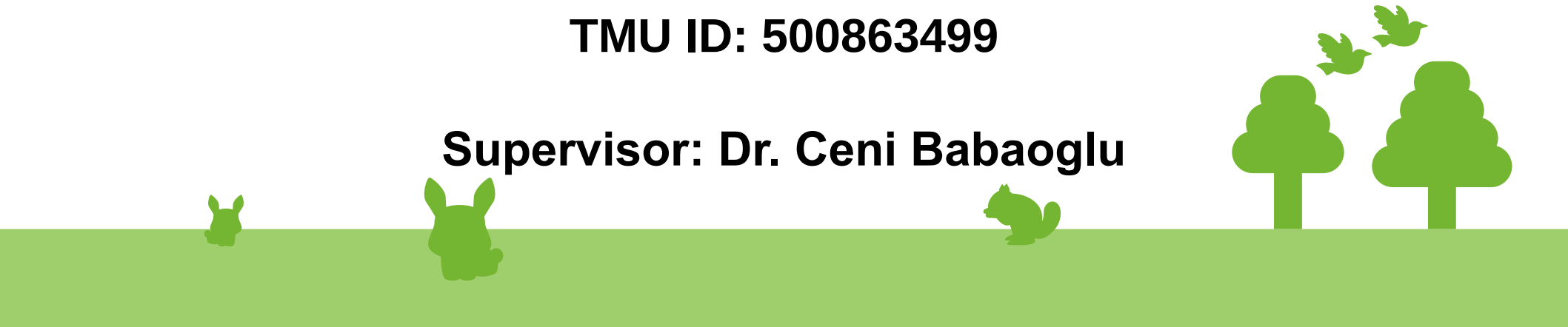
# Data Analytics Project

**Title: *Healthcare Analytics***: Predicting Stroke Risk using Machine Learning

A comprehensive data pipeline from raw data to interactive prediction.

