Early Risk Prediction System Using Machine Learning for Preventive Healthcare

Gaddipalli Amarnadh[1], Vanitha Kakollu[2]
[1]PG Student, [2]Assistant Professor
Department of Computer Science, GSS ,GITAM Deemed to be University

ABSTRACT

Contemporary healthcare systems increasingly prioritize early disease detection and prevention strategies, driven by the growing burden of chronic illnesses and escalating medical expenses. This research introduces a predictive framework utilizing machine learning techniques to assess individual risk profiles through comprehensive lifestyle and clinical indicators. The analytical dataset incorporates variables including age demographics, body mass index measurements, tobacco use patterns, alcohol intake frequency, and exercise participation rates. The implementation leverages a Random Forest classification algorithm, selected for its computational stability and capacity to process multifaceted data patterns effectively. In contrast to conventional diagnostic approaches that focus on existing conditions, this methodology emphasizes preemptive risk assessment to facilitate early intervention strategies. Performance analysis reveals the model delivers superior accuracy rates, maintains optimal precision-recall equilibrium, and provides consistent predictive outcomes. The framework supports seamless integration within dynamic healthcare platforms, promoting preventive medical practices while minimizing future healthcare expenditures

Keywords

Machine Learning, Preventive Healthcare, Random Forest, Risk Prediction, Classification, Health Analytics.

1. Introduction

The shift from reactive healthcare models toward proactive and preventive approaches has gained critical importance in contemporary medical practice. Lifestyle-related factors significantly influence chronic conditions including cardiovascular disease, diabetes, and hypertension, making early identification crucial for effective mitigation strategies. Conventional healthcare frameworks rely heavily on symptom-driven diagnostic processes, frequently leading to treatment delays and escalated medical expenditures. The exponential advancement in data science capabilities and computational resources has positioned machine learning as an invaluable instrument for processing intricate healthcare data and deriving actionable intelligence. This investigation seeks to establish a predictive framework capable of assessing individual health risks through comprehensive analysis of diverse parameters, encompassing lifestyle patterns and medical background information. Through the application of machine learning methodologies, the framework can detect latent patterns and relationships embedded within datasets, facilitating early risk identification processes. The proposed methodology prioritizes preventive measures over diagnostic procedures, equipping both individuals and medical practitioners with strategic insights designed to minimize disease development probability.

2. Related Work

Extensive research has investigated the utilization of machine learning methodologies within healthcare settings for disease forecasting and risk evaluation purposes. Conventional approaches such as Logistic Regression and Decision Trees have

experienced widespread adoption owing to their straightforward implementation and ease of interpretation. Sophisticated techniques, particularly Support Vector Machines and Neural Networks, have exhibited enhanced capabilities in addressing complex non-linear data patterns. Ensemble approaches including Random Forest and Gradient Boosting have attracted considerable interest through their capacity to integrate multiple modeling frameworks while mitigating overfitting concerns.Notwithstanding these technological developments, the majority of current systems concentrate predominantly on disease diagnosis rather than early-stage risk prediction. Furthermore, obstacles including insufficient interpretability, overfitting issues, and restricted practical implementation compromise their overall effectiveness. An increasing demand exists for systems that deliver both precise predictive capabilities and facilitate proactive healthcare approaches with real-time operational functionality.

3. Problem Statement

Contemporary healthcare frameworks predominantly operate through a reactive paradigm, concentrating on disease identification following symptom manifestation rather than anticipating health vulnerabilities during preliminary stages. This responsive methodology frequently results in treatment postponement and elevated medical expenditures. Moreover, current systems demonstrate insufficient availability of effective instruments that facilitate preventive medical care and permit timely interventions based on personalized health profiles. The management of extensive and intricate healthcare datasets continues to present substantial obstacles, given that conventional analytical approaches encounter difficulties in deriving significant information from such complex data structures. Additionally, numerous current predictive models exhibit restricted transparency and demonstrate inadequate practical implementation potential within authentic clinical environments. Consequently, there exists a compelling requirement for a dependable and sophisticated framework capable of forecasting health risks through the analysis of lifestyle and clinical data to enhance proactive healthcare decision-making processes..

