Stroke Prediction Dashboard: Insights and Preventive Analytics in Power BI

-RAMNATH BHAT

1) Problem Statement: Stroke Prediction Analysis

Stroke remains a critical health issue, leading to severe consequences such as disability and death. Preventive care for stroke involves predicting high-risk individuals early to administer timely intervention. This analysis aims to predict stroke likelihood in individuals based on health and lifestyle factors, providing a data-driven approach to stroke prevention.

Objective:

- Develop a model that identifies patients at high risk for stroke.
- Use Power BI to create a dashboard that displays insights from the data and predictive model.
- Facilitate preventive strategies in healthcare settings to reduce the incidence of strokes.

2) Data Requirements

For an effective stroke prediction model, the dataset should contain various demographic, lifestyle, and medical variables. Key columns include:

- **Patient ID**: Unique identifier for each patient.
- **Age**: A critical risk factor as stroke risk increases with age.
- **Gender**: Used to assess differences in stroke occurrence between males, females, and other genders.
- **Hypertension**: Presence of high blood pressure, a significant contributor to stroke risk.
- **Heart Disease**: Indicates pre-existing heart conditions, which can increase stroke likelihood.
- **Marital Status**: Marital status may correlate with lifestyle choices impacting stroke risk.
- **Occupation Type**: Occupation often reflects lifestyle and stress levels, which can influence stroke risk.

- **Residence Type**: Urban or rural residence, as urban areas might have more access to healthcare but also higher lifestyle-related stressors.
- **Average Glucose Level**: Elevated glucose levels can increase stroke risk, especially in diabetic patients.
- **BMI**: Body Mass Index, an indicator of overall health.
- **Smoking Status**: Smoking is a known risk factor for stroke, categorized as never smoked, formerly smoked, or currently smoking.
- **Stroke Outcome**: The target variable indicating whether a patient has had a stroke.

Adding columns like diet, physical activity levels, or genetic factors could provide more accuracy, though they may be harder to obtain.

3) Data Collection

Data can be obtained from various sources:

- **Primary Sources**:
    - **Hospitals and Health Records**: Electronic health records (EHRs) provide real-time and accurate patient data.
    - **Wearable Devices and Health Apps**: Track glucose, BMI, and other health metrics directly from patients.
- **Secondary Sources**:
    - **Public Datasets**: The *Stroke Prediction Dataset* from Kaggle or the *World Health Organization (WHO)* offers access to similar data.
    - **Research Papers and Medical Studies**: Aggregate data and findings from studies that explore stroke risk factors.

4) Data Validation

Ensuring data quality is essential to produce reliable insights and predictions. Validation steps include:

- **Accuracy Checks**: Validating that values are within acceptable ranges (e.g., age should be within a realistic human range).
- **Completeness**: Ensuring no missing or null values in critical fields such as Age, Hypertension, and Heart Disease.

- **Consistency**: Confirming uniformity in categorical values (e.g., all entries under Gender should be consistently labeled).
- **Integrity Checks**: Detecting duplicates, especially in Patient ID, to prevent erroneous records.
- **Timeliness**: Confirming that the data is recent enough to remain relevant, especially when medical practices and health profiles change over time.

5) Data Cleaning

Data cleaning enhances data quality by addressing missing values, outliers, and inconsistencies. Cleaning steps include:

- **Handling Missing Values**:
    - For non-critical missing values, imputation methods like mean, median, or mode can be used.
    - For critical columns (like BMI), deletion may be necessary if imputing values would compromise accuracy.
- **Outlier Detection and Removal**:
    - Identifying and addressing outliers in variables like glucose level or BMI, which can skew model predictions.
- **Data Transformation**:
    - **Label Encoding**: Applied to binary variables (e.g., Ever Married) to convert them into numeric format.
    - **One-Hot Encoding**: Applied to categorical variables with more than two categories (e.g., Work Type or Smoking Status).
- **Normalization**:
    - Scaling continuous variables (like BMI, glucose levels) to a standard range for balanced model training.
- **Dataset Balancing**:
    - Using techniques like SMOTE (Synthetic Minority Oversampling Technique) to handle class imbalance, ensuring that the minority class (stroke cases) is adequately represented.

6) Tools

1. **Python**: For data analysis, cleaning, feature engineering, and machine learning.
    - *Pandas* and *NumPy*: Data manipulation and cleaning.
    - *Scikit-Learn* and *XGBoost*: Model building and evaluation.
    - *Matplotlib* and *Seaborn*: Initial data visualizations.
2. **SQL**: For data extraction, filtering, and basic cleaning from databases, enabling structured access to large datasets.
3. **Excel**: For preliminary data exploration, quick data summaries, and sharing simplified reports with non-technical stakeholders.
4. **Power BI:**

- **Overview**: Power BI is a data visualization and business intelligence tool developed by Microsoft, allowing users to transform raw data into meaningful insights.
- **Data Connection and Transformation**: Power BI can connect to multiple data sources, clean, transform, and combine data for easier analysis.
- **Visualizations**: Power BI provides a variety of visuals, including line charts, bar charts, maps, and custom visuals for comprehensive data analysis.

