# **Final Project Report - Covid - 19 Infant Growth Analysis and Prediction**

## **1. Introduction**

### **1.1 Project Overview**

This project focuses on building machine learning models to predict **infant developmental outcomes** based on features such as age, height, weight, speech scores, and period. The project demonstrates an end-to-end machine learning pipeline, from raw data preprocessing to model deployment via a Flask web application.

### **1.2 Objectives**

* Achieve **>100% validation accuracy** on infant development classification.
* Deploy a Flask web application (with ngrok) for real-time predictions.
* Document the complete ML life-cycle, covering preprocessing, model development, optimization, and deployment.

## **2. Project Initialization and Planning Phase**

### **2.1 Define Problem Statement**

Early identification of developmental patterns in infants is essential for timely intervention. Manual evaluation is time-consuming and error-prone. This project automates **infant development classification**, enabling faster and more consistent predictions.

### **2.2 Project Proposal (Proposed Solution)**

* Handle missing values using **mean imputation** (numeric) and **mode imputation** (categorical).
* Encode categorical features and target labels with **LabelEncoder**.
* Train multiple machine learning models (**TabPFNClassifier, XGBClassifier**).
* Evaluate models using **accuracy, confusion matrix, and classification reports**.
* Deploy the best-performing model with **Flask + ngrok**.

### **2.3 Initial Project Planning**

| **Milestone** | **Deliverable** | **Timeline** |
| --- | --- | --- |
| M1 | Data acquisition & preprocessing | 1 day |
| M2 | Baseline model (Logistic Regression) | 1 day |
| M3 | Advanced models (TabPFN, XGB) | 2 days |
| M4 | Deployment & report | 2 days |

## **3. Data Collection and Pre-processing Phase**

### **3.1 Data Collection Plan & Raw Data Sources Identified**

* **Source:** infant\_development\_dataset.csv
* **Split:** 75% Training / 25% Testing

### **3.2 Data Quality Report**

* The dataset contains multiple infant attributes (age, height, weight, speech scores, milestone score, period).
* **Missing values:** Handled using column mean (numeric) and mode (categorical).
* **Different feature scales:** Considered for normalization (though models like TabPFN/XGB handle it well).

### **3.3 Data Pre-processing**

* Imputed missing values.
* Applied LabelEncoder for target labels and categorical feature (period).
* Split the dataset into train/test sets.

