IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Heart Disease Prediction Models through Data-driven Feature Selection: A Mafia-K-means Fusion Approach

Peter Dassu, DR. K.M. Abubakkar sithik,
Student, Assistant Professor,
Computer Science,
DMI st. Eugene, Chibombo, Zambia

Abstract : Heart disease remains a significant global health concern, necessitating accurate predictive models for timely diagnosis and intervention. In this paper, we propose a novel approach, merging Mafia and K-means algorithms, aimed at improving feature selection for predictive models. Leveraging the robustness of Mafia algorithms and the clustering efficacy of K-means, we iteratively select informative features from complex datasets related to heart health. Our fusion technique demonstrates superior predictive performance compared to traditional methods, showcasing enhanced accuracy, precision, and recall rates. The results underscore the potential of this approach in refining heart disease prediction models, offering a promising avenue for future research in medical data analysis and predictive

Introduction

Global healthcare is very much concerned with developing accurate predictive models for heart disease that aim to identify those at risk beforehand and allow for preventive interventions. Heart disease, consisting of a range of cardiovascular conditions, remains the top cause of death globally. Therefore, multifaceted nature of heart disease and intricate interplay of the risk factors call for specific predictive models suitable for different populations. In response to this pressing need, research " Enhancing Heart Disease Prediction Models through Data-driven Feature Selection: A Mafia-K-means Fusion Approach " takes a path that can be considered innovative since it brings together complex algorithms with focused clustering techniques towards making predictive models more effective.

The World Health Organization (WHO) underscores the staggering impact of cardiovascular diseases (CVDs) on global health, attributing an estimated 17.9 million deaths annually to these conditions, representing nearly a third of all global deaths. This profound global burden of CVDs encompasses diverse demographics, geographies, and socioeconomic strata, underscoring the complexity of mitigating its effects. Within the specific purview of Malawi, a country facing unique healthcare challenges rooted in limited resources, infrastructure, and a distinct disease profile, cardiovascular ailments emerge as a substantial contributor to the disease burden and mortality rates.

However, tailoring accurate predictive models to address the nuances of heart disease in such settings remains a formidable task. Challenges arise from data scarcity, variability in risk factors prevalent within specific populations, and limitations in existing methodologies to discern pertinent features indicative of cardiac risk. Hence, the confluence of data-driven strategies, computational intelligence, and an in-depth understanding of the local disease landscape becomes pivotal in augmenting the precision and applicability of predictive models.

The proposed research explores uncharted territory, harnessing the power of Mafia algorithms that excel in influential factors obtained in complex data sets, and combines cooperation with K-means approach of the strength of clusters. This fusion approach not only promises to overcome the limitations of traditional feature selection methods, but also seeks to improve the predictive accuracy of specialized cardiovascular models

NEED OF THE STUDY.

Heart disease is a prevalent and life-threatening condition globally. Timely and accurate prediction of heart disease risk is crucial for preventive healthcare measures. Traditional methods often rely on expert knowledge or limited sets of symptoms, potentially missing valuable predictive indicators. This project aims to revolutionize heart disease prediction by leveraging advanced data-driven techniques, specifically the fusion of Mafia and K-means algorithms for feature selection, within a sophisticated web-based system.

Existing heart disease prediction models may lack efficiency and accuracy due to either a limited scope of features or inadequate selection methods. The challenge lies in identifying the most relevant features from extensive medical datasets containing a plethora of symptoms, biomarkers, and patient information. This project addresses this issue by proposing a novel approach that harnesses the power of Mafia and K-means algorithms for optimized feature selection.

The project adopts a two-fold strategy: the integration of Mafia algorithms, known for their ability to uncover relationships between variables, and K-means clustering, renowned for partitioning data points into coherent groups. The fusion of these techniques allows for a comprehensive exploration of the dataset, identifying influential features for heart disease prediction. The resulting feature subset will be used to train and enhance predictive models

3.1Population and Sample

The World Health Organization (WHO) underscores the staggering impact of cardiovascular diseases (CVDs) on global health, attributing an estimated 17.9 million deaths annually to these conditions, representing nearly a third of all global deaths. This profound global burden of CVDs encompasses diverse demographics, geographies, and socioeconomic strata, underscoring the complexity of mitigating its effects. Within the specific purview of Malawi, a country facing unique healthcare challenges rooted in limited resources, infrastructure, and a distinct disease profile, cardiovascular ailments emerge as a substantial contributor to the disease burden and mortality rates.