Charts Used:

- **Univariate Analysis**:
    - **Bar Chart**: Used to display categorical variables like Gender and Marital Status.
    - **Histogram**: Visualizes distribution for continuous variables like Age and BMI.
- **Bivariate Analysis**:
    - **Scatter Plot**: Examines relationships between two continuous variables (e.g., BMI and Glucose Level).
    - **Side-by-Side Bar Chart**: Useful for comparing stroke occurrences across categories such as Smoking Status.
- **Multivariate Analysis**:
    - **Heatmaps**: Displays relationships among multiple variables, such as the correlation between age, BMI, and stroke occurrence.
    - **3D Scatter Plot**: Visualizes complex interactions, for example, among Age, BMI, and Glucose Level.

7) Dashboard

Dashboard Overview: Stroke Prediction Analysis

This dashboard enables healthcare professionals and analysts to understand demographic and health-related factors affecting stroke risk. By identifying high-risk groups and key predictive factors, it supports proactive interventions, helping reduce stroke incidence and improve patient outcomes. The interactive nature allows users to filter by specific demographics or health metrics, facilitating personalized insights tailored to particular patient profiles.

1. Overview Page

- **Total Stroke Cases**: A KPI card showing the number of stroke occurrences in the dataset, giving an immediate overview.
- **Gender Distribution**: A pie chart showing the proportion of male, female, and other genders in the dataset to analyze gender-based patterns in stroke risk.
- **Age Distribution**: A histogram displaying the age range of patients, with emphasis on age groups most associated with stroke risk.
- **Stroke Risk by Residence Type**: A bar chart comparing stroke occurrences across urban and rural settings, indicating the impact of environment on stroke risk.

2. Demographics Analysis

- **Gender vs. Stroke Occurrence**: A bar chart that breaks down stroke occurrences by gender, allowing easy comparison between male and female patients.
- **Marital Status Impact**: A pie chart or bar chart visualizing stroke rates among married vs. unmarried individuals, providing insights into lifestyle-related risk factors.
- **Occupation Type Analysis**: A bar chart showing stroke frequency by job type (e.g., private, self-employed, government job), highlighting occupational influences on stroke.

3. Health Metrics Correlation

- **Average BMI, Blood Pressure, and Glucose Levels**: Cards display these average health metrics among stroke patients, showing baseline health indicators.
- **Smoking Status vs. Stroke Occurrence**: A stacked bar chart or grouped bar chart showing stroke occurrence across different smoking statuses (current smoker, former smoker, non-smoker).
- **Hypertension and Diabetes Impact**: A heatmap illustrating the relationship between hypertension, diabetes, and stroke incidence, highlighting high-risk combinations.

8) Storytelling

What is Storytelling? Storytelling in data analysis involves crafting a narrative that guides users through insights in a meaningful way, allowing them to connect with and act on the data effectively.

Importance of Storytelling By presenting data insights in a logical, engaging sequence, storytelling transforms raw data into an accessible and actionable format. This approach ensures that healthcare professionals can quickly understand the key findings and take targeted actions based on the data.

Dashboard Storyline

The storyline for the stroke prediction dashboard follows a logical flow, allowing healthcare professionals to explore stroke risk factors and identify high-risk groups effectively:

1. **Starting with the Big Picture: Overview Page**
    - The story begins by highlighting the **Total Stroke Cases**, giving an immediate sense of the scope of the problem.
    - **Gender Distribution** shows whether stroke risk varies across genders, while the **Age Distribution** emphasizes age groups most affected, helping users quickly identify demographic patterns.
    - **Stroke Risk by Residence Type** adds context by showing how location (urban vs. rural) may influence stroke risk. These insights help establish a baseline understanding of stroke distribution across different population segments.
2. **Diving Deeper: Demographics Analysis**
    - The **Gender vs. Stroke Occurrence** chart further dissects the gender differences in stroke cases, helping healthcare professionals see how these demographic details contribute to stroke risk.
    - **Marital Status Impact** provides insights into lifestyle factors, showing how marital status may affect stroke occurrence.
    - **Occupation Type Analysis** examines the relationship between different job types and stroke frequency, identifying occupational factors that may elevate stroke risk. This deeper dive into demographic factors aids in building a holistic picture of at-risk groups.
3. **Exploring Health Metrics: Health Metrics Correlation**
    - Key health indicators such as **Average BMI, Blood Pressure, and Glucose Levels** are displayed, providing baseline health statistics for stroke patients.

- **Smoking Status vs. Stroke Occurrence** chart shows the relationship between smoking habits and stroke risk, helping healthcare providers understand the impact of lifestyle choices.
- **Hypertension and Diabetes Impact** heatmap illustrates how these conditions correlate with stroke incidence, identifying high-risk groups for more targeted interventions.

The Narrative Flow This dashboard provides a data-driven story that transitions from general information on stroke cases to specific demographic and health factors affecting stroke risk. By the end of the journey, healthcare professionals gain comprehensive insights into key risk factors and are equipped to make informed, targeted decisions to improve patient outcomes.