National University
of computer and emerging sciences

Final Report

CS422 Data Science

Semester Project

Wajih Hyder 21i-2957

Ali Ahmed 21k-4922

Taqi Raza 21k-4667

Section BCS-8A

Submitted to: M Nouman Durrani

Department of Computer Science BS(CS)

FAST-NUCES Karachi

Healthcare Stroke Prediction Project Report

1. Objective

The aim of this project is to predict whether a person has experienced a stroke using a machine learning model. The dataset used includes various health and demographic factors relevant to stroke risk.

---

2. Dataset Overview

- **Source**:
  - `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`

- **Features**:

  - **Demographics**: gender, age, ever_married

  - **Health Metrics**:
    - avg_glucose_level, bmi, hypertension,heart_disease

  - **Employment and Lifestyle**:
    - work_type, smoking_status, Residence_type

  - **Target**: `stroke` (binary classification)
- **Data Head:**

	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

---

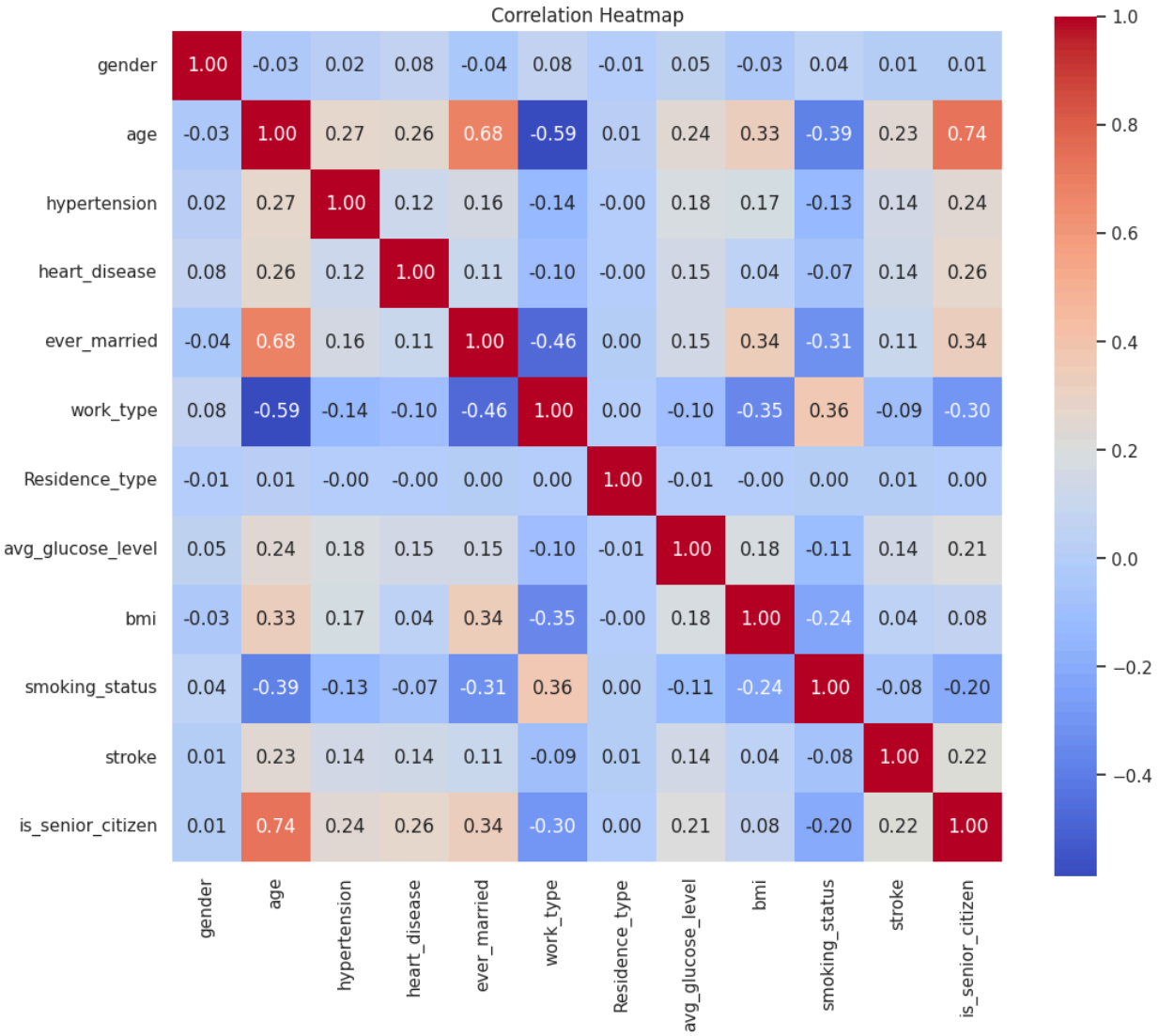
3. Data Preprocessing

- **Missing Values**:
    - bmi: Converted to numeric, and rows with null values were dropped.
    - smoking_status: Missing values were filled with 'Unknown'.
- **Irrelevant Columns**:
    - The id column was removed.
- **Encoding**:
    - Binary categorical variables (gender, ever_married, Residence_type) were label-encoded.
    - Other categorical features like work_type and smoking_status were mapped to numerical values.
- **Feature Engineering**:
    - A new binary feature is_senior_citizen was added to indicate if age >= 60.