3.2 Data and Sources of Data

For The methodology begins with the comprehensive acquisition of heterogeneous medical datasets encompassing patient demographics, clinical history, symptoms, laboratory tests, and confirmed heart disease cases. This diverse data compilation is subjected to rigorous preprocessing involving data cleaning, normalization, feature scaling, and handling missing values to ensure consistency and quality. The stage involves the meticulous acquisition and preprocessing of diverse medical datasets, a crucial step to ensure the integrity and reliability of subsequent analyses.

3.3 Theoretical framework

This project addresses this issue by proposing a novel approach that harnesses the power of Mafia and K-means algorithms for optimized feature selection.

The project adopts a two-fold strategy: the integration of Mafia algorithms, known for their ability to uncover relationships between variables, and K-means clustering, renowned for partitioning data points into coherent groups. The fusion of these techniques allows for a comprehensive exploration of the dataset, identifying influential features for heart disease prediction. The resulting feature subset will be used to train and enhance predictive models.

RESEARCH METHODOLOGY

Data Collection and Preprocessing

The methodology begins with the comprehensive acquisition of heterogeneous medical datasets encompassing patient demographics, clinical history, symptoms, laboratory tests, and confirmed heart disease cases. This diverse data compilation is subjected to rigorous preprocessing involving data cleaning, normalization, feature scaling, and handling missing values to ensure consistency and quality. The stage involves the meticulous acquisition and preprocessing of diverse medical datasets, a crucial step to ensure the integrity and reliability of subsequent analyses.

Data Acquisition

The data collection process is initiated by gathering heterogeneous medical datasets from various sources, including hospitals, clinics, research repositories, and healthcare databases. These datasets encapsulate a wide array of information, such as:

- **Patient Demographics:** Including age, gender, ethnicity, and socioeconomic status.
- **Clinical History:** Chronic conditions, previous medical interventions, and family medical history.
- **Symptoms:** Comprehensive records of reported symptoms and their temporal patterns.
- **Laboratory Tests:** Results from diagnostic tests, blood work, and imaging studies.
- **Confirmed Heart Disease Cases:** Patient records indicating the presence or absence of heart disease.

This inclusive data compilation ensures a holistic representation of factors contributing to heart disease risk.

Data Preprocessing

The acquired datasets undergo a rigorous preprocessing phase to address challenges associated with real-world healthcare data:

- **Data Cleaning:** Identify and handle outliers, errors, and inconsistencies within the datasets to mitigate the impact of noisy data on subsequent analyses.
- **Normalization:** Standardize numerical features to a common scale, preventing the dominance of certain features due to their larger magnitudes.
- **Feature Scaling:** Normalize features to a standardized range, ensuring that no particular feature disproportionately influences subsequent analyses.
- **Handling Missing Values:** Implement strategies such as imputation or removal of missing data points to maintain dataset completeness.

Exploration through Mafia Algorithm

In this phase, the method uses sophisticated Mafia algorithms, notably Apriori and FP-Growth, to deepen the mining of association rules in a broad set of medical data These algorithms are powerful tools for revealing complex and multivariate associations, and shed light on potential determinants of cardiovascular disease.

Apriori Algorithm

• **Association rule mining:** Apriori algorithm is a classical method for identifying common structures in transactional databases. It typically operates through associations between co-occurrences.

• **Identifying common features**: Through medical databases, Apriori identifies common symptoms, risk factors, and other characteristics that occur frequently at the same time. For example, it may reveal that certain symptoms or risk factors are frequently observed together in patients diagnosed with cardiovascular disease.

