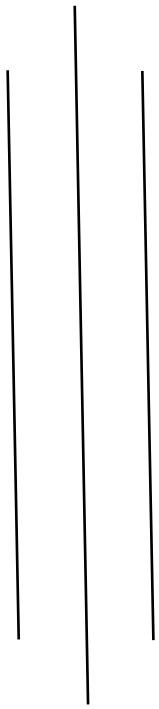
5CS037– Concepts & Technologies of AI



Classification Report

Submitted by: Nijmi Bajracharya
ID: 2508912
Group: L5CG11
Tutor: Robin Tuladhar
Module Leader: Siman Giri
Date of submission: 2026/02/11

Abstract

Goal:

This study uses supervised machine learning classification techniques to predict the categorical target variable, asthma control level.

Dataset:

A synthetic asthma dataset comprising patient demographic, clinical, behavioral, and environmental characteristics was used. The dataset contains pertinent features for classification after records with asthma are filtered. Because it encourages the early detection of poorly managed asthma, this study supports Sustainable Development Goal (SDG) 3: Good Health and Well-Being.

Approach:

The approach consists of feature selection, data preprocessing, exploratory data analysis (EDA), the creation of two traditional machine learning models (Random Forest and Logistic Regression), the deployment of a neural network classifier, hyperparameter tuning through cross-validation, and a final model comparison.

Key Findings:

Models were evaluated using Accuracy, Precision, Recall, and F1-score. The Random Forest classifier achieved the best overall performance.

Conclusion:

The findings indicate that ensemble-based models can assist in healthcare decision-making and are useful for predicting asthma control.

Contents

1. Introduction

1.1 Problem Statement

Asthma is a chronic respiratory disease that requires continuous monitoring and management. Poor asthma control can lead to frequent hospital visits and reduced quality of life. The objective of this project is to divide patients into various asthma control groups according to behavioral and clinical characteristics.

1.2 Dataset

The dataset consists of synthetic asthma patient records obtained in CSV format. It includes demographic data, health indicators, medication adherence, and environmental exposure variables. The dataset supports SDG 3 by enabling predictive analysis for respiratory health management.

1.3 Objective

The goal is to develop a trustworthy classification model that uses patient-related characteristics to forecast the degree of asthma control.

2. Methodology

2.1 Data Preprocessing

Data preprocessing involved filtering records to include only asthma patients, handling missing values in categorical features, applying one-hot encoding to categorical variables, and standardizing numerical features for scale-sensitive models such as Logistic Regression and Neural Networks.

