Project Title:

" Brain Stroke Prediction Using Machine Learning and Statistical Analysis"

Objective:

The goal of this project is to predict whether a person is likely to suffer a stroke (1 = Yes, 0 = No) using demographic and health-related features. Both statistical analysis (e.g., hypothesis testing) and machine learning (Decision Tree, Random Forest) are used to identify important risk factors and build predictive models.

Dataset Description:

- Source: Brain Stroke dataset(Kaggle)
- Type: Classification
- Target Variable: stroke (0 = No, 1 = Yes)
- Independent Variables: gender, age, hypertension, heart disease, marital status, work type, residence type, glucose level, BMI, smoking status.

Data Cleaning and Preparation

- Checked for missing values:
  → bmi had missing values → Imputed using mean.
- Converted categorical variables using label encoding.
- No duplicate rows.
- Ensured correct data types (e.g., age is float).

Exploratory Data Analysis (EDA)

Univariate Analysis(One variable at a Time)

- Most people are in the “never smoked” category.
- Majority are from the urban area and work in the private sector.
- Stroke occurred in only a small portion of the population class imbalance.
- Age and glusose are right-skewed, bmi is normally distributed.

Visuals used:

Countplots, histogram, boxplots for numeric variables(age, bmi,glucose).

Bivariate Analysis (Relationship with Stroke)

Feature	Observation
age vs stroke	Stroke is more common among people over 60.
hypertension	Higher proportion of stroke among hypertensive individuals.
heart_disease	Slightly higher stroke rate among those with heart disease.

work_type	Self-employed and private workers show more stroke cases.
smoking_status	Former smokers and current smokers have higher stroke percentages.

Visuals used:

Boxplots, stacked bar charts, grouped countplots.

Statistical Analysis

Why Chi-Square Test?

Because we want to test if **categorical variables** (like gender, smoking, hypertension) are associated with the binary outcome stroke.

Chi-square helps **test independence between two categorical variables**.

Significant Associations:

Feature	p-value	Result
Hypertension	< 0.05	Significant
Smoking Status	< 0.05	Significant
Work Type	> 0.05	Not significant
Heart Disease	< 0.05	Significant

Pearson Correlation (for numerical features):

Feature	Correlation with Stroke
Age	0.2465 (weak positive)
avg_glucose_level	0.13 (weak)
BMI	~0.04 (very weak)

Machine Learning Model Building

Algorithms Used:

- **Decision Tree Classifier**
- **Random Forest Classifier**

➢ Split Data: 80% training, 20% testing
➢ Target Variable: stroke
➢ Evaluation Metrics: Accuracy, Precision, Recall, F1 Score

Model Comparison Table:

Metric	Decision Tree	Random Forest
Accuracy	91%	95%
Precision	60%	72%
Recall	43%	65%
F1-Score	50%	68%

Observation:

Random Forest performed better than Decision Tree across all metrics and is chosen as the final model.

Final Conclusion:

- Stroke is most associated with age, hypertension, heart disease, and smoking status.
- Random Forest Classifier is recommended for predicting stroke risk.
- The project can be extended using techniques like SMOTE (for class imbalance) or logistic regression for interpretability.