The model correctly classifies 90% of all actual positive cases, but misclassified 10% of them as negative.
- Precision Score : 0.90 - Among all cases that the model predicts as positive, 90% are indeed positive.

GridSearchCV tuned (training performance):

- F1 Score: 1.00 - The model identifies positive samples correctly 100% of the time, while also minimizing false positives.
- Accuracy Score: 1.00 - The model makes correct predictions for 100% of all cases
- Recall Score: 1.00 - The model correctly classifies 100% of all actual positive cases, and does not miss any negative classes.
- Precision Score :1.00 - Among all cases that the model predicts as positive, 100% are indeed positive.

References

[1] “Diabetes.” World Health Organization. April 10th, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Overview,hormone%20that%20regulates%20blood%20glucose.

[2] “F1 Score in Machine Learning: Intro & Calculation.” V7labs. April 10th, 2024. [Online]. Available: https://www.v7labs.com/blog/f1-score-guide

[3] “How to explain the ROC curve and ROC AUC score?” Evidentlyai. April 10th, 2024. [Online]. Available: https://www.evidentlyai.com/classification-metrics/explain-roc-curve

[4] “Accuracy vs. precision vs. recall in machine learning: what's the difference?” Evidentlyai. April 14th, 2024. [Online]. Available: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall#:~:text=Accuracy%20is%20a%20metric%20that,often%20the%20model%20is%20right%3F

[5] S. Mondal. “Regression Analysis | Beginners Comprehensive Guide (Updated 2024)”. Analyticsvidhya.com. April 14th, 2024. [Online] Available: https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/#:~:text=Linear%20Regression%20is%20used%20to,Logistic%20regression%20provides%20discreet%20output.

[6] “What is a decision tree?” IBM. Apr. 14th, 2024. [Online]. Available: https://www.ibm.com/topics/decision-trees

[7] “What is random forest?” IBM. Apr. 14th, 2024. [Online]. Available: https://www.ibm.com/topics/random-forest

