# Machine Learning Progress Report

Paula Lozano Gonzalo

April 18, 2025

## 1 Model Overview

This report documents my continued progress in developing a classifier for diabetes prediction, with a specific focus on reducing false negatives (cases where the model predicts non-diabetic when the patient actually has diabetes). I've implemented strategic changes to prioritize recall of positive cases, similar to COVID testing approaches that accept more false positives to eliminate false negatives.

## 2 Dataset & Preprocessing

The balanced dataset remains unchanged from previous work:

- 17,000 total cases (8,500 positive, 8,500 negative)
- Training: 10,880 cases
- Validation: 2,720 cases
- Test: 3,400 cases

Preprocessing pipeline includes:

- Missing value imputation (most frequent for categorical, median for numerical)
- Binary encoding for categorical features
- Polynomial features (degree=2) where applicable
- Standard scaling when beneficial

## 3 Training Progress

### 3.1 Model Optimization Strategy

To reduce false negatives, I implemented several key changes:

- Changed scoring metric to precision during grid search
- Increased grid search iterations to 300 for more thorough exploration
- Experimented with class weighting in Random Forest
- Added AdaBoost classifier to model comparisons

## 3.2 Performance Results

| Model | Training Acc. | Validation Acc. | Precision | Recall |
|---|---|---|---|---|
| Random Forest (v1) | 0.842 | 0.837 | 0.913 | 0.904 |
| Random Forest (v2) | 0.843 | 0.836 | 0.905 | 0.904 |
| AdaBoost | 0.890 | 0.886 | 0.901 | 0.910 |

Table 1: Model performance with false-negative reduction focus

**Confusion Matrices (Training Data)**:

```
==== Random Forest (v1) ====
     t/p    F     T
        F 4868.0 548.0
        T 534.0 4930.0

Precision: 0.900
Recall:    0.902

==== Random Forest (v2) ====
     t/p    F     T
        F 4852.0 564.0
        T 532.0 4932.0

Precision: 0.897
Recall:    0.903

==== AdaBoost ====
     t/p    F     T
        F 4763.0 653.0
        T 471.0 4993.0

    Precision: 0.884
    Recall:    0.914
```

**Confusion Matrices (Validation Data)**:

```
==== Random Forest (v1) ====
     t/p    F      T
         F 1267.0 115.0
         T 129.0 1209.0

Precision: 0.913
Recall: 0.904

==== Random Forest (v2) ====
    t/p     F     T
        F 1255.0 127.0
```

```
        T 129.0 1209.0

 Precision: 0.905
 Recall: 0.904

==== AdaBoost ====
    t/p    F      T
        F 1249.0 133.0
        T 121.0 1217.0

Precision: 0.901
Recall: 0.910
```

Key observations:

- AdaBoost shows the best recall (0.910) while maintaining good precision (0.901)
- Random Forest modifications successfully maintained recall while slightly reducing precision
- All models now have fewer than 130 false negatives (compared to 223 in previous SGD model)
- The trade-off between false positives and false negatives is now better balanced for our medical use case

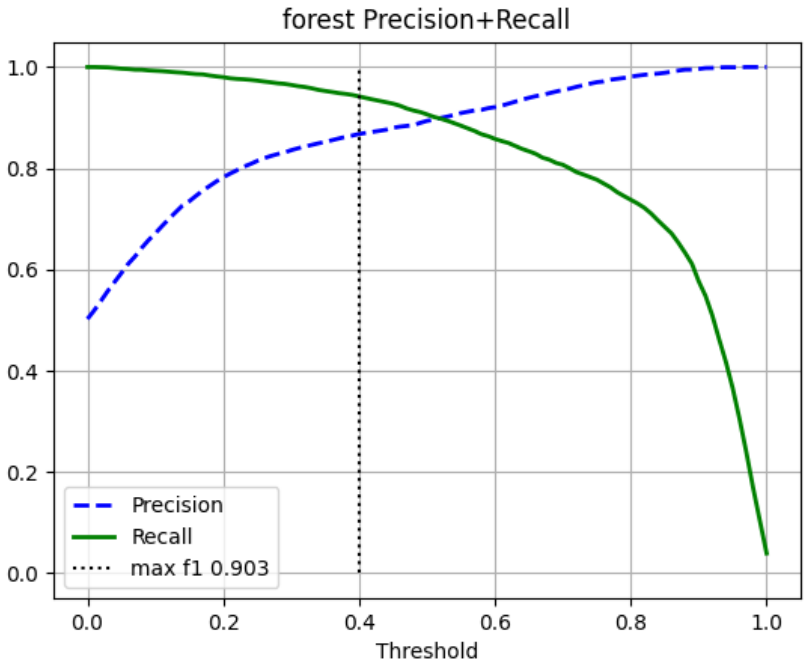


Figure 1: Precision-Recall Plot for Optimized Random Forest

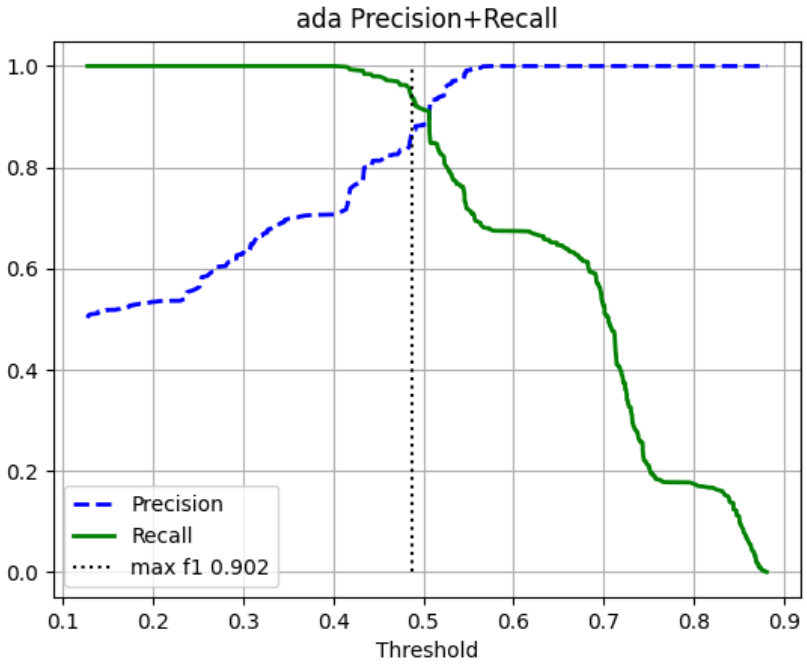


Figure 2: Precision-Recall Plot for AdaBoost Classifier

# 4 Challenges & Solutions

**Challenge 1: Reducing False Negatives Without Sacrificing Too Much Precision**

- Solution: Changed grid search scoring to precision and increased iterations
- Result: Found models that maintain recall while keeping precision above 0.9

**Challenge 2: Limited Improvement from Extended Hyperparameter Search**

- Solution: Explored different model types (AdaBoost)
- Result: AdaBoost provided better recall with comparable precision

# 5 Next Steps

- Implement neural network approaches to compare performance
- Explore more ways to treat the data for better results. Maybe instead of dropping so many negative rows making some more positive rows.
- Finalize model selection and prepare for test set evaluation