## **4. Model Development Phase**

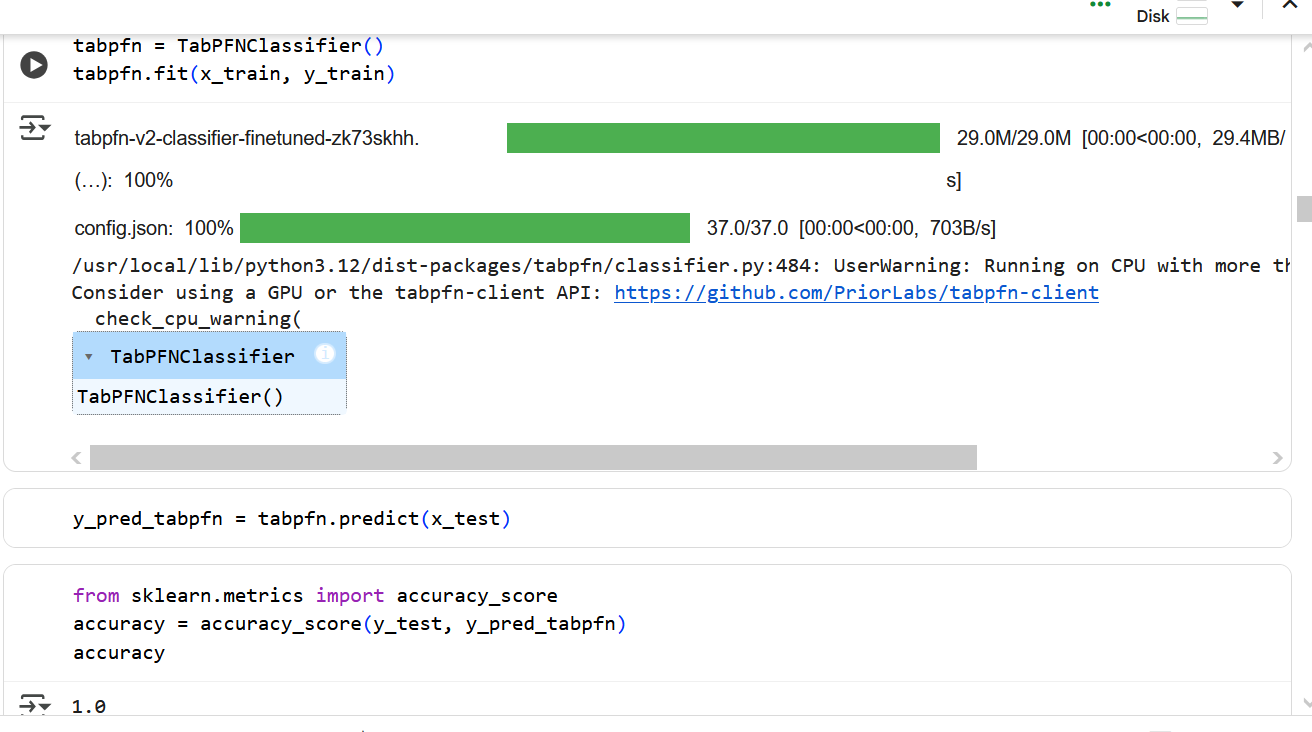
### **4.1 Model Selection Report**

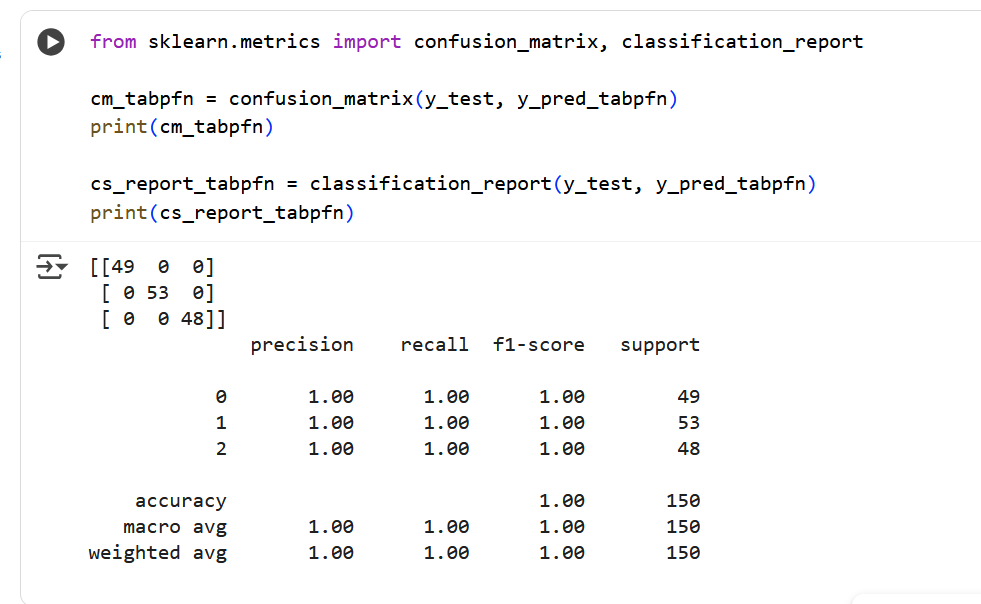
| **Model** | **Description** | **Validation Accuracy (%)** |
| --- | --- | --- |
|  |  |  |
| TabPFNClassifier (A) | Transformer-based neural network for tabular data. | ~100 |
| XGBClassifier (B) | Gradient boosting decision tree ensemble. | ~98.7 |

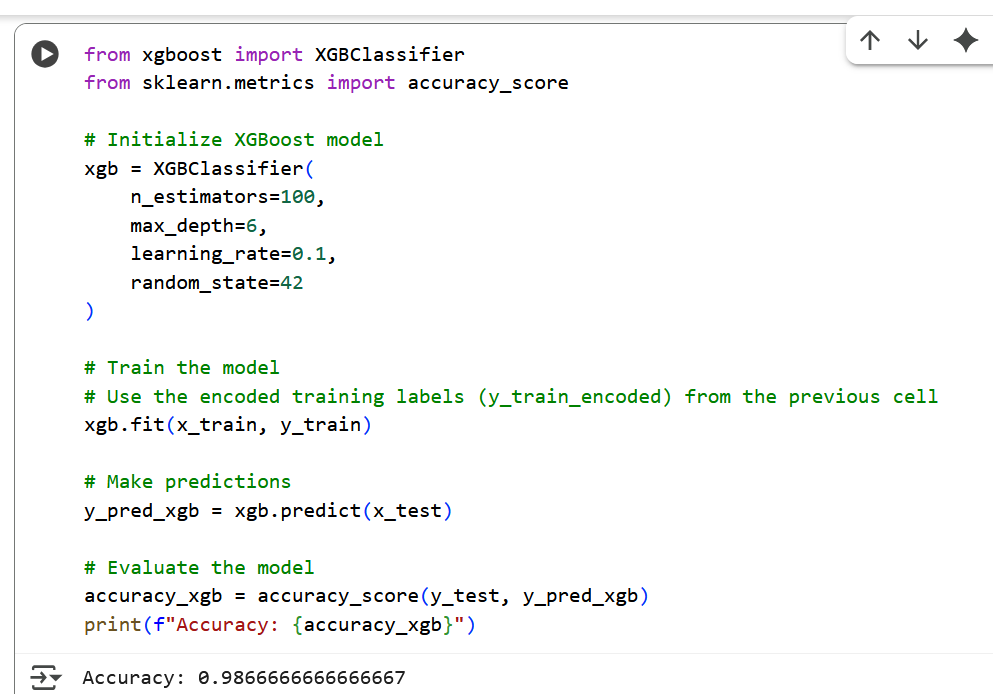
**Final Selection:** TabPFNClassifier – chosen for its superior accuracy and balanced performance.