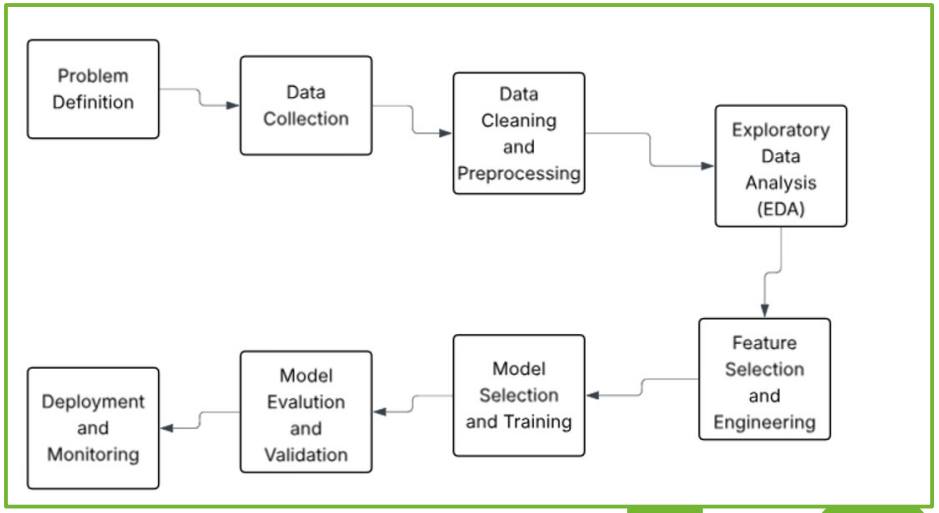
**Md. Islam**
**TMU ID: 500863499**

**Supervisor: Dr. Ceni Babaoglu**



# Agenda

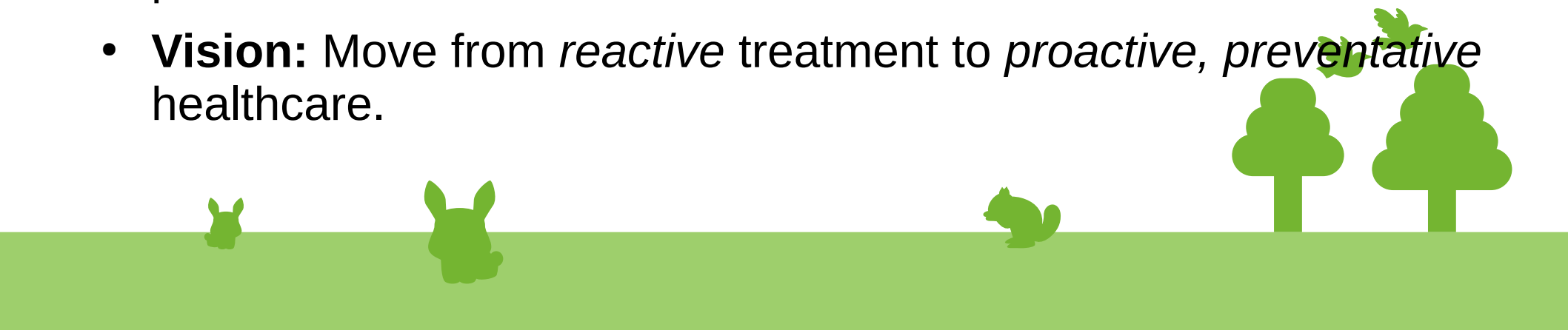
- *Problem Statement & Objective*
- *Dataset Overview (Health-Stroke)*
- *Data Cleaning & Outlier Handling (Data Preprocessing)*
- *Exploratory Data Analysis (EDA)*
- *Feature Preparing & Handling Imbalance (SMOTE)*
- *Model Selection Strategy (10-Fold Cross-Validation)*
- *Final Model Evaluation (Confusion Matrix)*
- *Key Risk Factors (Feature Importance)*
- *Practical Application: Interactive Tool*
- *Conclusion*



# Problem Statement & Objective

Challenge: Stroke is a leading cause of death and disability globally.

- **Key Insight:** Early identification of high-risk individuals enables preventative care.
- **Our Goal:** Build a smart, data-driven tool to estimate a person's stroke risk.
- **Vision:** Move from *reactive* treatment to *proactive, preventative* healthcare.



# Project Deliverables

**The Model:**

- "Which machine learning algorithm (With several models to choose from) works best for this specific prediction task?",

**The Drivers:**

- Beyond just making a prediction, What does it tell us about the drivers of stroke risk?

# Heart_Stroke - Dataset

Example Dataset:

| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 9046 | Male | 67.0 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked | 1 |
| **1** | 51676 | Female | 61.0 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | NaN | never smoked | 1 |
| **2** | 31112 | Male | 80.0 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked | 1 |
| **3** | 60182 | Female | 49.0 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 |
| **4** | 1665 | Female | 79.0 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24.0 | never smoked | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5105** | 18234 | Female | 80.0 | 1 | 0 | Yes | Private | Urban | 83.75 | NaN | never smoked | 0 |
| **5106** | 44873 | Female | 81.0 | 0 | 0 | Yes | Self-employed | Urban | 125.20 | 40.0 | never smoked | 0 |
| **5107** | 19723 | Female | 35.0 | 0 | 0 | Yes | Self-employed | Rural | 82.99 | 30.6 | never smoked | 0 |
| **5108** | 37544 | Male | 51.0 | 0 | 0 | Yes | Private | Rural | 166.29 | 25.6 | formerly smoked | 0 |
| **5109** | 44679 | Female | 44.0 | 0 | 0 | Yes | Govt_job | Urban | 85.28 | 26.2 | Unknown | 0 |

5110 rows × 12 columns



# Dataset Overview

***Source & Structure:***

- **Data Source:** Healthcare Stroke Dataset (CSV).
- **Initial Shape:** Approximately 5,100 records, 12 columns.
- **Target Variable:** stroke (Binary: 0 = No, 1 = Yes).

Key Features:

- **Demographic:** Gender, Age, Residence Type, Marital Status.
- **Medical:** Hypertension, Heart Disease, Average Glucose Level, BMI.
- **Lifestyle:** Work Type, Smoking Status.

```
Dataset Shape: (5110, 12)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
None
```

# Data Cleaning & Preprocessing Strategy

- **Handling Missing Values:**
  - Detected missing values in bmi.
  - Imputed using the median value to be robust against outliers.
- **Dropping Irrelevant Data:**
  - Removed the id column (provides no predictive power).
- **Handling Inconsistencies:**
  - Removed rows with 'Other' gender (too few samples).
  - Merged 'Unknown' smoking status into 'never smoked' for consistency.