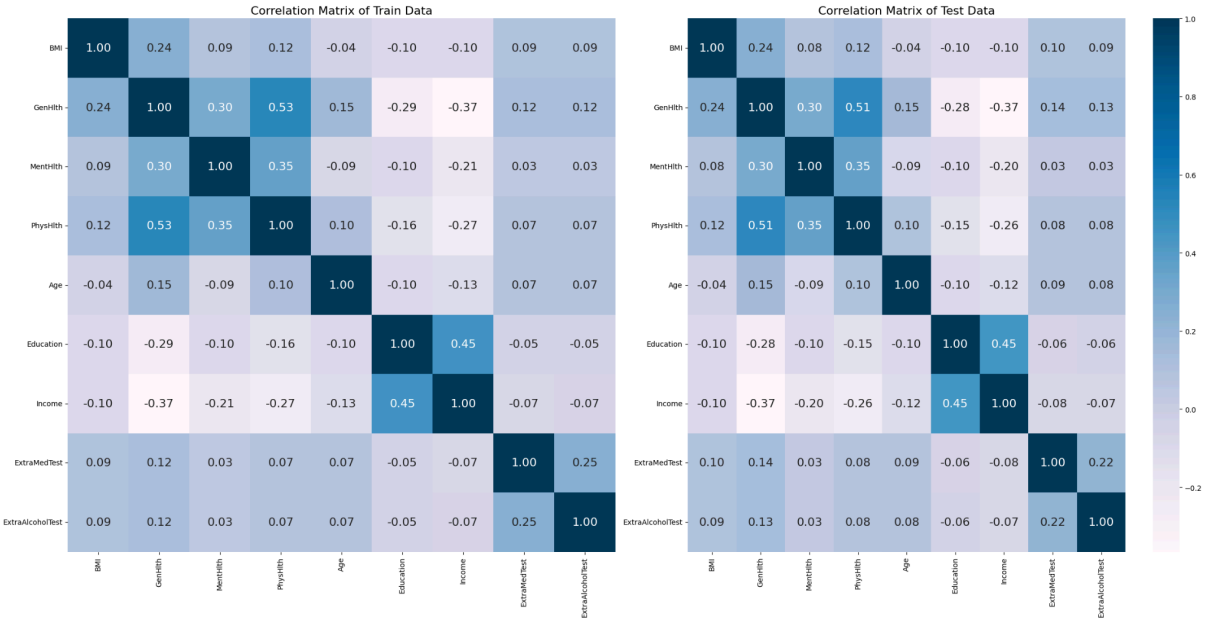

Fig. 1 Correlation Matrix

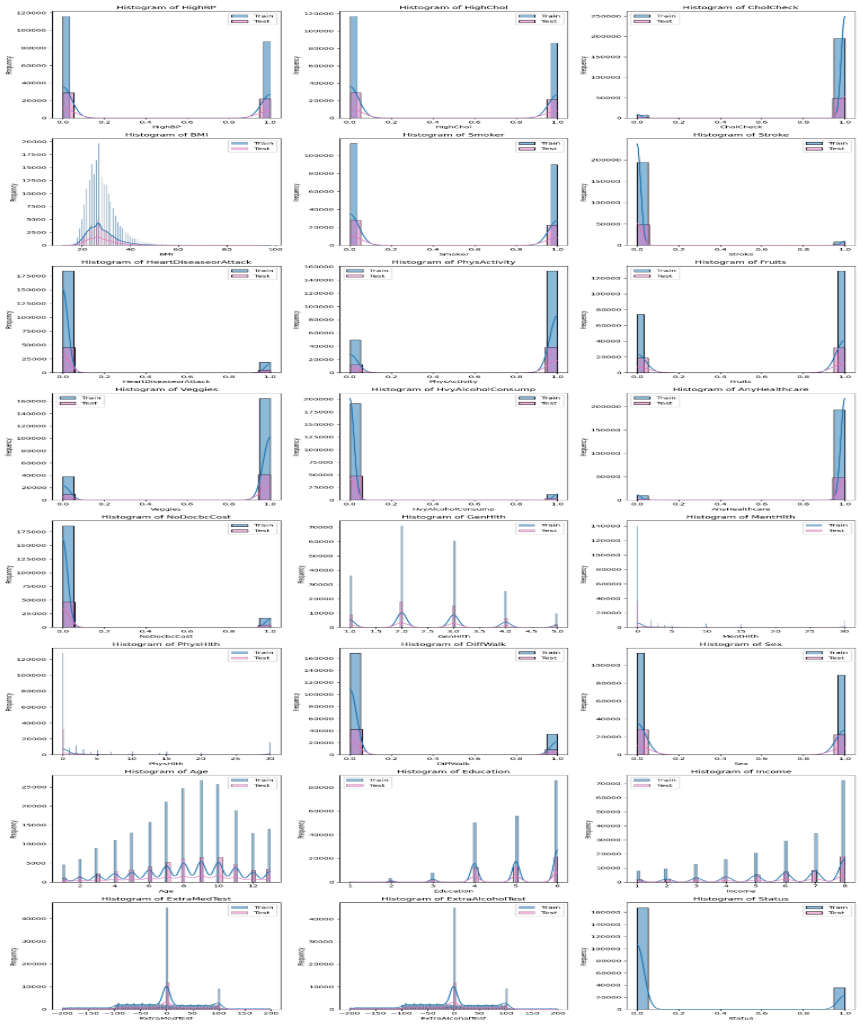

Fig. 2 Histogram Plot

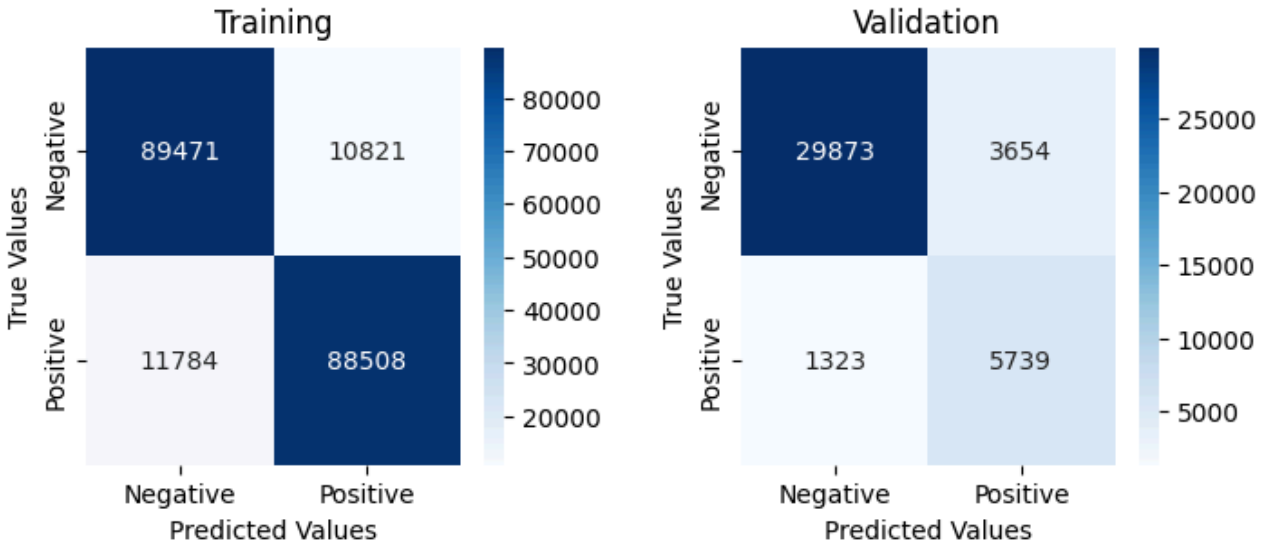