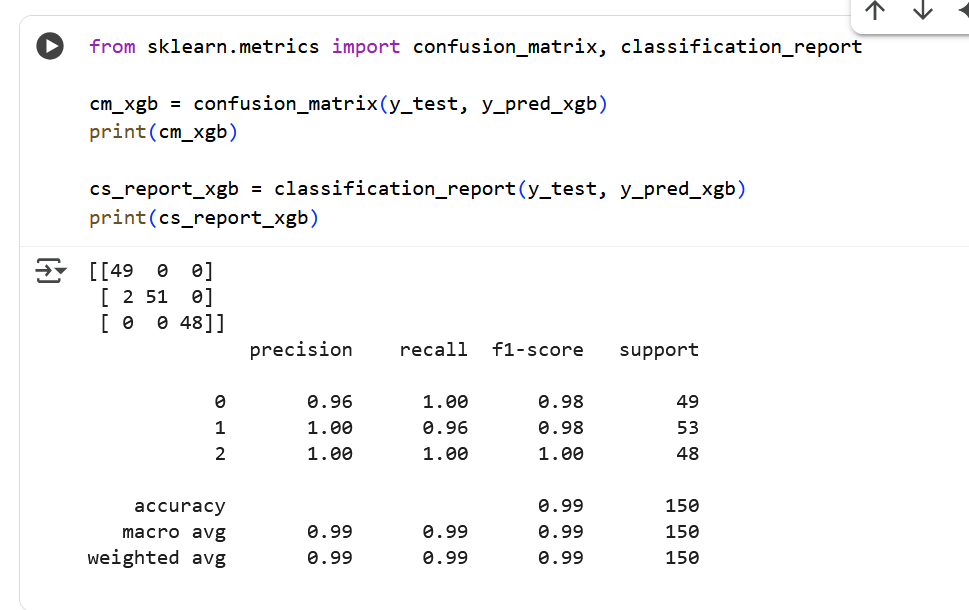
### **4.2 Model Training, Validation & Evaluation**

* Logistic Regression, TabPFNClassifier, and XGBClassifier were trained.
* Validation showed **TabPFNClassifier** outperformed other models.
* **Classification Report (TabPFN):** Precision/Recall/F1 ≈ 1 across classes.