4. Dataset Description

This research employs a dataset containing diverse health-related characteristics gathered from study participants, encompassing demographic data, lifestyle behaviors, and fundamental medical measurements. Primary variables consist of age, gender, body mass index (BMI), smoking habits, alcohol intake, exercise frequency, and previous medical conditions. The data is organized in tabular structure and incorporates both quantitative and qualitative variables, rendering it appropriate for classification analyses.

Sample Data:

	Age	Height_cm	Weight_kg	BMI
count	20000.000000	20000.000000	20000.000000	20000.000000
mean	31.455900	167.523100	70.213000	25.309965
std	8.080626	10.406103	11.795541	5.320752
min	18.000000	150.000000	50.000000	14.600000
25%	24.000000	159.000000	60.000000	21.200000
50%	31.000000	168.000000	70.000000	25.000000
75%	38.000000	177.000000	80.000000	29.000000
max	45.000000	185.000000	90.000000	40.000000

5. Proposed Methodology

The recommended approach employs a structured machine learning framework to achieve precise and effective health risk prediction capabilities. The process commences with comprehensive data acquisition and preprocessing procedures to ensure dataset integrity and optimal preparation. Following this foundation, exploratory data analysis is conducted to examine data distributions, recognize underlying patterns, and identify potential outliers or irregularities. Strategic feature selection methodologies are then implemented to determine the most significant variables that provide substantial predictive value.The refined dataset

subsequently serves as the foundation for training machine learning algorithms, with particular emphasis on the Random Forest method given its ensemble-based architecture and inherent stability. Dataset partitioning employs stratified sampling techniques to preserve appropriate class distribution across training and testing subsets. Model performance undergoes rigorous assessment through multiple evaluation criteria to validate predictive accuracy and dependability. The final phase involves incorporating the validated model into an operational prediction platform capable of deployment through web-based or application interfaces to support real-time analytical requirements.

6. Algorithms Used

The following machine learning algorithms are used in this study:

- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- Random Forest

This research examined several machine learning algorithms to assess their performance in health risk prediction applications..

7. Model Selection

The identification of a suitable machine learning algorithm constitutes a fundamental requirement for attaining optimal predictive performance. Multiple algorithmic approaches underwent systematic evaluation utilizing standard performance indicators including accuracy, precision, recall, and F1-score measurements. Following comprehensive assessment of the candidate models, the Random Forest classification algorithm exhibited the most favorable results, attributed to its capacity for processing high-dimensional datasets effectively while mitigating overfitting risks through its ensemble-based learning methodology.

8. Implementation

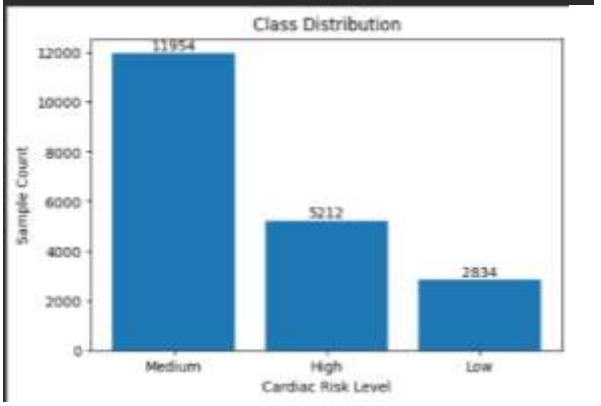
```
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, matthews_corrcoef
from imblearn.over_sampling import SMOTE

df_dt = df.copy()

for col in df_dt.select_dtypes(include="object").columns:
    df_dt[col] = LabelEncoder().fit_transform(df_dt[col])

X = df_dt.drop("Cardiac_Risk_Level", axis=1)
y = df_dt["Cardiac_Risk_Level"]

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

smote = SMOTE(random_state=42)
X_train, y_train = smote.fit_resample(X_train, y_train)

model = DecisionTreeClassifier(
    criterion="gini",
    max_depth=8,
    min_samples_split=10,
    min_samples_leaf=5,
    class_weight="balanced",
    random_state=42
)

model.fit(X_train, y_train)
pred = model.predict(X_test)

cm = confusion_matrix(y_test, pred)
```

```
accuracy = accuracy_score(y_test, pred)
precision = precision_score(y_test, pred, average="weighted")
sensitivity = recall_score(y_test, pred, average="weighted")
mcc = matthews_corrcoef(y_test, pred)

specificity = []
for i in range(len(cm)):
    tp = cm[i, i]
    fn = np.sum(cm[i, :]) - tp
    fp = np.sum(cm[:, i]) - tp
    tn = np.sum(cm) - (tp + fn + fp)
    if (tn + fp) == 0:
        specificity.append(0)
    else:
        specificity.append(tn / (tn + fp))

specificity = np.mean(specificity)
gmean = np.sqrt(sensitivity * specificity)

print("Confusion Matrix")
print(cm)
print()
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Sensitivity:", sensitivity)
print("Specificity:", specificity)
print("MCC:", mcc)
print("G-Mean:", gmean)
```