Fig. 3 Logistic Regression with Linear Features

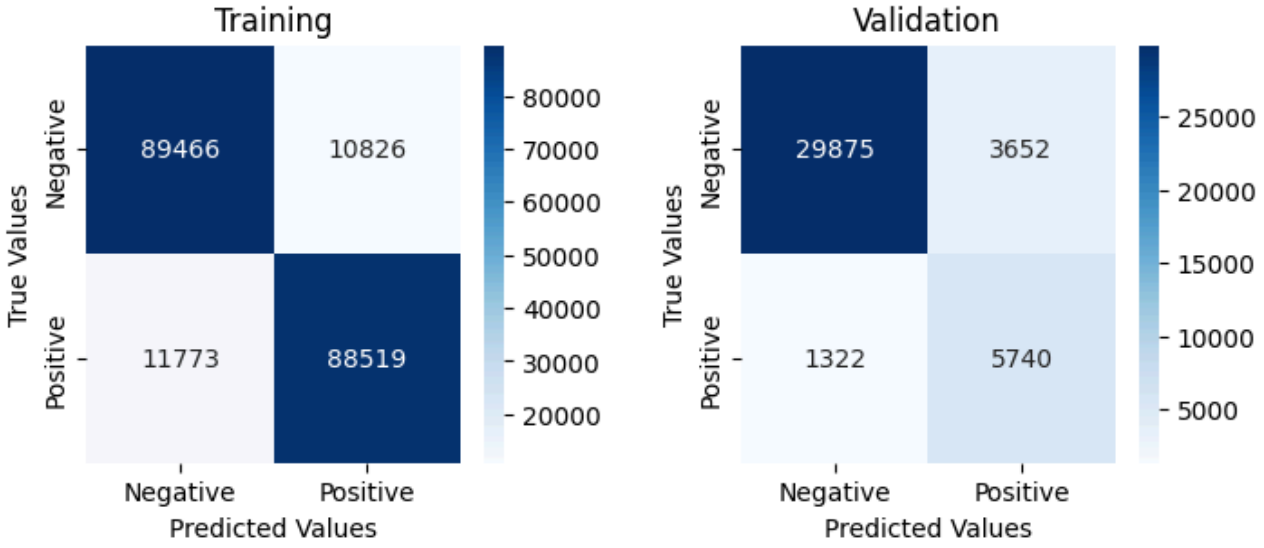

Fig. 4 Logistic Regression with Linear Features (GridSearchCV tuning)