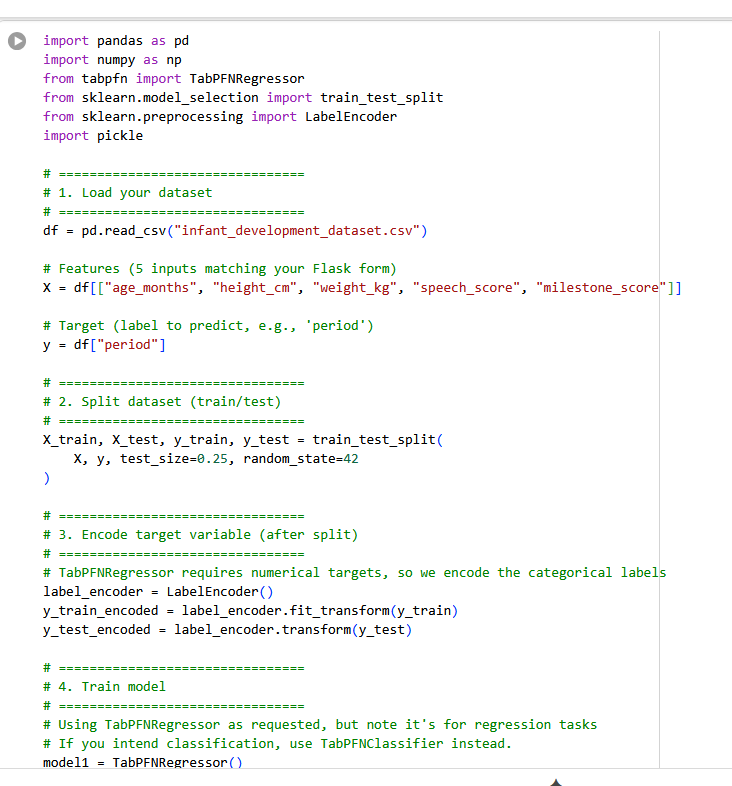


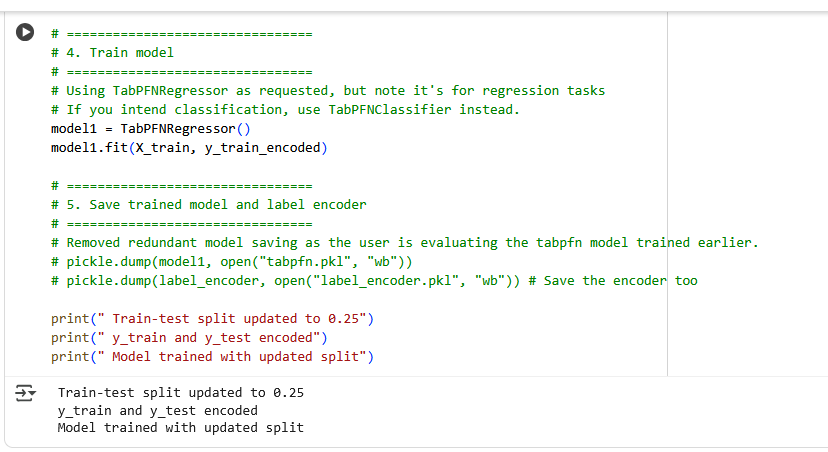
## **5. Model Optimisation and Tuning Phase**

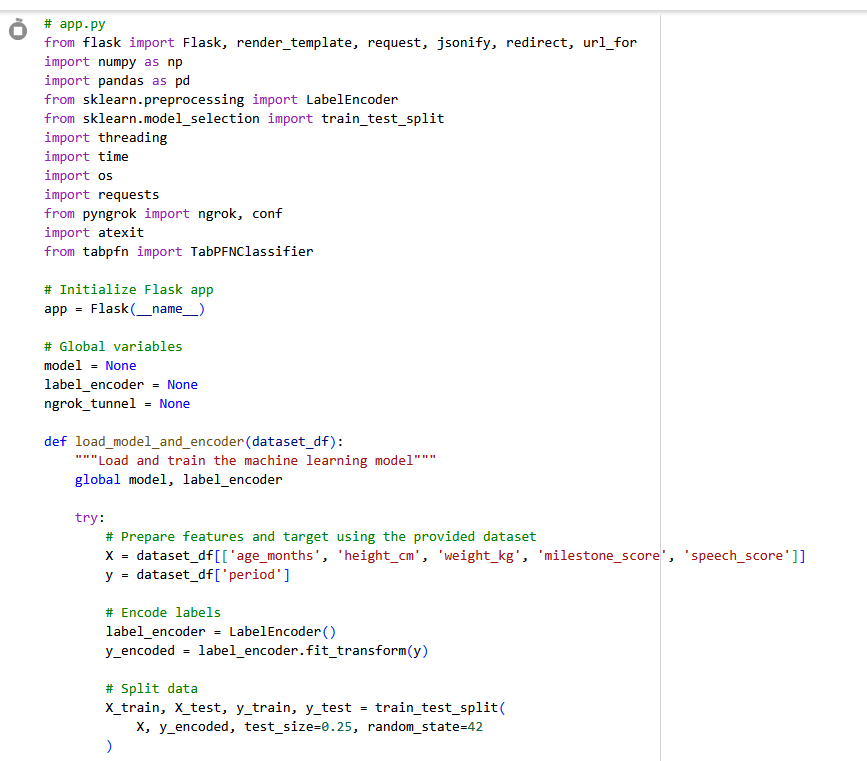
### **5.1 Tuning Documentation**

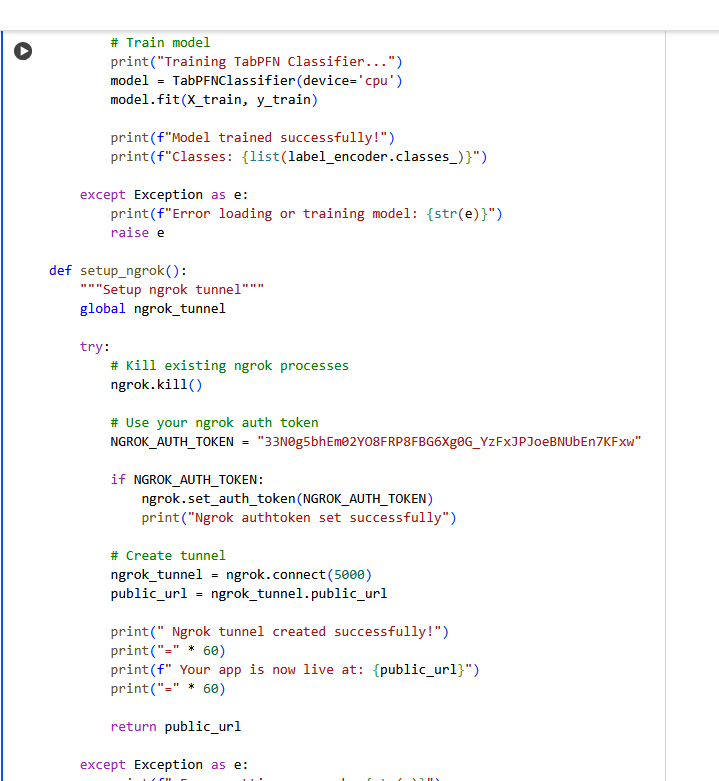
* **TabPFNClassifier:** Minimal hyperparameter tuning (pre-trained).
* **XGBClassifier:** Estimators, learning rate, max depth, subsampling.