---

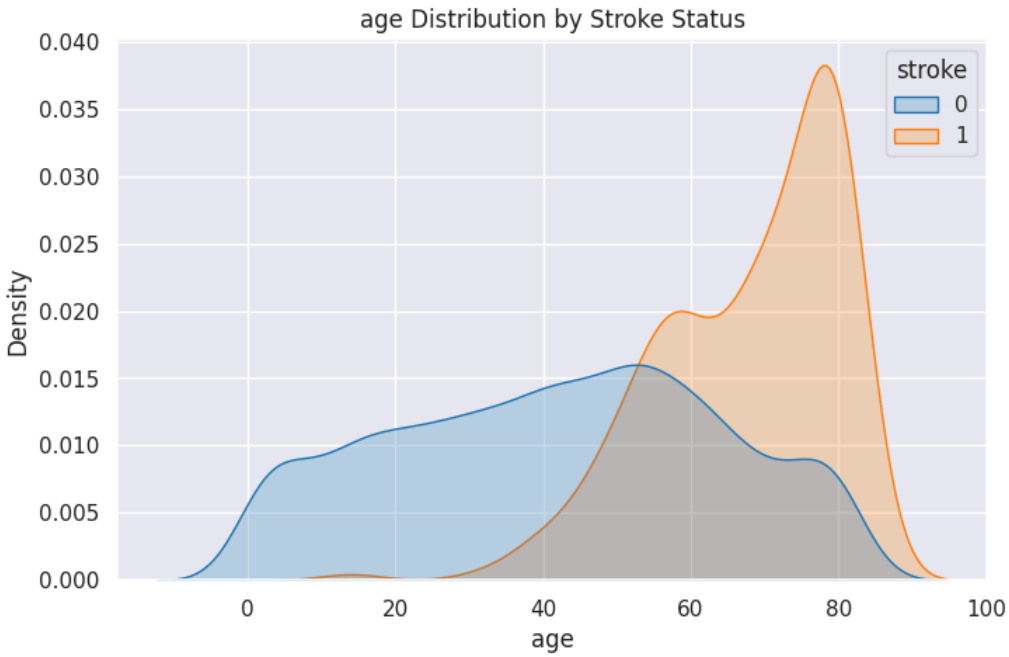
4. Exploratory Data Analysis (EDA)

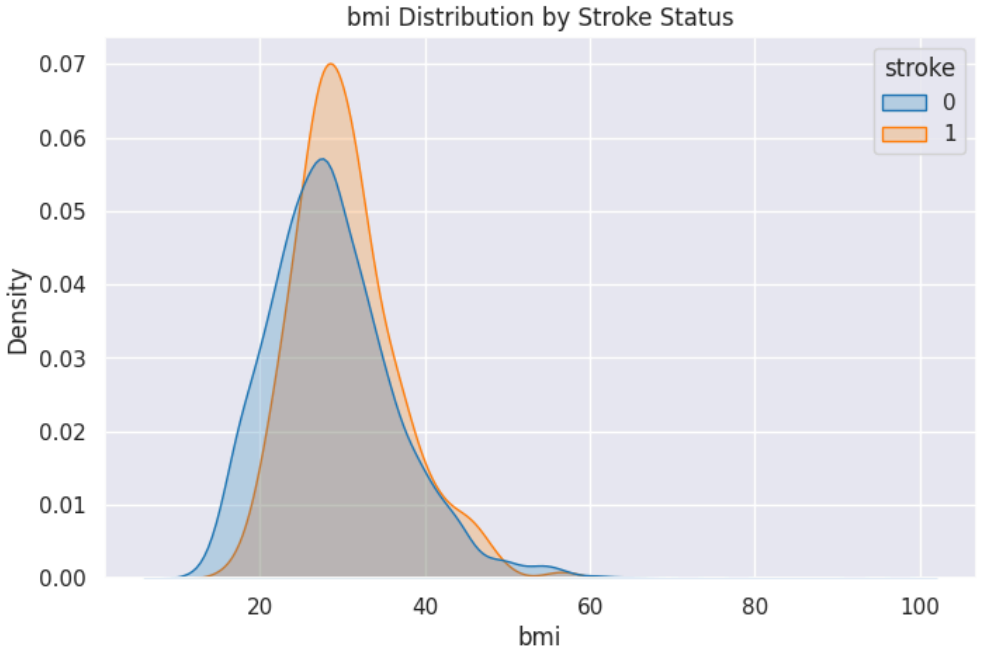
- **Correlation Matrix**:
    - A heatmap was generated to analyse feature correlations with each other and with the target.

