SYNOPSIS

TITLE – MOST STREAMED SONGS ON SPOTIFY

Vishal C Halkodu_AF0377844

ABOUT THE TOPIC (DATASET):

Data Structure

The dataset should be organized in a tabular format, where each row represents a song and the artists. The columns might include:

- **track_name**: The name of the track.
- **artist(s)_name**: The names of the artists featured in the track.
- **artist_count**: The number of artists involved in the track.
- **released_year**: The year the track was released.
- **released_month**: The month the track was released.
- **released_day**: The day the track was released.
- **in_spotify_playlists**: The number of Spotify playlists that include the track.
- **in_spotify_charts**: The number of Spotify charts that include the track.
- **streams**: The total number of streams the track has received on Spotify.
- **in_apple_playlists**: The number of Apple Music playlists that include the track.
- **in_apple_charts**: The number of Apple Music charts that include the track.
- **in_deezer_playlists**: The number of Deezer playlists that include the track.
- **in_deezer_charts**: The number of Deezer charts that include the track.
- **in_shazam_charts**: The number of Shazam charts that include the track.
- **bpm**: The beats per minute of the track.
- **key**: The musical key of the track.
- **mode**: The mode of the track (e.g., Major or Minor).
- **danceability_%**: The danceability percentage of the track.
- **valence_%**: The valence (musical positivity) percentage of the track.
- **energy_%**: The energy percentage of the track.
- **acousticness_%**: The acousticness percentage of the track.
- **instrumentalness_%**: The instrumentalness percentage of the track.
- **liveness_%**: The liveness percentage of the track.
- **speechiness_%**: The speechiness percentage of the track.

Data Preprocessing

1. Data Cleaning: Handle missing values and ensure data consistency.

2. Normalization: Normalize numeric features if required (e.g., age, cholesterol levels).
3. Encoding Categorical Variables: Encode categorical variables (e.g., chest pain type, thalassemia) using one-hot encoding or label encoding.

Analysis Techniques

1. Time Series Analysis
- Trend Analysis: Identify trends in heart disease diagnosis rates over time.
- Seasonality Detection: Detect seasonal patterns in heart disease occurrences.
- Anomaly Detection: Identify anomalies in the data, which could indicate unusual spikes or drops in heart disease cases.

2. Predictive Modelling
- Classification Models: Use logistic regression, decision trees, or more advanced models like Random Forest, Gradient Boosting Machines, or neural networks to predict the likelihood of heart disease based on patient attributes.
- Survival Analysis: Analyze the time until a heart disease event occurs, using techniques like Kaplan-Meier estimation or Cox proportional hazards models.

3. Correlation Analysis
- Risk Factors: Explore correlations between various risk factors (e.g., cholesterol level, blood pressure) and the presence of heart disease.
- Demographic Analysis: Analyze correlations between demographic variables (e.g., age, gender) and heart disease incidence.

4. Cluster Analysis
- Patient Segmentation: Use clustering algorithms (e.g., K-means, hierarchical clustering) to segment patients into groups based on similarities in their attributes and risk factors.
- Risk Profiling: Identify common characteristics of high-risk groups.

Implementation Steps

1. Data Ingestion: Load the dataset into a data analysis environment (e.g., Python, R).
2. Preprocessing: Clean and prepare the data for analysis.
3. Exploratory Data Analysis (EDA): Conduct EDA to understand data distribution and initial patterns.
4. Modeling: Develop and validate models for classification, time series analysis, and clustering.
5. Visualization: Create visualizations to communicate insights effectively (e.g., correlation heatmaps, survival curves, cluster visualizations).
6. Reporting: Summarize findings in reports or dashboards for stakeholders.

Data set: https://onyxdata.ck.page/a12261b1fb

Technologies: pandas, Microsoft Excel, Microsoft PowerBi, seaborn, matplotlib.

Software Requirements:

Operating System – Windows, Linux and mac

IDLE – Jupyter Notebook

Hardware Requirements:

RAM – Minimum 4GB

Processor – Minimum intel i3