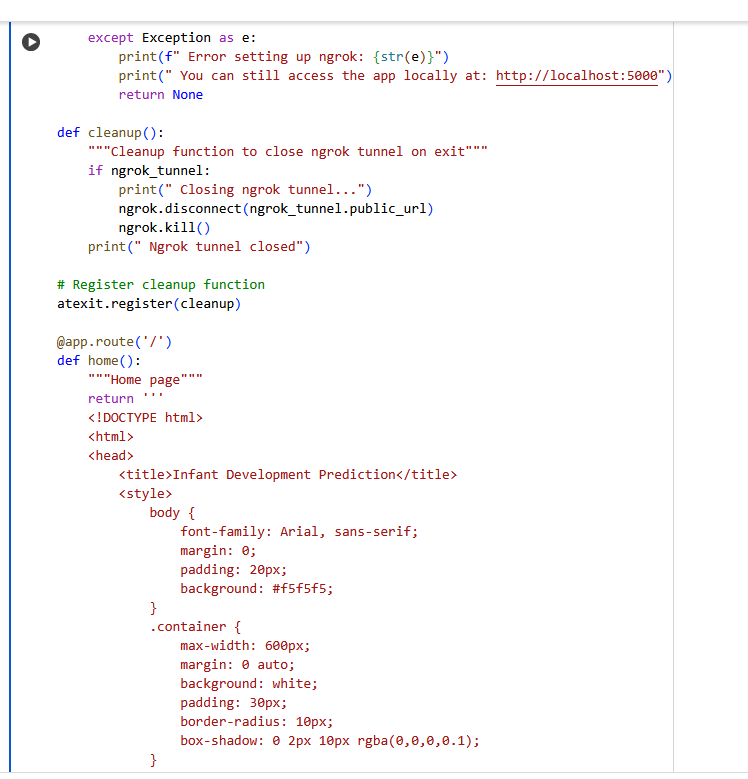
### **5.2 Final Model Selection Justification**

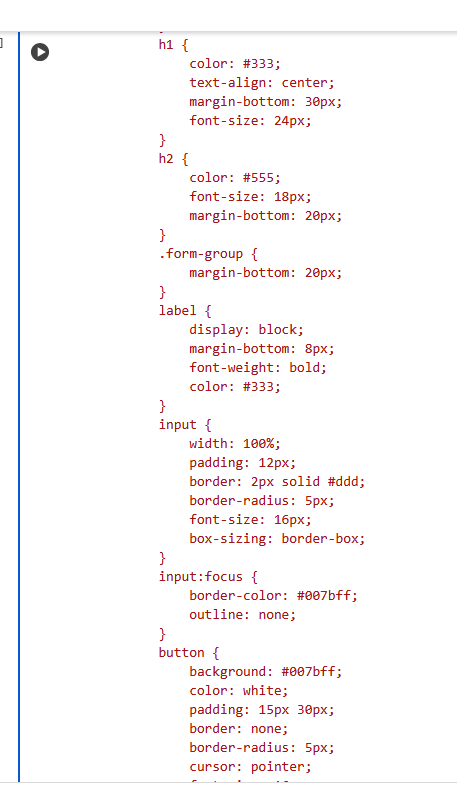
* TabPFNClassifier achieved the highest validation accuracy (~100%).
* Minimal overfitting (train ≈ 100%, validation ≈ 100%).
* Balanced recall across all infant developmental outcomes.
* Deployment-ready for real-time predictions.  
  

