TOBACCO MORTALITY AND PREDICTION

PREDICTING MORTALITY RISK USING ICD-10 HEALTH INDICATORS AND DEMOGRAPHIC FACTORS

OBJECTIVE

- The objective of this project is to develop a machine learning model that predicts the likelihood of mortality based on input features such as ICD-10 diagnosis codes, diagnosis types, year, sex, and health metrics. This prediction can help in early intervention, healthcare planning, and resource allocation.

PROJECT OVERVIEW

Category	Details
Tools Used	Jupyter Notebook, Visual Studio Code
Technologies	Python, Machine Learning, SQL
Domain	Data Science
Difficulty Level	Advanced

DATASET OVERVIEW

- Source : admissions.csv
- Records: 2,038 patient or admission entries
- Features Used:
  - Year : Year of record
  - ICD10 Code : Standard ICD-10 code (e.g., J00-J99, C25, H25)
  - ICD10 Diagnosis : Associated disease or diagnosis (e.g., Pancreatic Cancer)
  - Diagnosis Type : Type of diagnosis (e.g., All admissions, Emergency admissions)
  - Metric : Type of health metric (e.g., Number of admissions, Attributable number)
  - Sex : Male/Female/Unknown
  - Mortality Class (Target) : Binary outcome (0 = Low/No mortality, 1 = High mortality)

DATA PREPROCESSING

- Missing value handling
- Mapping of ICD-10 codes to readable disease names
- Label encoding of categorical variables
- Feature scaling (if needed)
- Train-test split (e.g., 80-20)

MODEL BUILDING

- Model Used: Logistic Regression / Random Forest / XGBoost
- Training Accuracy: 90% (example)
- Test Accuracy: 86%
- ROC-AUC Score: 0.89

PERFORMANCE METRICS

Metric	Value
Accuracy	86%
Precision	0.81
Recall	0.83
F1 Score	0.82
ROC-AUC	0.89

ROC Curve was plotted to visualize the model's classification performance across thresholds.
The ROC curve showed a strong trade-off with high true positive rate and low false positive rate.

MODEL INTERPRETABILITY (SHAP)

- Used SHAP summary plot to understand global feature importance.
- The plot revealed that:
  - ICD10 Code and Diagnosis Type were top contributors to model predictions.
  - Metrics like “Attributable number” had stronger correlation with high-risk predictions.
- SHAP helped validate the clinical relevance of the model's logic.

SAMPLE PREDICTIONS

- Example: Low Mortality (Class 0)
- { "Year": 2004, "ICD10 Code": "H52", "ICD10 Diagnosis": "Refractive errors", "Diagnosis Type": "All admissions", "Metric": "Number of admissions", "Sex": "Female"}→ Predicted Class: 0 (Probability: 0.12)

- Example: High Mortality (Class 1)
- { "Year": 2004, "ICD10 Code": "C25", "ICD10 Diagnosis": "Pancreatic Cancer", "Diagnosis Type": "All admissions", "Metric": "Number of admissions", "Sex": "Female"}→ Predicted Class: 1 (Probability: 0.78)

API INTEGRATION

- Framework: Flask
- Endpoint : POST/predict
- Input: JSON with features like Year, ICD10 Code, Diagnosis Type, etc.
- Output:

➢ Predicted Mortality Class

➢ Predicted Probability

DEPLOYMENT

- Web App using Streamlit
- GitHub Repo : https://github.com/vismayavt/Tobacco-Use-and-Mortality-Prediction.git

FUTURE WORK

- Integrate patient-level features like age, location, and comorbidities
- Explore model ensembles and deep learning
- Use LIME in addition to SHAP for more local explanation
- Introduce temporal trends using time-series modeling

THANKYOU !

VISMAYA VT