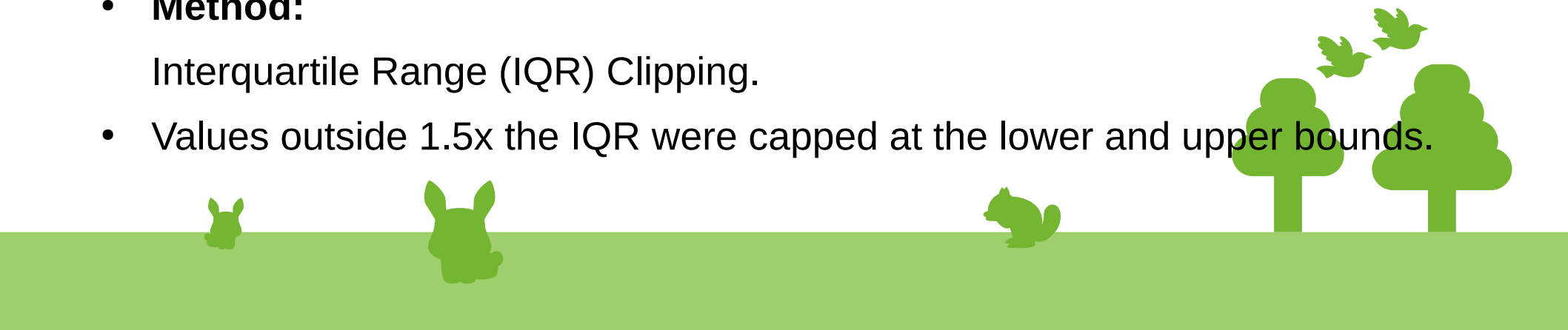
| | age | hypertension | heart_disease | avg_glucose_level | bmi | stroke |
|---|---|---|---|---|---|---|
| count | 5109.000000 | 5109.000000 | 5109.000000 | 5109.000000 | 5109.000000 | 5109.000000 |
| mean | 43.229986 | 0.097475 | 0.054022 | 106.140399 | 28.863300 | 0.048738 |
| std | 22.613575 | 0.296633 | 0.226084 | 45.285004 | 7.699785 | 0.215340 |
| min | 0.080000 | 0.000000 | 0.000000 | 55.120000 | 10.300000 | 0.000000 |
| 25% | 25.000000 | 0.000000 | 0.000000 | 77.240000 | 23.800000 | 0.000000 |
| 50% | 45.000000 | 0.000000 | 0.000000 | 91.880000 | 28.100000 | 0.000000 |
| 75% | 61.000000 | 0.000000 | 0.000000 | 114.090000 | 32.800000 | 0.000000 |
| max | 82.000000 | 1.000000 | 1.000000 | 271.740000 | 97.600000 | 1.000000 |

# Handling Outliers (Numerical Data)

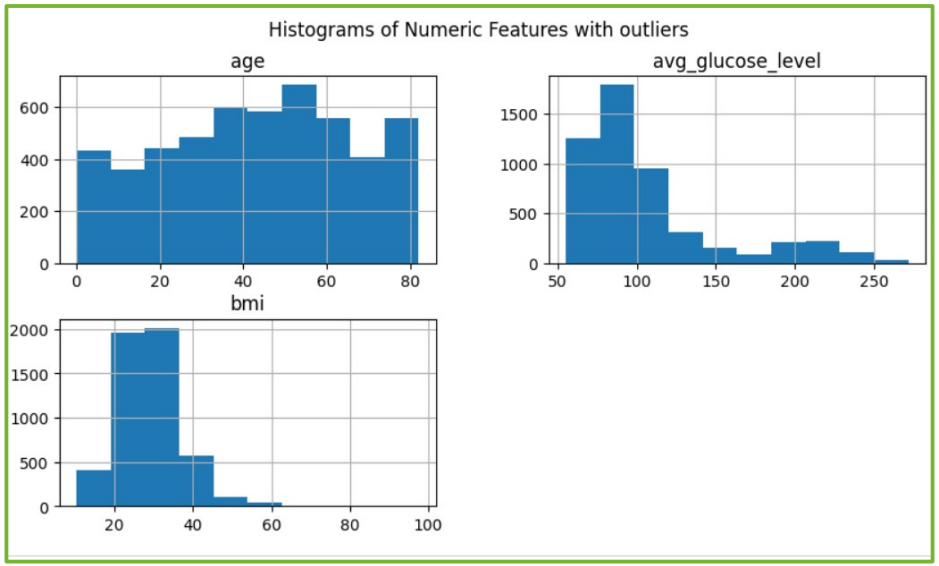
Features affected: age, avg_glucose_level, bmi.:

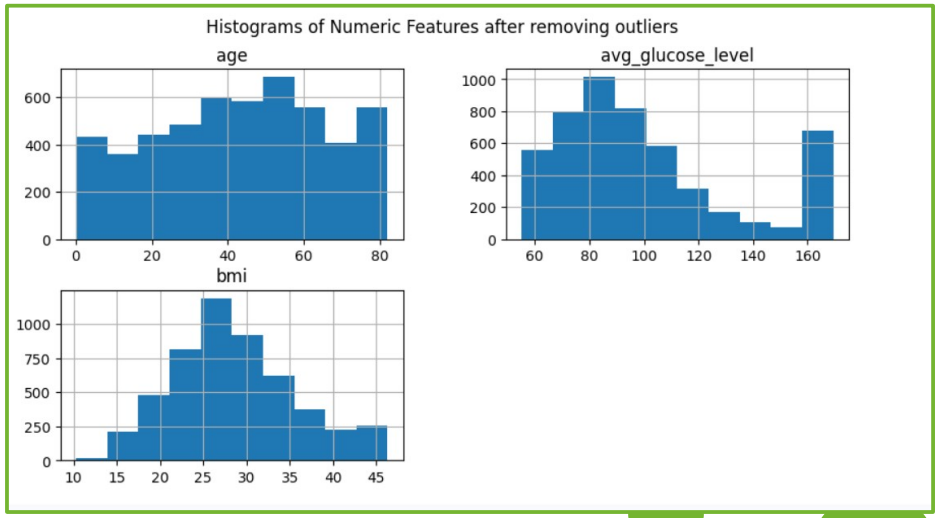
- **Distribution pattern:**
  - Age: The distribution is relatively uniform.
  - Avg_glucose_level: This feature is highly right-skewed
  - BMI: The BMI distribution is close to normal but slightly right-skewed, which aligns with the global trend towards overweight/obesity.
- **Method:**

  Interquartile Range (IQR) Clipping.
- Values outside 1.5x the IQR were capped at the lower and upper bounds.



# Exploratory Data Analysis (Numerical Data)

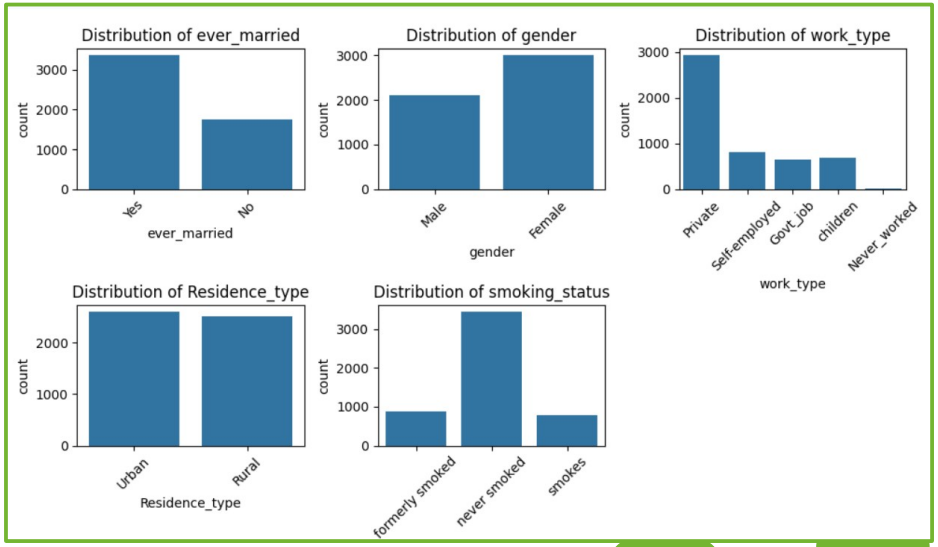


# Exploratory Data Analysis (Categorical Data)

**Key Observations:**

- **Gender & Marriage:** Balanced gender distribution; higher proportion of married individuals.
- **Work Type:** Majority employed in the private sector.
- **Residence:** Roughly equal split between Urban and Rural.
- **Smoking Status:** Large proportion of 'never smoked' (after merging unknowns).



# Feature Preparing

**Binary Encoding (0/1):**

- Converted yes/no features: ever_married, hypertension, heart_disease.