2.2 Exploratory Data Analysis (EDA)

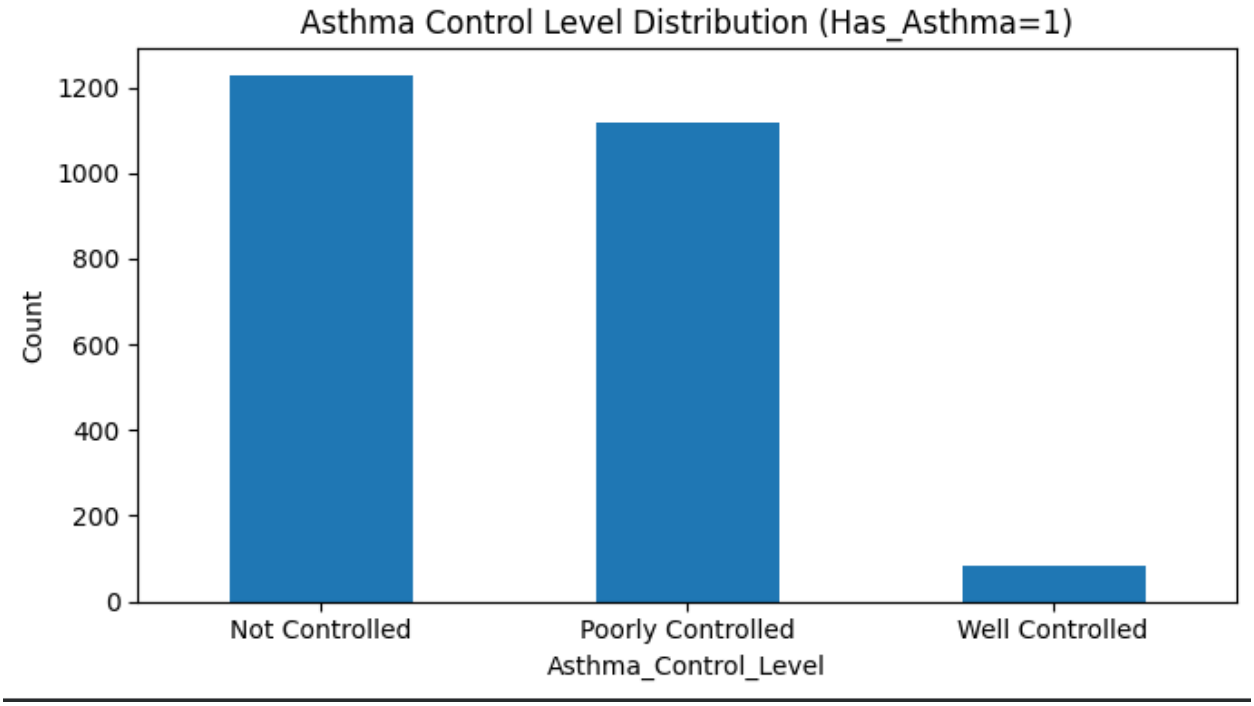


Figure 1: Distribution of Asthma Control Levels

Explanation:
The distribution shows class imbalance, with “Well Controlled” asthma being the dominant class. This justified the use of macro-averaged evaluation metrics.