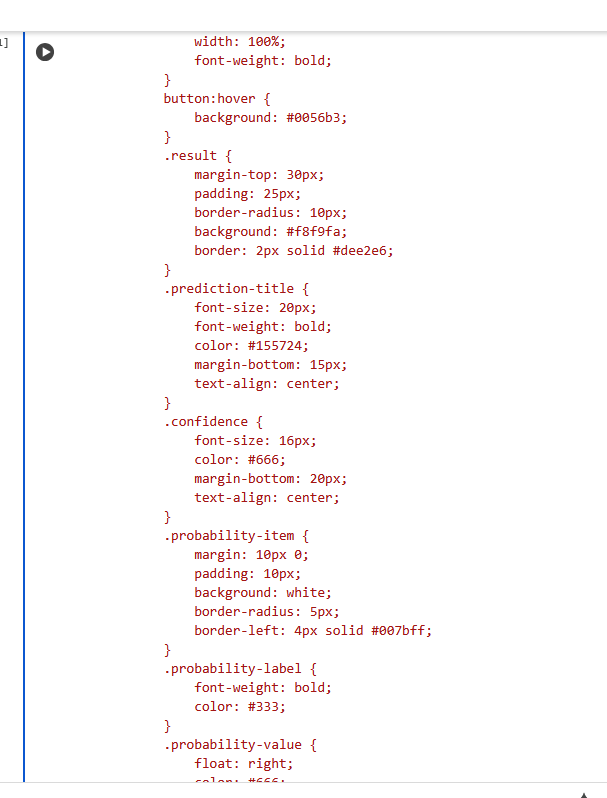








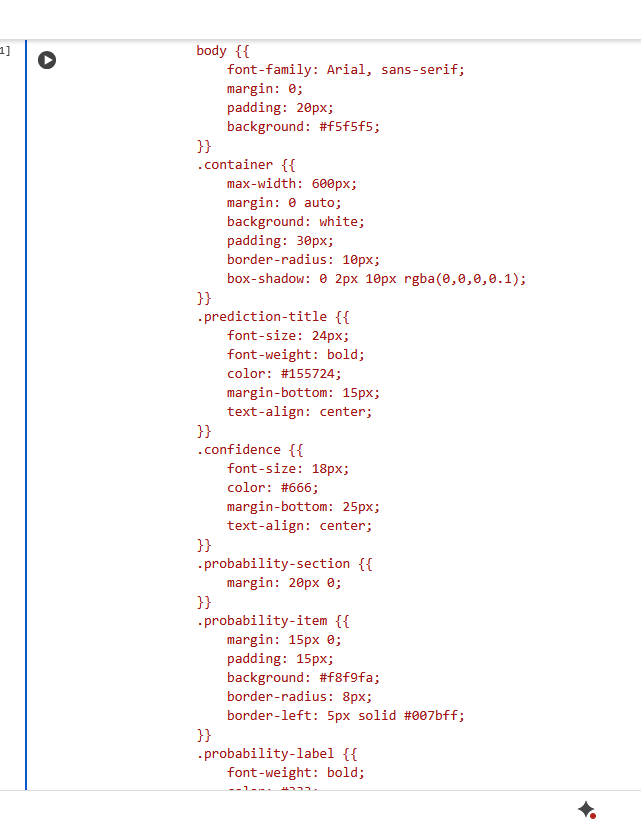










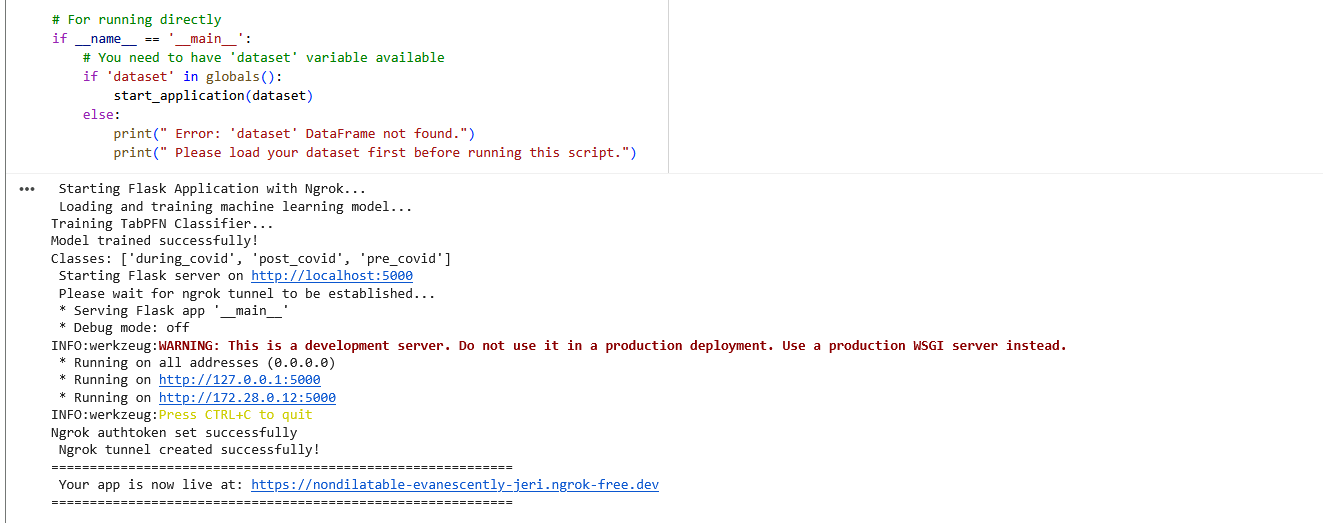










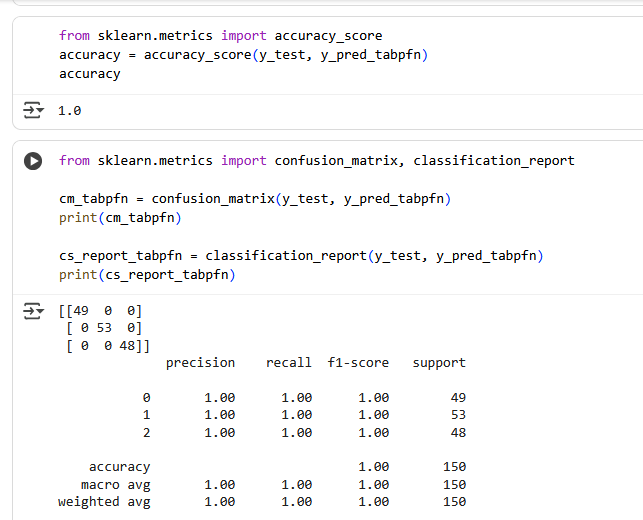