**One-Hot Encoding:**

- Applied to multi-category nominal features to avoid introducing false ordinal relationships.
- Features: gender, work_type, Residence_type, smoking_status.
- Note: Used drop_first=True to prevent multicollinearity (the dummy variable trap).

| bmi | stroke | gender_Male | work_type_Never_worked | work_type_Private | work_type_Self-employed | work_type_children | Residence_type_Urban | smoking_status_never smoked | smoking_status_smokes |
|---|---|---|---|---|---|---|---|---|---|
| 36.6 | 1 | True | False | True | False | False | True | False | False |
| 28.1 | 1 | False | False | False | True | False | False | True | False |
| 32.5 | 1 | True | False | True | False | False | False | True | False |
| 34.4 | 1 | False | False | True | False | False | True | False | True |
| 24.0 | 1 | False | False | False | True | False | False | True | False |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 28.1 | 0 | False | False | True | False | False | True | True | False |
| 40.0 | 0 | False | False | False | True | False | True | True | False |
| 30.6 | 0 | False | False | False | True | False | False | True | False |
| 25.6 | 0 | True | False | True | False | False | False | False | False |
| 26.2 | 0 | False | False | False | False | False | True | True | False |

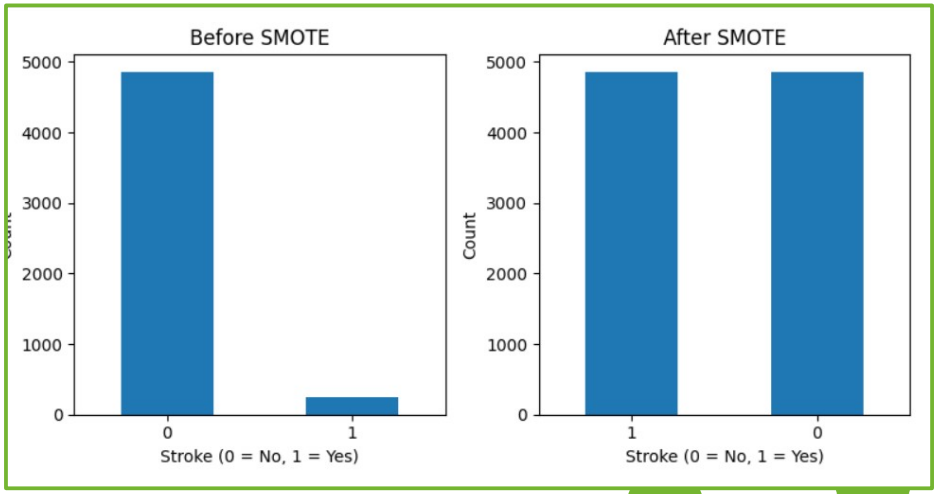
# The Critical Challenge: Class Imbalance

**The Problem:**

- Strokes are rare events.
- Initial Distribution: Approx. **95% "No Stroke" vs. 5% "Stroke".**
- A model could achieve 95% accuracy by simply predicting "No" every time—this is useless.

The Solution: **SMOTE** (Synthetic Minority Over-sampling Technique)

- Applied SMOTE to create synthetic examples of the minority class ("Stroke").
- Result: A perfectly balanced training dataset (50/50 split), allowing models to learn stroke characteristics effectively.