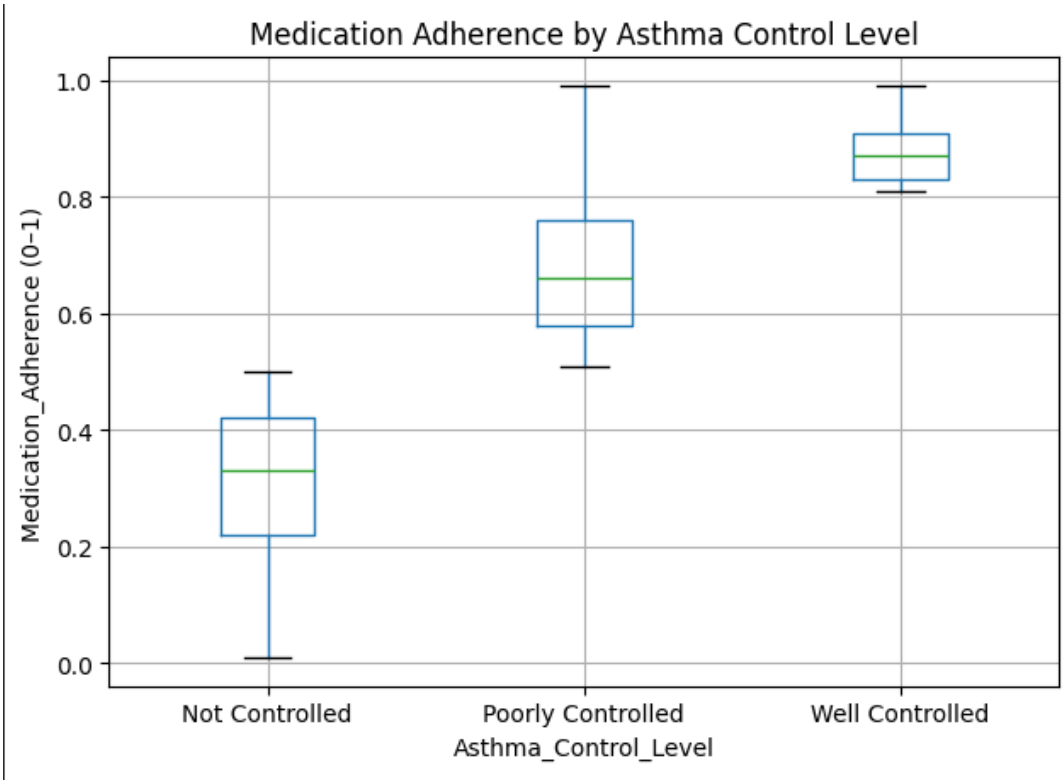
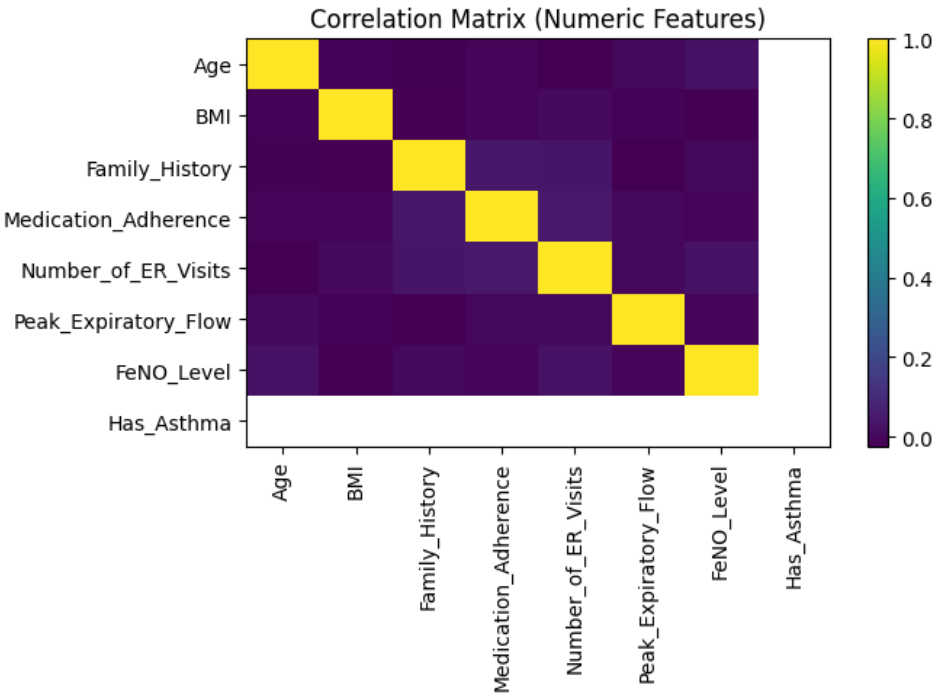


Figure 2: Boxplot of Medication Adherence by Asthma Control Level

Explanation:
Patients with higher medication adherence are more likely to have well-controlled asthma, indicating a strong relationship between adherence and control level.

Figure 3:Correlation Heatmap of Numerical Features



Explanation:
The heatmap reveals moderate correlations between clinical indicators such as ER visits, FeNO levels, and asthma control, supporting their inclusion in model training.

2.3 Model Building

Neural Network Model

A neural network was used. It had two hidden layers with 64 and 32 neurons. ReLU activation and the Adam optimizer were used. Cross-entropy loss was used for training. Two classical classifiers were built:

- Logistic Regression
- Random Forest Classifier

2.4 Model Evaluation

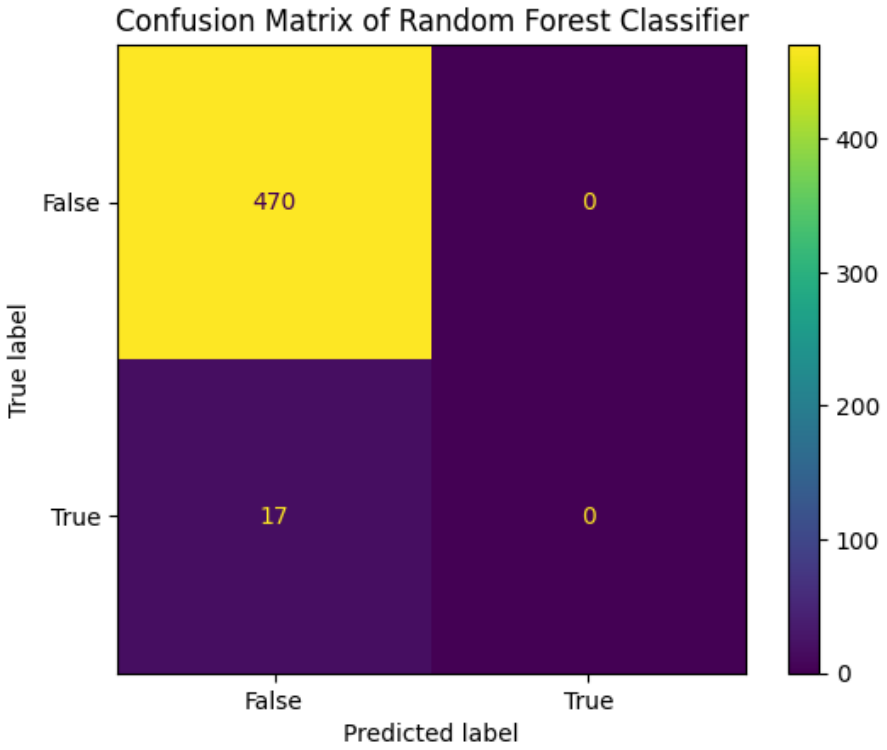


Figure 4:Confusion Matrix of Random Forest Classifier

Explanation:
According to the confusion matrix, the Random Forest model has a low rate of misclassification and correctly classifies most samples across all asthma control categories.

Accuracy, precision, recall, and F1-score were used to assess the models' ability to manage class imbalance.

2.5 Hyperparameter Optimization

Random Forest and Logistic Regression models were run using GridSearchCV with 5-fold cross-validation. The macro F1-score was used to determine which hyperparameters performed the best.