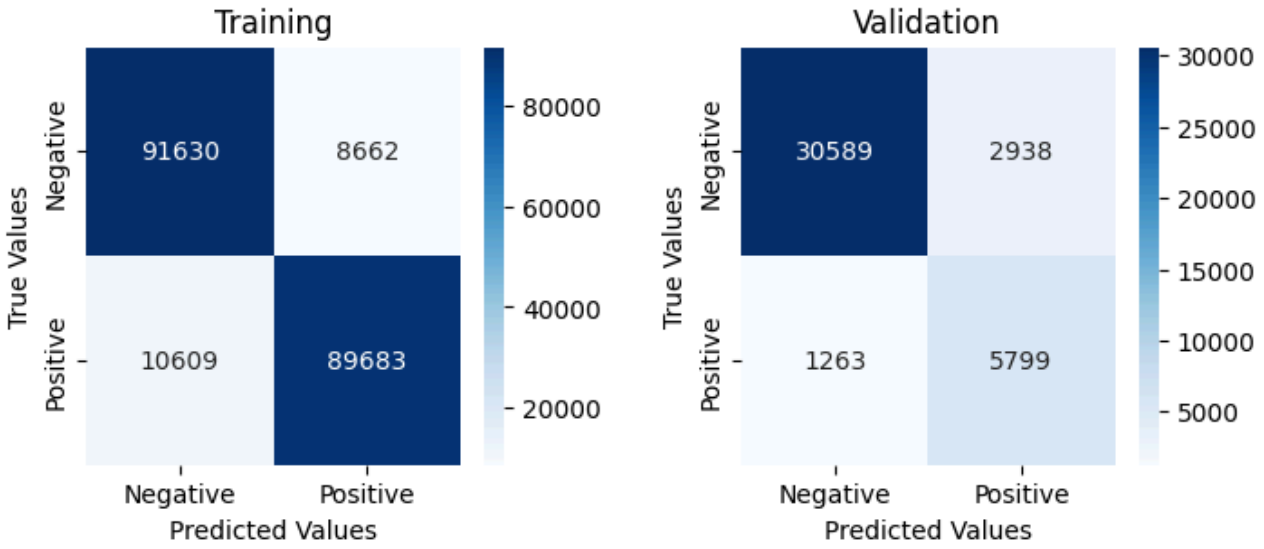

Fig. 4 Logistic Regression with Polynomial degrees of 2

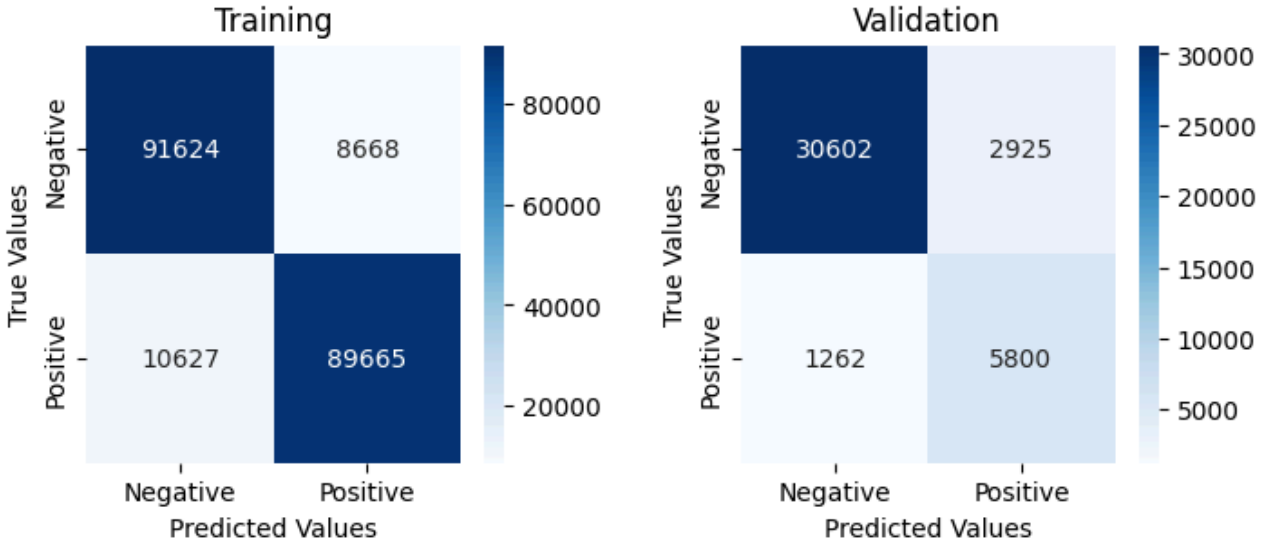

Fig. 5 Logistic Regression with Polynomial degrees of 2 (GridSearchCV tuning)