# Model Selection Strategy

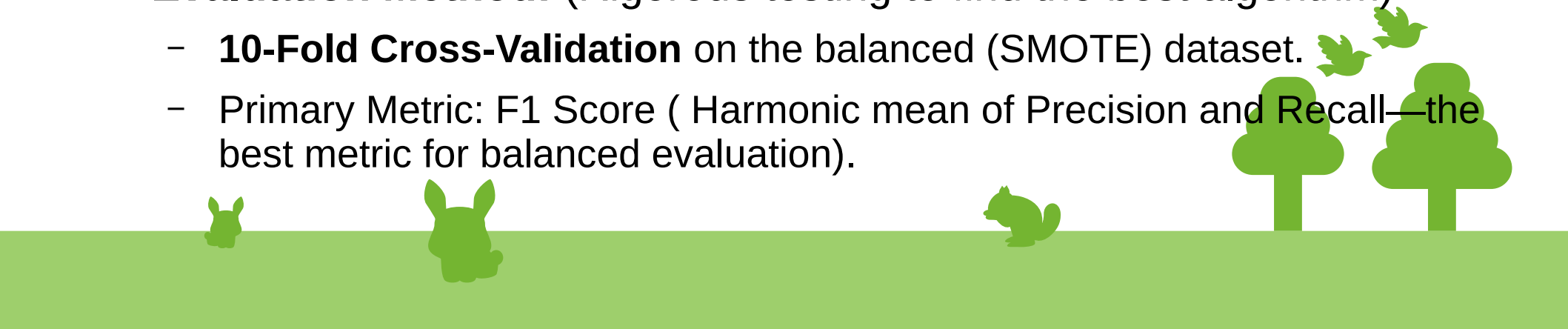
- **Models Tested:**

  - Logistic Regression - Decision Tree Classifier - Random Forest Classifier

- **Performance Matrics:**

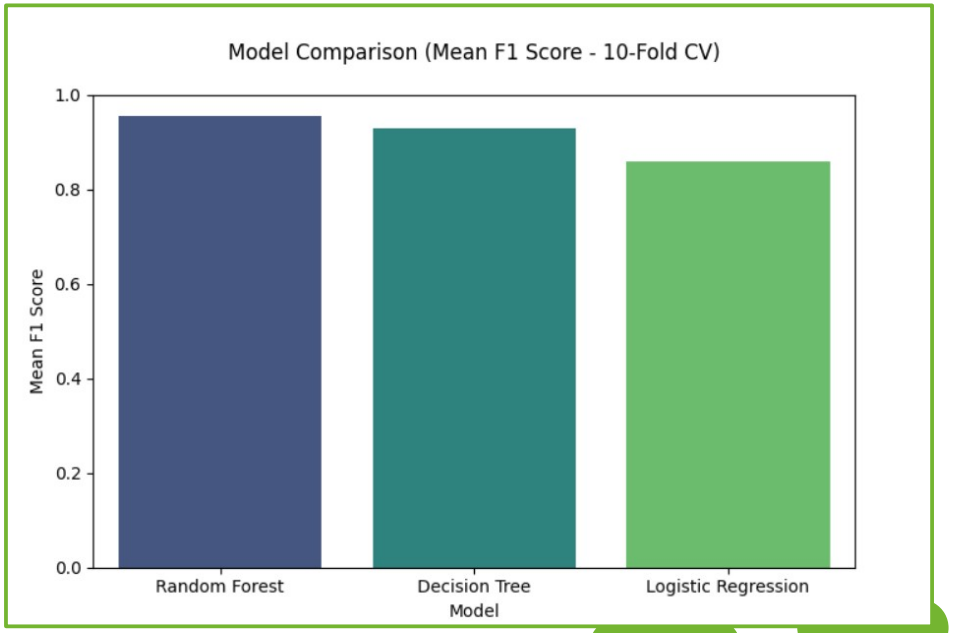
  - Accuracy - Precision - Recall - F1 Score

- **Evaluation Method:** (Rigorous testing to find the best algorithm)
  - **10-Fold Cross-Validation** on the balanced (SMOTE) dataset.
  - Primary Metric: F1 Score ( Harmonic mean of Precision and Recall—the best metric for balanced evaluation).



# Models Performances

```
===== Cross-Validation Results =====

                Model  Mean Accuracy  Std Accuracy  Mean Precision  \
2       Random Forest         0.9563        0.0414          0.9514
1       Decision Tree         0.9319        0.0292          0.9214
0  Logistic Regression         0.8599        0.0383          0.8554

   Std Precision  Mean Recall  Std Recall  Mean F1 Score  Std F1 Score
2         0.0107       0.9607      0.0797         0.9556        0.0451
1         0.0122       0.9428      0.0759         0.9296        0.0397
0         0.0168       0.8654      0.0797         0.8587        0.0489
```