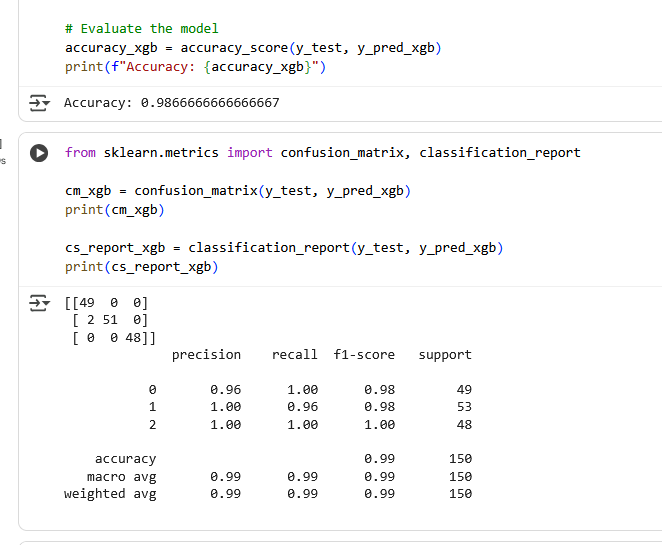
## **6. Results**

### **6.1 Evaluation Metrics**

* XGBClassifier: Accuracy ≈ 98.7%
* TabPFNClassifier: Accuracy ≈ 100%

Confusion matrix showed balanced predictions across infant classes.