• **Auxiliary reliability measure**: The algorithm computes the auxiliary reliability measure of the observed association, which reflects the frequency of occurrence and the strength of the relationship between the factors.

FP-Growth Algorithm

• **Efficient mining of frequency patterns**: FP Growth is a new algorithm designed for mining frequent objects efficiently without explicit mining of target objects, making it particularly useful for data sets with many properties

• **Creating an FP-Tree**: FP-Growth is an FP-Tree, a data structure representing the most frequent patterns in the dataset. This tree compresses the data set to a more manageable size and makes it easier to remove frequent items.

• **Pattern generation:** When pruning the FP tree, FP grows for more frequently arranged features, revealing important associations and patterns related to cardiovascular disease.

Uncovering Relationships and Dependencies

- **Identification of Association Rules:** Both Apriori and FP-Growth algorithms reveal association rules, elucidating dependencies and correlations between symptoms, risk factors, and occurrences of heart disease.
- **Intricate Relationships:** These association rules highlight intricate relationships, such as co-occurring symptoms or combinations of risk factors, providing valuable insights into potential predictors or indicators of heart disease.

Strategic Feature Selection via K-means Clustering

In this step, the method combines K-means clustering, a powerful unsupervised learning algorithm, to divide the dataset into cohesive groups based on feature similarity This strategic feature selection process aims to identify and exclude features clusters showing significant association with cardiomyopathy

Unsupervised Clustering with K-means Algorithm

• **Feature similarity evaluation**: K-means clustering algorithm divides the dataset into 'K' clusters, where 'K' is pre-defined based on domain knowledge or determined by optimization methods.

• **Iterative process:** K-means works iteratively to minimize the sum of squared distances between data points and their cluster center points. The categories are divided into groups based on the area of focus of each group.

Identification of Feature Clusters Associated with Heart Disease

- **Group Definition**: Each resulting group contains objects that exhibit similarities in their patterns or properties. These groups may have symptoms, biomarkers, or other characteristics associated with cardiovascular events.
- **Correlation analysis**: Characteristics within different groups are examined for significant associations or associations with cardiovascular events. Statistical methods or correlation coefficients can predict the strength of these associations.

*. Selection of Pertinent Features**

- **Earmarking Relevant Clusters:** Clusters demonstrating substantial correlations with heart disease or containing features of clinical significance are earmarked for further analysis and inclusion in subsequent stages of predictive modeling.
- **Feature Subset Identification:** Features within these earmarked clusters serve as a targeted subset for subsequent analyses, contributing to the development of predictive models aimed at accurate heart disease risk assessment.

. Significance for Predictive Modeling

- **Reduced Dimensionality:** Strategic feature selection via K-means clustering aids in reducing the dimensionality of the dataset by identifying clusters of relevant features. This reduction simplifies subsequent analyses and enhances computational efficiency.
- **Identification of Predictive Features:** The identified clusters and their constituent features serve as a pivotal foundation for subsequent model development. These features hold potential predictive power for accurately assessing heart disease risk.

The strategic selection of features through K-means clustering represents a critical phase in the methodology, providing a focused subset of features that exhibit significant correlations with heart disease. These selected features will inform the subsequent stages of model development, enhancing the accuracy and relevance of predictive models for heart disease risk assessment.

Fusion of Algorithms for Feature Subset Extraction

The fusion step in this approach is an important step that combines insights from Mafia and K-means algorithms. By combining the associations revealed by the Mafia algorithm with the cumulative relationships discovered through K, this fusion process aims to combine the strengths of both methods, concluding on a refined feature subset important for accurate prediction of cardiovascular disease

Integration of Association Insights from Mafia Algorithms

- **Association Rule Integration:** Insights derived from Mafia algorithms, such as Apriori or FP-Growth, unveil associations and dependencies among various features within the dataset.
- **Association Strength:** These association rules encapsulate the strength and significance of relationships between symptoms, risk factors, and occurrences of heart disease, providing valuable insights into potential predictors.