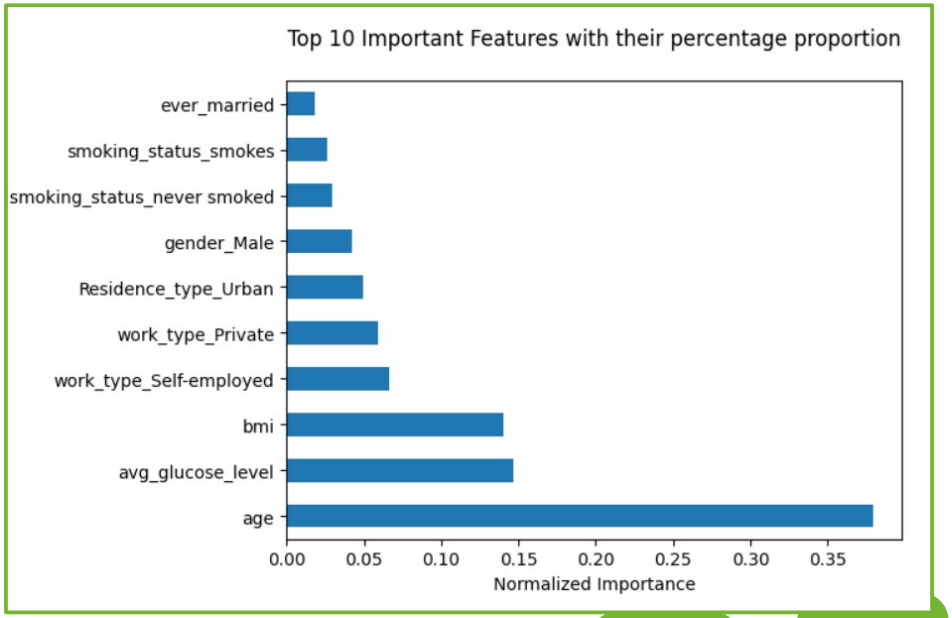


# Key Stroke Risk Factors

- **Feature Importance :**

  The model (Random Forest) identified the following as the most significant predictors of stroke risk:

- **Interpretation:** Typically, Age, Glucose Level, and BMI are top drivers, alongside specific work types or health conditions.

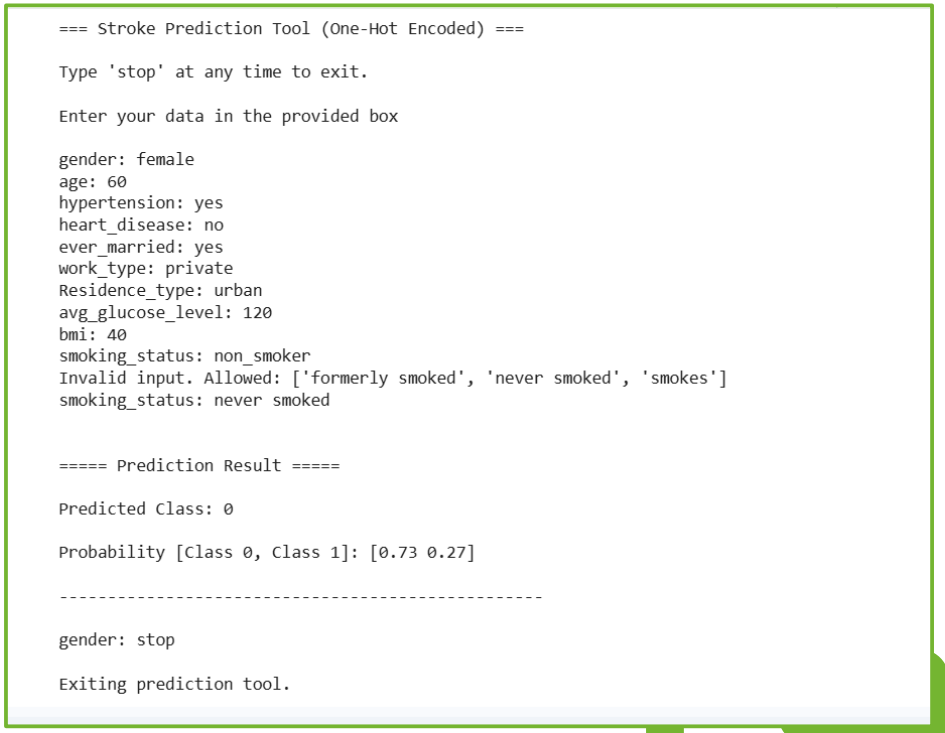


# Practical Application: Interactive Tool

We developed a Python-based interactive function that mimics real-world deployment.

**How it works:**

- **User Input:** A clinician or user enters patient details (Age, BMI, Smoking Status, etc.).
- **Preprocessing Pipeline:** The tool applies the exact same encoding (one-hot) and alignment used during training.
- **Prediction:** The trained Random Forest model outputs a classification (Stroke/No Stroke) and a probability score.



# Best Model Performance

**Train dataset:**

- Data Split: 80% Training / 20% Testing (Stratified).