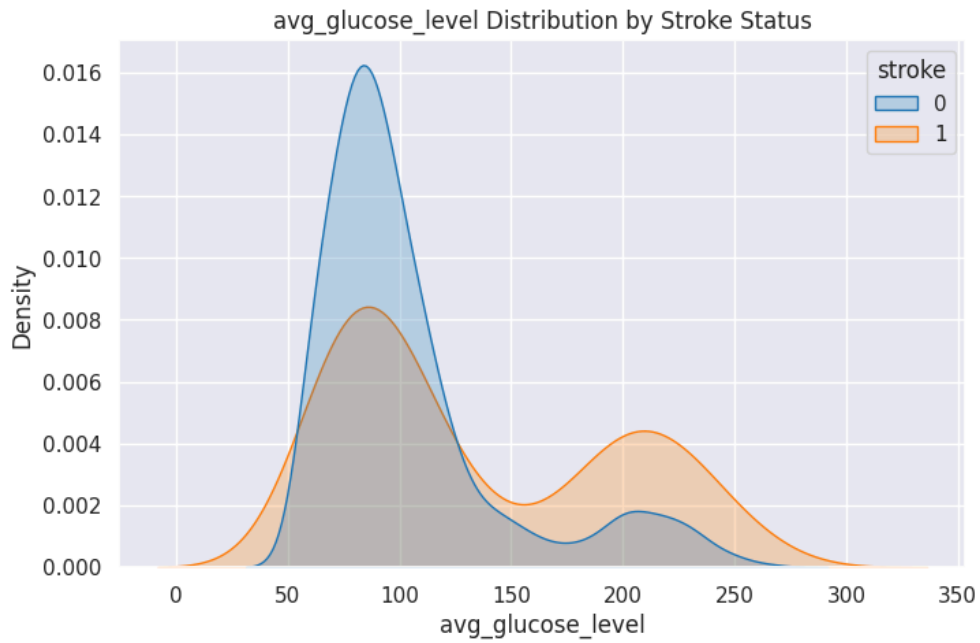


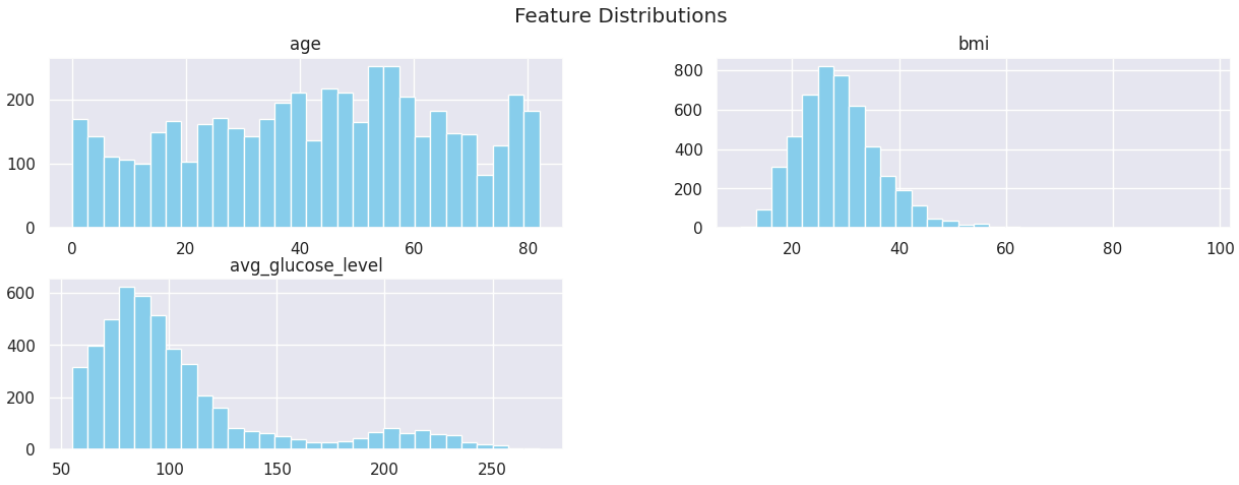
- **Distributions**:
    - KDE plots showed the distributions of age, bmi, and avg_glucose_level by stroke status.

- Histograms illustrate the general distribution of these features.









---

5. Modeling

- **Model Used**: Random Forest Classifier (n_estimators=100)
- **Train-Test Split**: 80% training, 20% testing using stratification on the target variable.

---

6. Evaluation Metrics

- **Confusion Matrix**: Used to evaluate prediction accuracy in terms of TP, FP, TN, and FN.
- **Classification Report**: Provided precision, recall, F1-score, and accuracy for the classification model.

```
print(confusion_matrix(y_test, y_pred))

[[939   1]
 [ 42   0]]
```

```
[ ] print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.96      1.00      0.98       940
           1       0.00      0.00      0.00        42

    accuracy                           0.96       982
   macro avg       0.48      0.50      0.49       982
weighted avg       0.92      0.96      0.94       982
```

---

7. Conclusion

- A complete pipeline from data loading to model evaluation was implemented.
- Exploratory analysis helped identify important patterns and correlations.
- The Random Forest model achieved results that can be improved further by:
    - Applying model tuning techniques
    - Trying advanced models like XGBoost or Gradient Boosting
    - Handling class imbalance if necessary