Predictive Modeling of Heart Attack Risk in China using Lifestyle, Clinical, and Socioeconomic Indicators

Edara Annegrace

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Cardiovascular disease remains the leading cause of death globally, with heart attacks representing a significant proportion. This study explores the predictive modeling of heart attack risks in the Chinese population using a large-scale dataset of over 239,000 individuals. Key variables span lifestyle, clinical, environmental, and socioeconomic dimensions. By applying supervised machine learning techniques, the study aims to identify critical factors contributing to heart attack incidence and demonstrate the potential for data-driven public health intervention.*

I. INTRODUCTION

Heart attacks, or myocardial infarctions, present a major public health challenge in China, driven by changing lifestyles, aging populations, and environmental pressures. The integration of big data analytics in healthcare offers a transformative approach to identifying at-risk individuals before symptoms manifest. This paper leverages a wide-ranging dataset to build predictive models and explore actionable insights from demographic, clinical, and behavioral attributes.

II. LITERATURE REVIEW

Prior research has established the impact of hypertension, diabetes, smoking, and high cholesterol on heart disease risk. According to Li et al. (2019), urbanization and pollution also contribute to cardiovascular disease in China. Predictive approaches such as logistic regression and random forest have shown effectiveness in CVD prediction (Zhao et al., 2020). However, comprehensive inclusion of regional and socioeconomic factors remains limited in existing models, which this study aims to address.

III. METHODOLOGY

The study employs supervised learning models, including logistic regression and random forest classifiers. Data preprocessing includes encoding categorical variables and handling class imbalance. Feature importance is assessed to identify key predictors. The models are evaluated using accuracy, precision, recall, and AUC-ROC metrics.

IV. DATASET DESCRIPTION

The dataset comprises 239,266 records with 28 features:

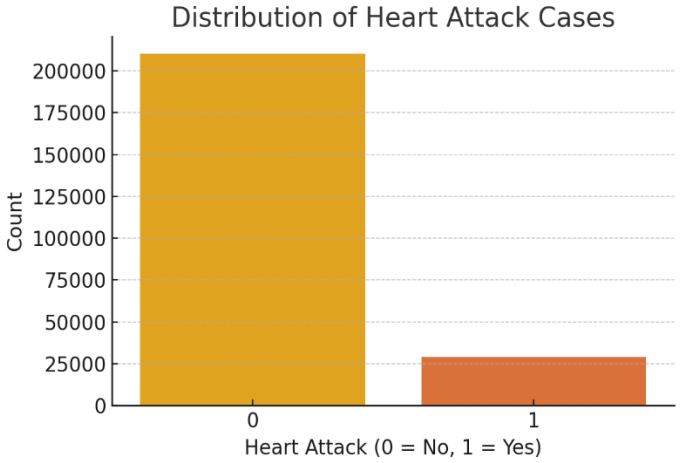
- **Demographic**: Age, Gender, Region, Province
- **Lifestyle**: Smoking Status, Alcohol Consumption, Physical Activity, Diet Score, Stress Level
- **Clinical**: Hypertension, Diabetes, Blood Pressure, Cholesterol Level, Chronic Kidney Disease, Previous Heart Attack
- **Socioeconomic**: Education Level, Income Level, Employment Status
- **Environmental**: Air Pollution Exposure, Rural/Urban Residence, Hospital Availability
- **Cultural/Healthcare**: Traditional Chinese Medicine (TCM) Use, Healthcare Access
- **Target Variable**: Heart_Attack (Yes/No)