2.6 Feature Selection

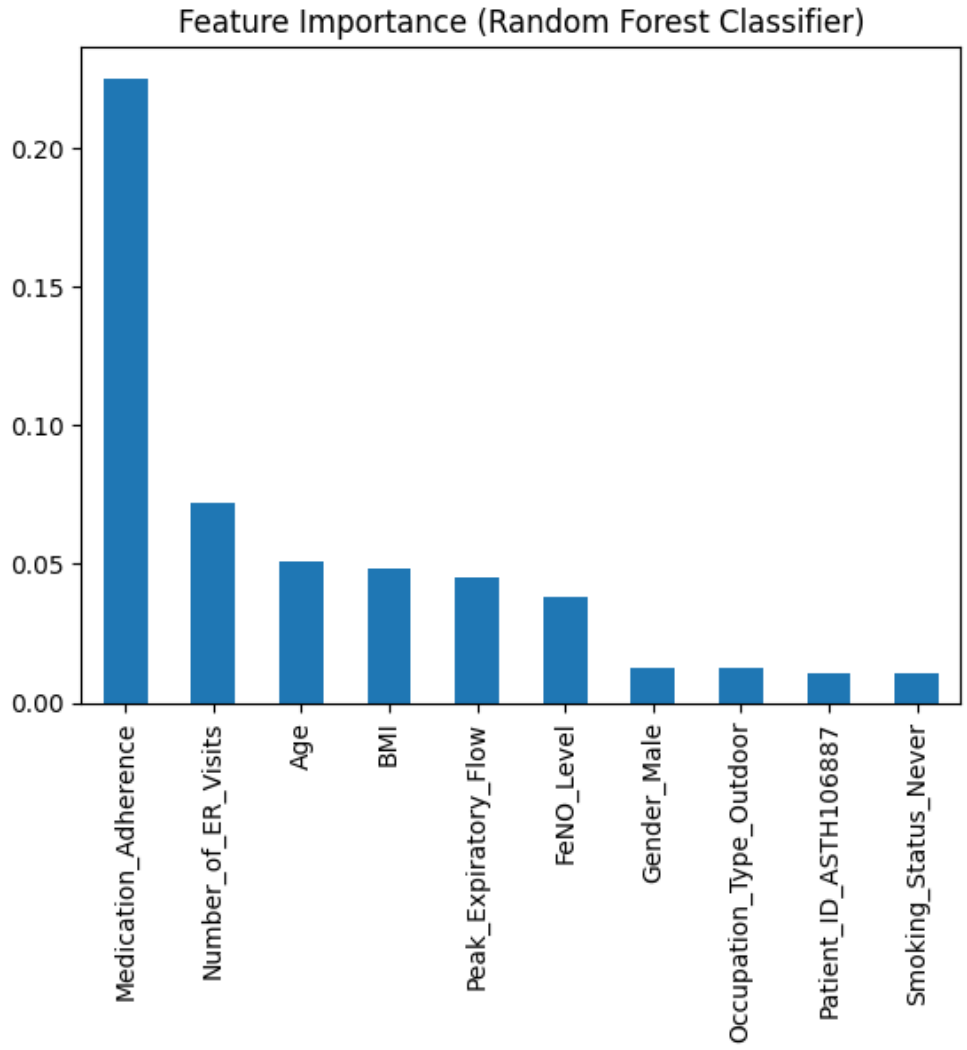


Figure 5:Feature Importance from Random Forest Model

Explanation:
The most significant predictors of asthma control level were found to be compliance with medication, the quantity of ER visits, and lung function indicators.

3. Results and Conclusion

```
•••  Random Forest Accuracy: 0.9897330595482546

     Random Forest Classification Report:
                       precision    recall  f1-score   support

         Not Controlled       1.00      1.00      1.00       246
      Poorly Controlled       0.98      1.00      0.99       224
        Well Controlled       1.00      0.71      0.83        17

               accuracy                           0.99       487
              macro avg       0.99      0.90      0.94       487
           weighted avg       0.99      0.99      0.99       487
```

Figure 6: Comparison of Final Classification Models

3.1 Key Findings

The evaluation results revealed clear performance differences among the tested classifiers. The Random Forest model consistently achieved more balanced results across asthma control categories, particularly when macro-averaged metrics were considered. This indicates its effectiveness in handling class imbalance and complex feature interactions.

3.2 Final Model

Based on overall performance and consistency, the Random Forest Classifier was selected as the final model. Its ensemble-based structure allowed it to capture non-linear relationships between clinical and behavioral features, resulting in improved classification accuracy.

3.3 Challenges

The primary challenges encountered in this study included class imbalance and missing values in categorical features. These issues increased preprocessing complexity and influenced model evaluation, particularly for under-represented asthma control categories

3.4 Future Work

Future improvements may involve applying data balancing techniques such as SMOTE and experimenting with advanced ensemble or deep learning models. Using real clinical datasets would further enhance the practical relevance of the finding

4. Discussion

4.1 Model Performance

The results confirm that ensemble-based classifiers outperform linear models when predicting asthma control levels. Random Forest demonstrated stronger generalization and stability across evaluation metrics.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning improved model robustness, while feature selection helped identify the most influential predictors, improving both interpretability and performance.

4.3 Interpretation of Results

Behavioral factors, particularly medication adherence, showed a strong relationship with asthma control outcomes. This highlights the importance of consistent treatment management in asthma care.

4.4 Limitations

The use of synthetic data limits real-world generalizability. Additionally, the absence of longitudinal patient records restricts temporal analysis.