Synthesizing Clustered Relationships from K-means

- **Clustered Feature Relationships:** K-means clustering identifies groups of features exhibiting similarity and potential correlations related to heart disease.
- **Cluster Significance:** Features clustered together demonstrate associations and patterns that collectively contribute to the understanding of heart disease indicators.

*Comprehensive Feature Subset Extraction**

- **Synergistic Fusion:** The fusion process combines the associations revealed by the Mafia algorithms with the clustered relationships identified through K-means, creating a comprehensive and refined feature subset.
- **Discriminative Features:** This fusion synthesizes the strengths of both algorithms, extracting discriminative features that encompass intricate associations, dependencies, and clustered relationships crucial for heart disease prediction.

Enhanced Predictive Power

- **Robust Predictive Features:** The resultant feature subset, enriched through the fusion of Mafia and K-means outputs, possesses enhanced predictive power, encapsulating a comprehensive set of discriminative features crucial for heart disease assessment.
- **Informative Insights:** This refined feature subset serves as a foundation for subsequent stages of model development, contributing significantly to the accuracy and robustness of predictive models for heart disease risk assessment.

The fusion of insights obtained from Mafia and K-means algorithms represents a synergistic approach, leveraging the strengths of both techniques to extract a refined feature subset. This comprehensive feature selection process lays the groundwork for the subsequent development of predictive models aimed at accurate heart disease prediction.

Implementation of Web-based System

The culmination of this methodology lies in the development of a sophisticated web-based platform that integrates the trained predictive models derived from the refined feature subset. This intuitive interface empowers healthcare professionals and individuals to access real-time heart disease risk assessments by conveniently inputting pertinent data

System Architecture Design

- **User Interface Planning:** Designing an intuitive and user-friendly interface that facilitates seamless data input and interpretation of results.
- **Backend Infrastructure:** Developing a robust backend system to handle data processing, model computations, and result generation.

Integration of Trained Predictive Models

- **Model Integration:** Implementing the trained predictive models, developed using the refined feature subset, into the web-based system's architecture.
- **Real-time Processing:** Configuring the system to perform real-time computations based on user inputs, allowing immediate heart disease risk assessments.

User Accessibility and Security Measures

- **User Access Control:** Implementing secure user authentication protocols to ensure restricted access only to authorized healthcare professionals and individuals.
- **Data Security:** Adhering to stringent data security measures to safeguard sensitive patient information, employing encryption and secure transmission protocols.

User Interface Development

- **Input Mechanism:** Creating an intuitive interface for users to input relevant patient data, including demographics, symptoms, and medical history.
- **Result Display:** Designing an output display mechanism that provides easily interpretable heart disease risk assessments based on the predictive models' analysis of the input data.

IV. RESULTS AND DISCUSSION

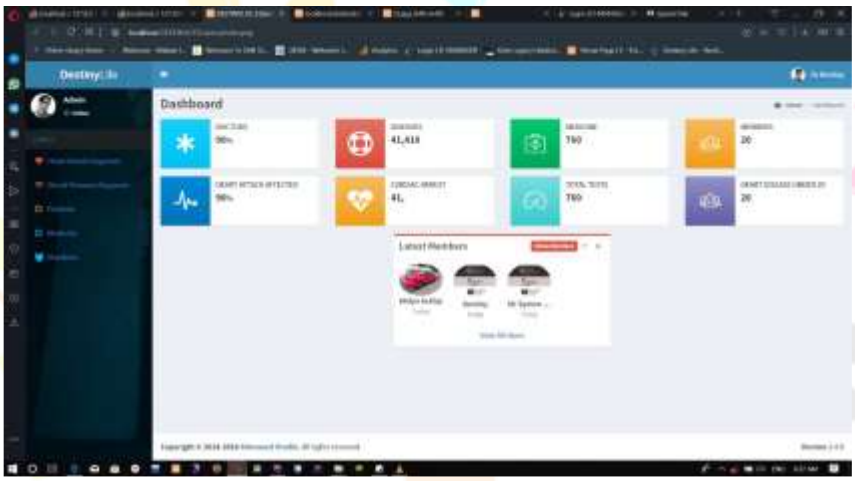
4.1 Results of Descriptive Statics of Study Variable

Authentication

Before the user accesses the system, they are required to login. This is where the user enters the log in details given by the system administrator either through email or written depending on the user. After the user fills the form, authorization and authentication occurs.