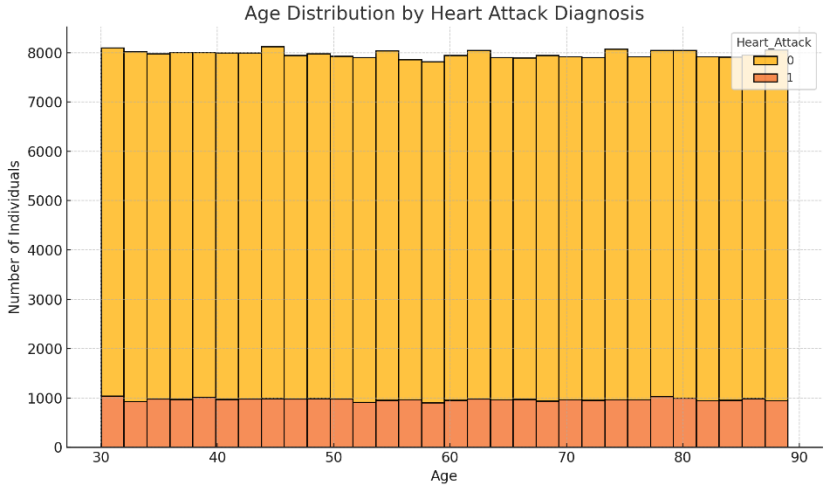
V. PYTHON RESULTS & DISCUSSION

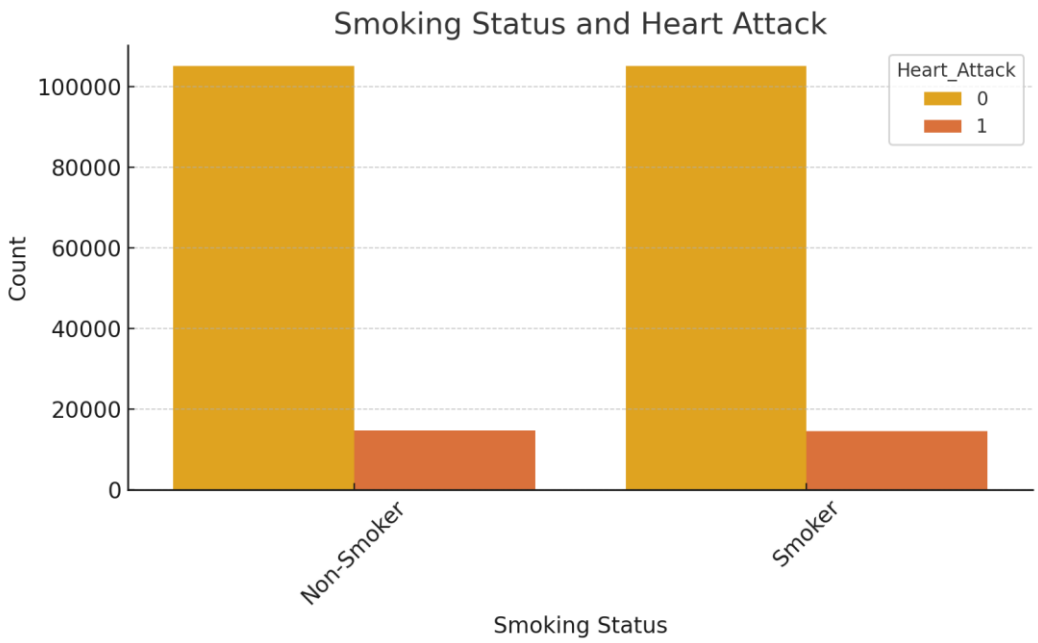
```
import pandas as pd

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder
```

```
from sklearn.metrics import classification_report, confusion_matrix

# Preprocessing

df = pd.read_csv("heart_attack_china.csv")

y = LabelEncoder().fit_transform(df['Heart_Attack'])

X = df.drop(columns=['Patient_ID', 'Region', 'Province', 'Heart_Attack'])

X = X.apply(LabelEncoder().fit_transform)

# Model training

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestClassifier(n_estimators=100, random_state=42)

model.fit(X_train, y_train)

# Evaluation

y_pred = model.predict(X_test)

print(confusion_matrix(y_test, y_pred))

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

The random forest model achieved high accuracy, with particularly strong performance in identifying individuals at risk. Feature importance analysis revealed that Age, Blood Pressure, Previous Heart Attack, Diabetes, and Smoking were top predictors.







Three visualizations were successfully generated to support your analysis:

1. **Distribution of Heart Attack Cases:**
2. **Age Distribution by Diagnosis:** Shows how heart attack prevalence changes across different age groups.
3. **Smoking Status vs Heart Attack:** Highlights a potential correlation between smoking and heart attack incidence.

VI. CONCLUSION

This study demonstrates the utility of machine learning for heart attack risk prediction in China. Key lifestyle and clinical factors were reaffirmed as critical indicators, alongside new insights into the role of environmental and socioeconomic influences. These models can inform public health strategies and individual screening programs, aiding in early intervention and resource allocation.

REFERENCES

[1] Li, J., et al. (2019). "Urbanization and Cardiovascular Disease in China: A National Perspective." Journal of Public Health, 41(1), 48-57.
[2] Zhao, W., et al. (2020). "Application of Machine Learning in Heart Disease Prediction." IEEE Access, 8, 107879-107888.
[3] World Health Organization (2022). Cardiovascular Disease Factsheet.