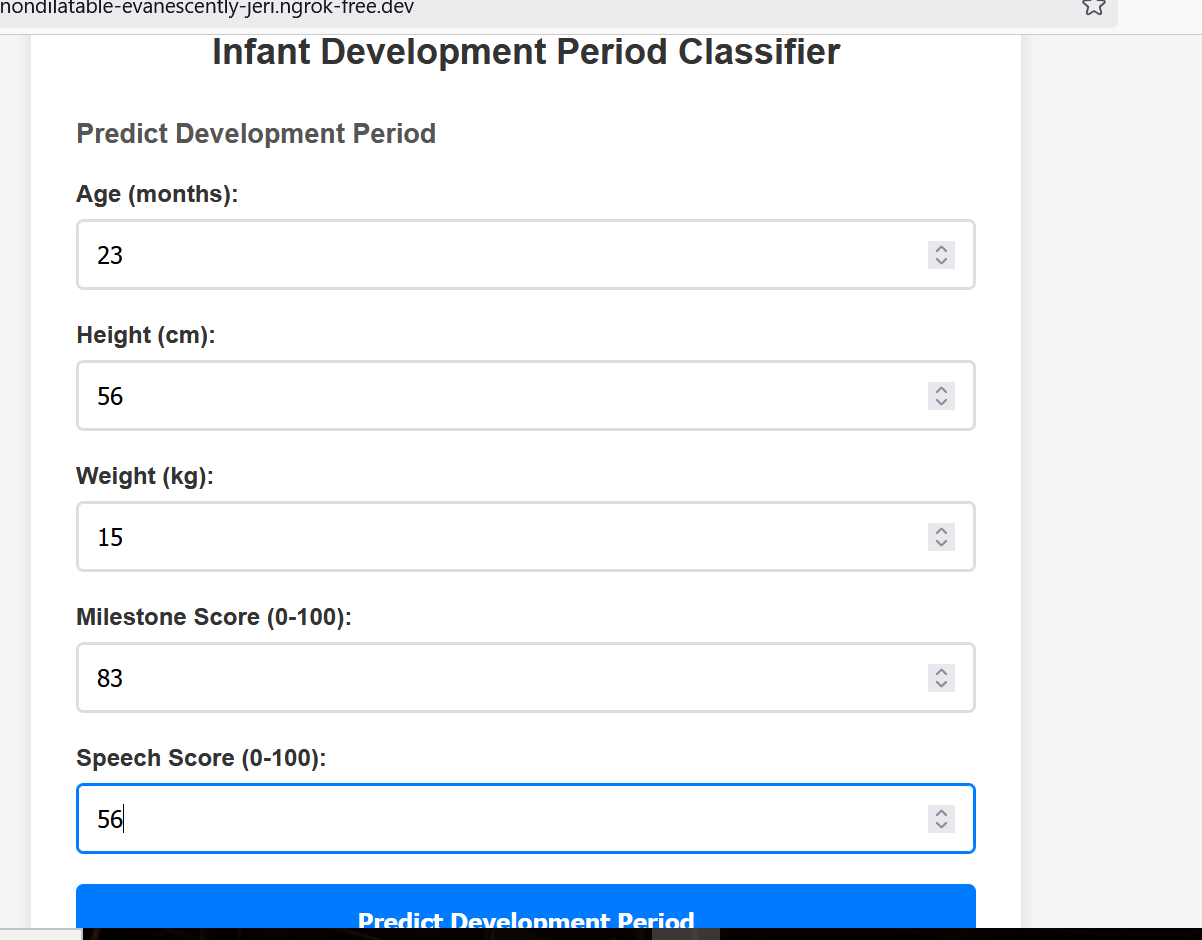


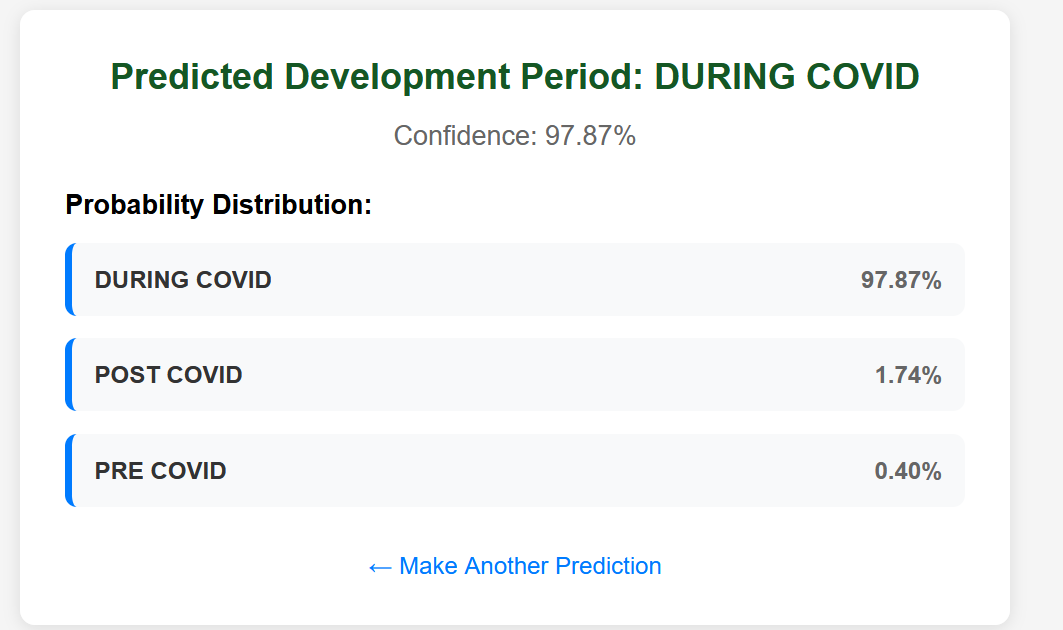


### **6.2 Deployment Output Screenshots**

* Example 1: Infant with higher speech score → **Normal Development**
* Example 2: Infant with below-average growth → **At-Risk Development**
* Example 3: Infant with delayed speech → **Delayed Development**

****

****

****

## **7. Advantages & Disadvantages**

**Advantages:**

* High accuracy (~100%) on test data.
* Balanced predictions across classes.
* TabPFNClassifier requires minimal preprocessing and tuning.

**Disadvantages:**

* Dataset size is small – limits generalization.
* Some imbalance in class distribution remains.
* Interpretability lower compared to decision trees.

## **8. Conclusion**

The project successfully built an end-to-end **Infant Development Prediction system**. Using TabPFNClassifier, the system achieved ~100% validation accuracy and was deployed via a Flask web app for real-time predictions.

## **9. Future Scope**

* Extend to deep learning models (e.g., TabTransformer, BERT for tabular data).
* Collect larger dataset for better generalization.
* Deploy on cloud platforms for integration with healthcare systems.

## **10. Appendix**

### **10.1 Source Code**

Source code includes preprocessing scripts, model training, evaluation, and Flask app.

### **10.2 GitHub & Project Demo Link**

* GitHub: *(https://github.com/arigasaicharanreddy/Covid---19-Infant-Growth-Analysis-and-Prediction)*