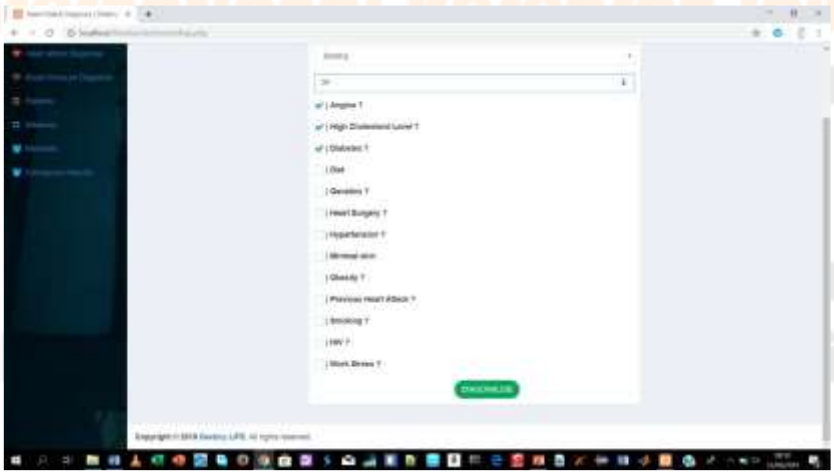
DASHBOARD

The dashboard shows interaction between the users of the system and the graphical user interaction which has been designed to be user friendly



DIAGNOSIS PAGE

THE DIAGNOSIS PAGE ALLOWS THE DOCTOR TO SELECT THE PARAMETERS SET AND THERE BY PERFORMING THE DIAGNOSIS.