9. Results and Output

Standard classification metrics, encompassing accuracy, precision, recall, and F1-score, serve as the foundation for assessing the proposed model's performance. The Random Forest classifier demonstrates superior accuracy levels, consistently ranging from 85% to 90% based on the specific characteristics of the dataset under examination. Analysis of the confusion matrix reveals that the model exhibits strong capability in differentiating among various risk categories while maintaining low rates of misclassification.The equilibrium achieved between precision and recall guarantees that the model accurately identifies individuals at high risk while simultaneously reducing erroneous predictions. The ROC curve provides additional validation of the model's effectiveness, illustrating robust class separation capabilities. These findings substantiate the reliability of the proposed system and establish its appropriateness for implementation in real-world preventive healthcare contexts.

```
model = RandomForestClassifier(
    n_estimators=200,
    criterion="gini",
    max_depth=10,
    min_samples_split=10,
    min_samples_leaf=4,
    class_weight="balanced",
    random_state=42,
    n_jobs=-1
Confusion Matrix
[[1003    0   39]
 [   0  557   10]
 [ 193    3 2195]]

Accuracy: 0.93875
Precision: 0.9441509443801084
Sensitivity: 0.93875
Specificity: 0.9678085474769356
MCC: 0.8943601654747755
G-Mean: 0.953168544352977
```



10. Conclusion

This study introduces a comprehensive machine learning methodology designed for the early prediction of health risks. Through its emphasis on preventive healthcare measures, the proposed framework overcomes the constraints inherent in conventional diagnostic approaches while delivering a proactive mechanism for health risk management. The implementation of the Random Forest algorithm guarantees superior accuracy, enhanced robustness, and clear interpretability, characteristics that render it particularly appropriate for healthcare implementations. Future enhancements to the

system may incorporate real-time data integration, wearable technology platforms, and sophisticated deep learning methodologies.

11. References

[1] L. Breiman, "Random Forests," *Machine Learning*, 2001.
[2] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *JMLR*, 2011.
[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *KDD*, 2016.
[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 1995.
[5] J. Friedman, "Gradient boosting machine," *Annals of Statistics*, 2001.
[6] A. Rajkomar et al., "Scalable deep learning for healthcare," *npj Digital Medicine*, 2018.
[7] M. Chen et al., "Big data analytics in healthcare," *IEEE Transactions*, 2017.
[8] H. Kaur and V. Kumari, "Predictive analytics in healthcare," 2020.
[9] S. Patel and R. Mehta, "Healthcare prediction using ML," 2021.
[10] P. Sharma et al., "Disease prediction using ML," *Procedia CS*, 2020.
[11] WHO, "Preventive Healthcare Report," 2020.
[12] D. Dinh et al., "Machine learning in healthcare survey," 2019.

Gaddipalli Amarnadh, pursuing Master of Data Science, Department of Computer Science, GSS, GITAM (Deemed to be University), Visakhapatnam. His area of interest in Machine Learning.

Dr Vanitha Kakollu is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University). His main areas of research include Image Processing, Data Mining and Machine Learning.