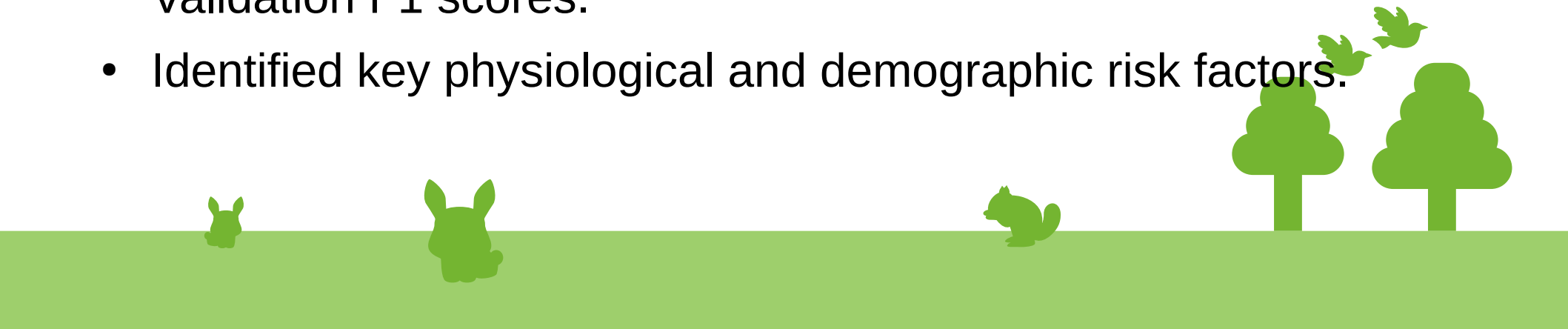
- **Performance** of the **"Best Model"** (Random Forest) on test data.

```
Best Model: Random Forest

Test Set Evaluation:

[[919  53]
 [ 37 935]]
              precision    recall  f1-score   support

           0       0.96      0.95      0.95       972
           1       0.95      0.96      0.95       972

    accuracy                           0.95      1944
   macro avg       0.95      0.95      0.95      1944
weighted avg       0.95      0.95      0.95      1944


===== Confusion Matrix Breakdown =====
True Negatives (TN): 919
False Positives (FP): 53
False Negatives (FN): 37
True Positives (TP): 935

===== Additional Evaluation Metrics =====
Accuracy:  0.9537
Precision: 0.9464
Recall:    0.9619  (Sensitivity)
Specificity: 0.9455
```

# Conclusion & Future Work

- **Summary:**

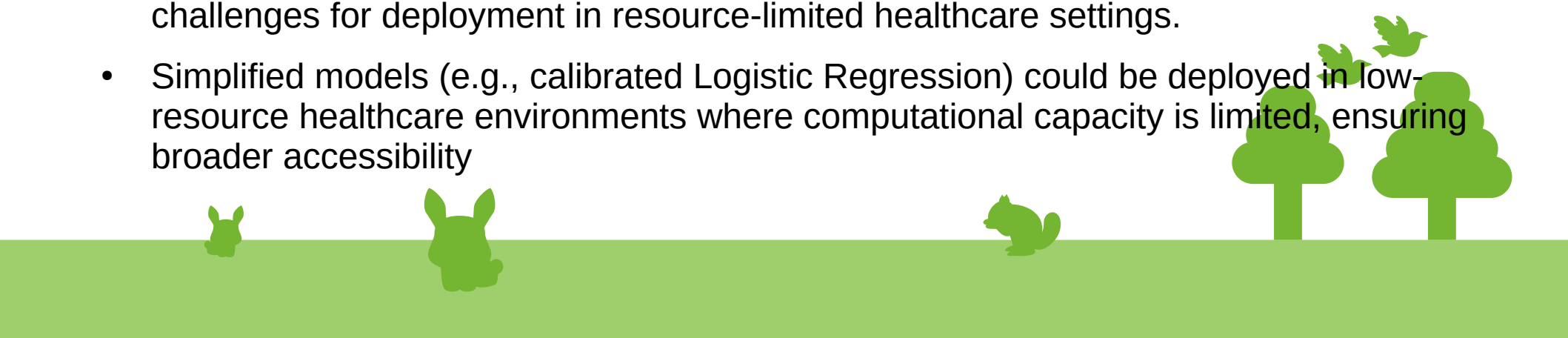
- Successfully built an end-to-end pipeline handling missing data, outliers, and severe class imbalance via SMOTE.
- Random Forest outperformed other models based on Cross-Validation F1 scores.
- Identified key physiological and demographic risk factors.



# Limitations

**Limitations:**

- No project is perfect, and we're aware of ours:
- Dataset relatively small and contains synthetic samples after SMOTE
- Some features self-reported (e.g., smoking status had 30% “Unknown”)
- No temporal or lifestyle depth (exercise, diet, family history)
- While Random Forest, it required greater computational resources. This may pose challenges for deployment in resource-limited healthcare settings.
- Simplified models (e.g., calibrated Logistic Regression) could be deployed in low-resource healthcare environments where computational capacity is limited, ensuring broader accessibility



# Future Work

So, what's next? **We'd love to**:

- Test more powerful models.
- Source a larger, more diverse, and clinically rich dataset.
- Turn our command-line tool into a user-friendly web app.

# Thank you for your time!

- I'm happy to answer any questions.
- **Thank you!**