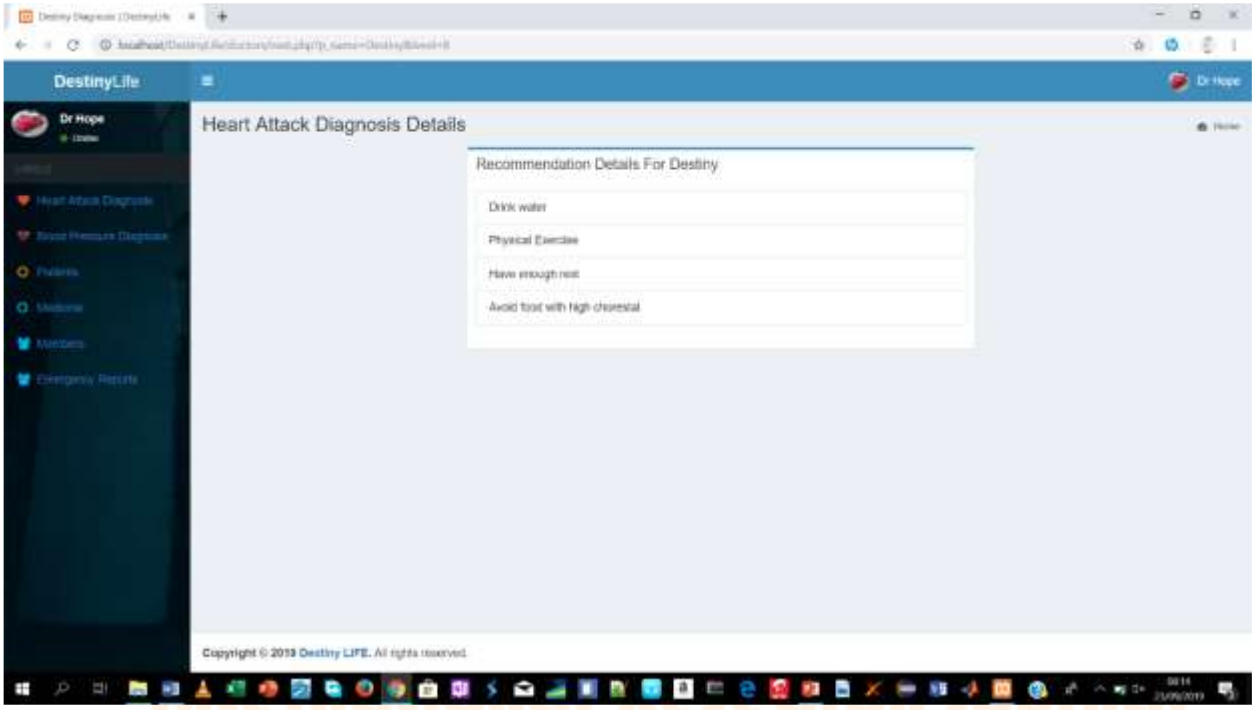


RESULTS PAGE

The results page shows the results of the diagnosis. It shows the results in the form of percentage level. The percentage level above 80 percentage depicts the higher probability that that the patient might have a heart disease

RECOMMENDATION

After the results page, the system is able to give recommendation to the patient based on the diagnosis results.



I. ACKNOWLEDGMENT

In the voyage of completing this project, my profound thanks go to the **ALMIGHTY GOD** for the guiding light that illuminated my path without hindrance.

I extend heartfelt appreciation to **Rev.Fr. Dr. J.E.ARUL RAJ, OMI**, and the distinguished DMI group of Institutions, Zambia. Their unwavering support has been the bedrock on which this study found its wings.

To **Dr. T.X.A. ANANTH,** President of the University Council, my gratitude knows no bounds for the manifold ways in which he supported this endeavor.

Special recognition is due to **Dr. IGNATIUS A. HERMAN**, the Director of Education, whose encouragement provided the wind beneath my wings.

In the intricate web of assistance and guidance, my thanks go to **Rev.Sr.FATHIMA MARY, Dr.R.KAVITHA, Dr.R.SAKTHIVEL, and Fr.T.AMALRAJ**, along with the entire department, for their invaluable contributions.

A distinctive acknowledgment is reserved for **DR.K.M Abubakkar Sithik**, whose guidance and support wove the threads of this project into a coherent whole.

A special acknowledgment is extended to **Mr. Shannawazi Dassu**, whose unwavering belief, support and encouragement were like rays of sunshine on the darkest days of this journey.

REFERENCES

1. Janet valade(2009). Php5 for dummies. Wiley Publishing, Inc.
2. David powers(2015). Php solutions.dynamic web design made easy
3. www.w3schools.com/css-doc/index.html -official documentation for css
4. www.w3shools.com /php- php tutorials
5. V. Manikantan and S. Latha ,"Predicting the analysis of heart disease symptoms using medicinal data mining methods",International Journal of Advanced Computer Theory and Engineering, vol. 2 ,pp.46-51,2013.
6. Shadab Adam Pattekari and Alma Parveen," Prediction system for heart disease using naïve bayes", International Journal of Advanced Computer and Mathematical Sciences, vol.3,pp 290-294,2012.
7. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive data mining for medical diagnosis: an overview of heart disease prediction" International Journal of Computer Science and Engineering, vol. 3 ,2011
8. R. Agrawal,T Imielinski ,and A. Swami ,'Mining association rules between sets of items in large databases'
9. Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", International Conference on Computer Research and Development, ISBN: 978-1-61284-840-2,2011
10. M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", International Journal of Engineering Science and Technology vol.2, pp.5370- 5376,2010.
11. Douglas Burdick, Manuel Calimlim, Johanne Gehrke,"MAFIA: A Maximal Frequent Item set Algorithm For Transactional Databases", Proceedings of the 17th International Conference on Data Engineering.
12. K.Srinivas, Dr. G.Raghavendra Rao and Dr. A. Govardhan, "ASurvey On Prediction Of Heart Morbidity Using Data Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, No.3, May 2011
13. S.Vijayarani, M. Divya, " An Efficient Algorithm for
14. Generating Classification Rules", IJCST ,vol. 2, Issue 4, 2011

International Research Journal
IJNRD
Research Through Innovation