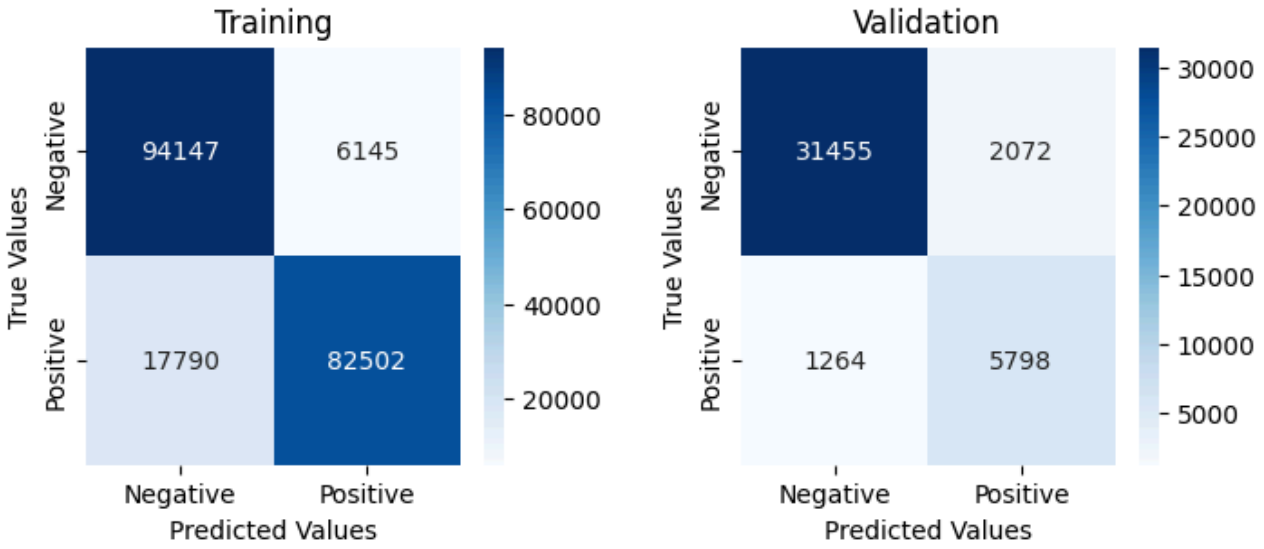

Fig. 6 Decision Tree

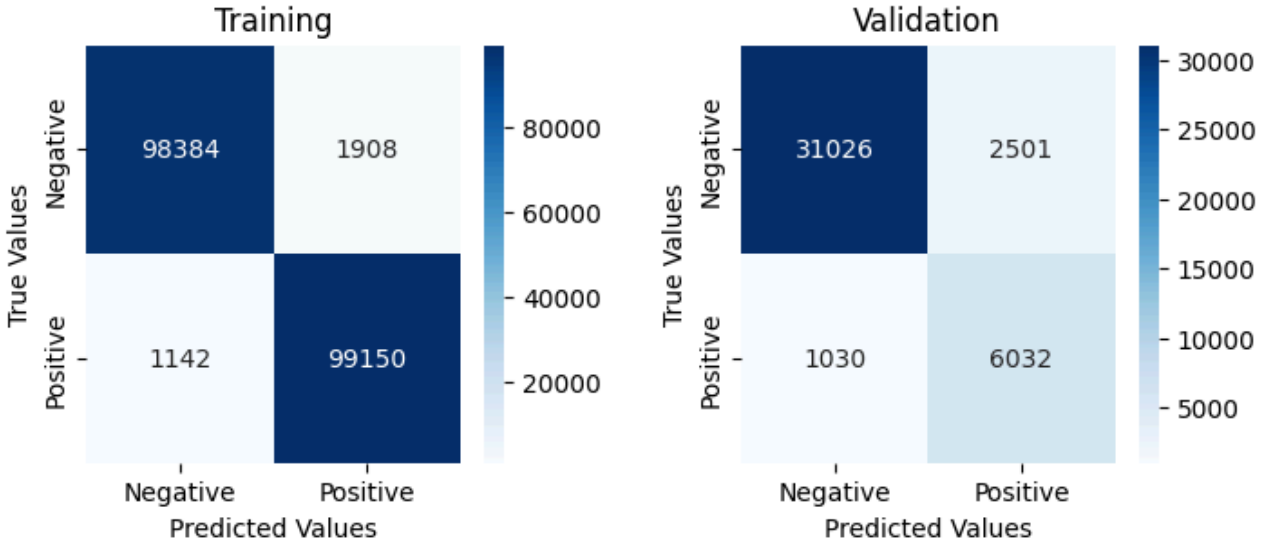

Fig. 7 Decision Tree (GridSearchCV tuning)