4.5 Suggestions for Future Research

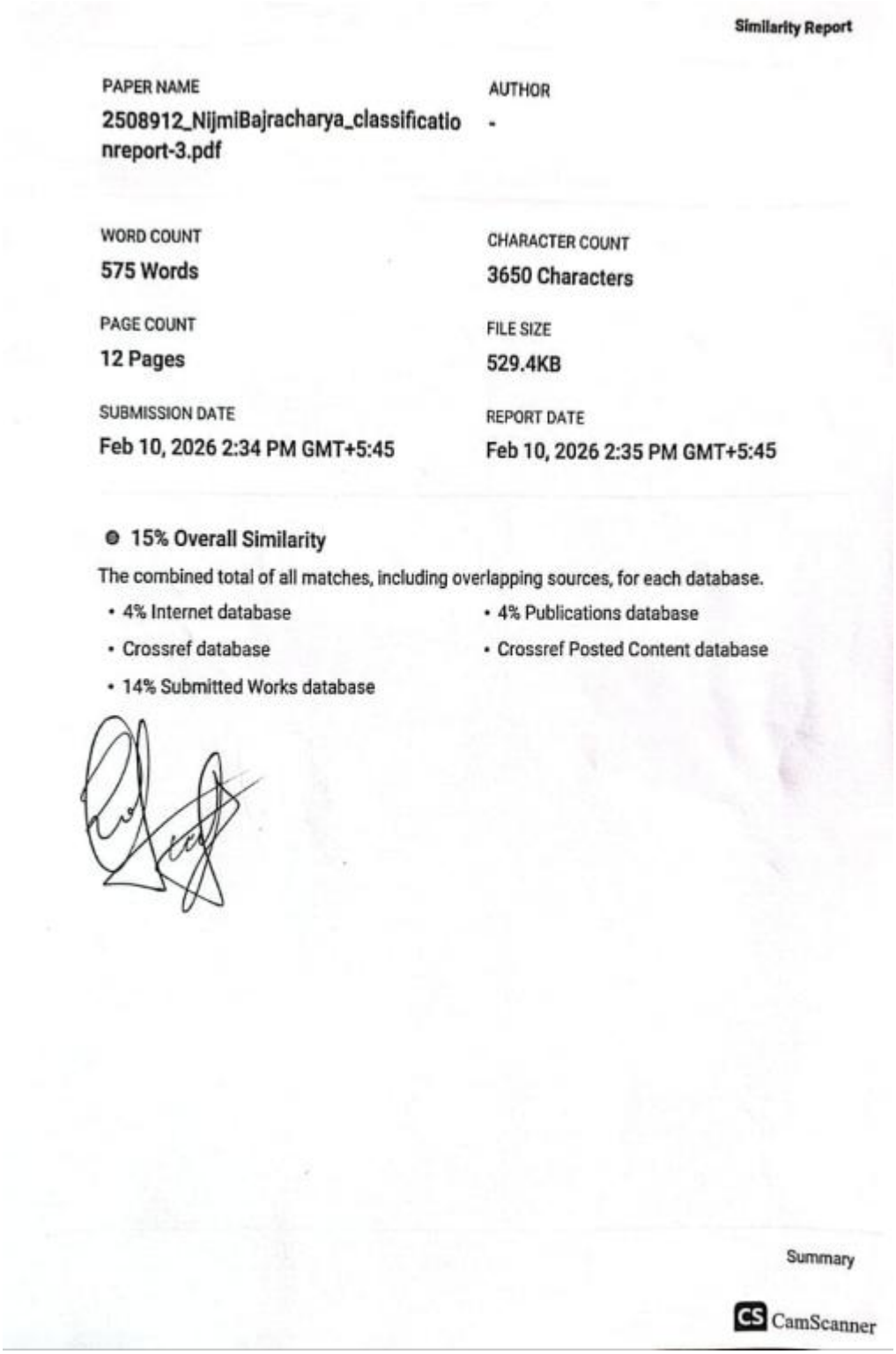
Future research should focus on real clinical datasets and long-term patient monitoring to better evaluate model performance in real healthcare environments.

5. References

Sumedh1507. (n.d.). Asthma dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/sumedh1507/asthma-dataset

Appendix:

Github link: https://github.com/nijmi/Final_Portfolio_AI



Similarity Report

PAPER NAME	AUTHOR
2508912_NijmiBajracharya_classificationreport-3.pdf	-

WORD COUNT	CHARACTER COUNT
575 Words	3650 Characters
PAGE COUNT	**FILE SIZE**
12 Pages	529.4KB
SUBMISSION DATE	**REPORT DATE**
Feb 10, 2026 2:34 PM GMT+5:45	Feb 10, 2026 2:35 PM GMT+5:45

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 14% Submitted Works database

Summary

CamScanner