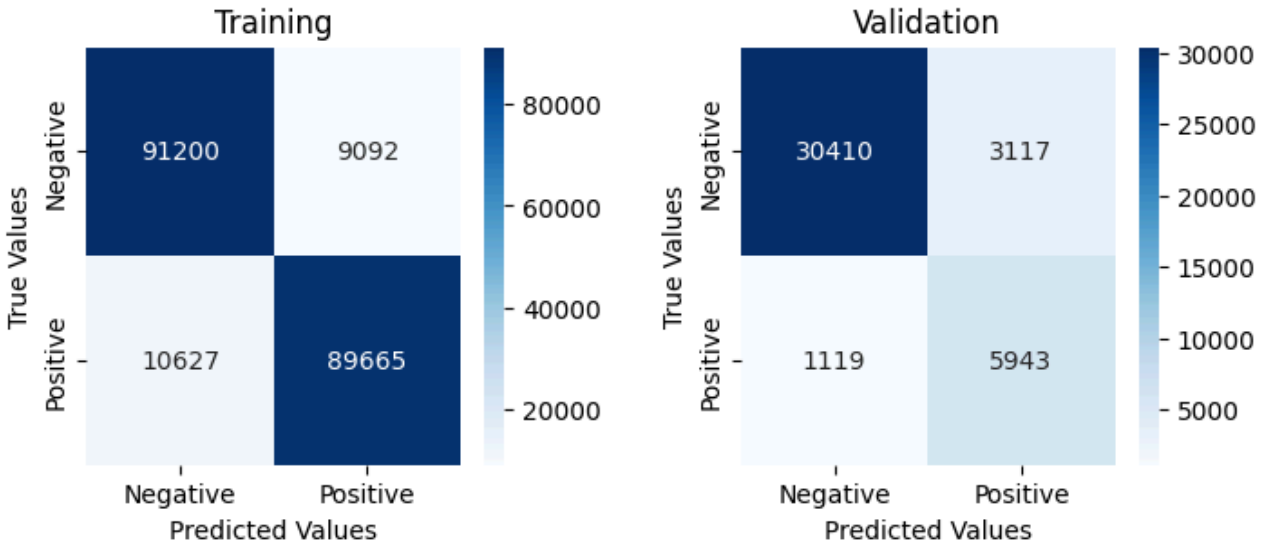

Fig. 8 Random Forest

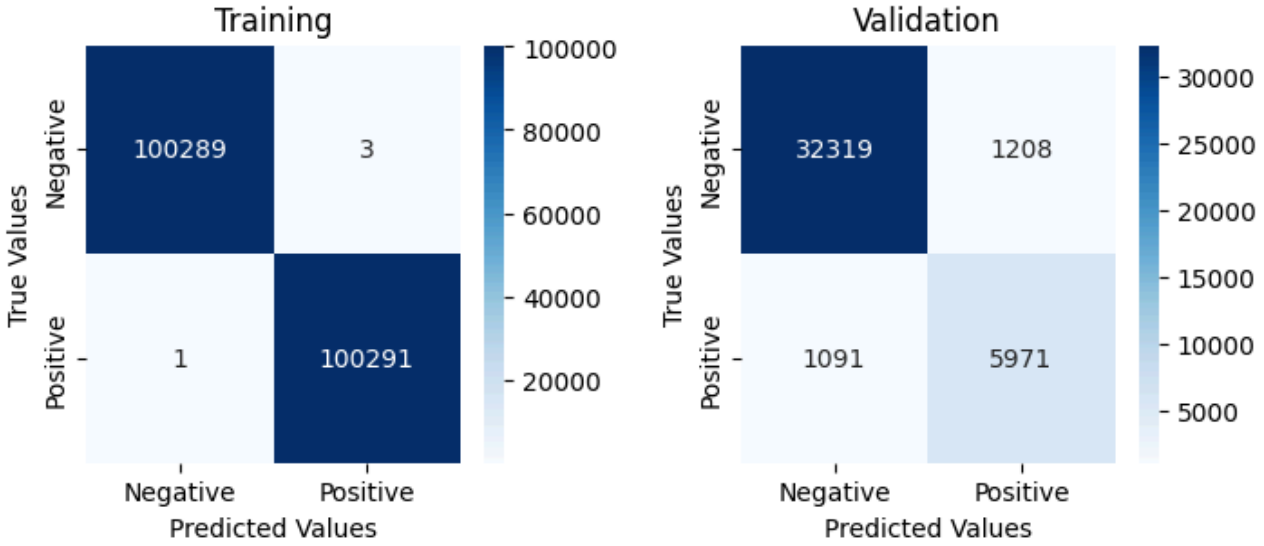

Fig. 9 Random Forest (GridSearchCV tuning)