International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Prediction of disease in the HealthCare using Machine Learning

Mrs. Yerraginnela Shravani1, Dr Ashesh K2
1PhD-Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
2Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
shravanilokeshwar@gmail.com1, imasheshk@kluniversity.in2

Abstract
Heart disease is one of the complex diseases and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. In this article, we proposed an efficient and accurate system to diagnosis heart disease and the system is based on machine learning techniques. Predicting diseases in healthcare using machine learning often encounters imbalanced datasets where the number of instances of one class (e.g., diseased patients) is significantly lower than the other (e.g., non-diseased patients). Addressing imbalanced data is crucial as models trained on such datasets tend to favor the majority class, leading to biased predictions.in this project also doing imbalce data to balance data using Random Under and over Sampler. after conversion of data into balanced data.

Keywords: Heart disease, Diagnosis, Machine learning, Classification algorithms, Data imbalance, Random Under Sampler, Random Over Sampler, SMOTE, Balanced data, Healthcare, Cardiology.

I. INTRODUCTION

In the realm of modern healthcare, the accurate prediction and timely identification of diseases have become increasingly vital. Among these diseases, heart disease stands out as a complex and prevalent health issue, impacting a substantial portion of the global population. Given its widespread prevalence and the potential severity of its consequences, the need for effective tools to diagnose heart disease cannot be overstated, particularly within the field of cardiology.

This article introduces an innovative approach to address this pressing concern—a system that harnesses the power of machine learning techniques for the precise diagnosis of heart disease. The application of machine learning in healthcare has garnered considerable attention due to its potential to revolutionize medical diagnostics. In this context, our system represents a significant advancement towards more efficient and accurate disease prediction.

By employing these techniques, we aim to enhance the system's capacity to discern patterns and relationships within complex medical data, ultimately leading to more reliable diagnoses.

The consequences of imbalanced data in this context are significant. A model trained on imbalanced data may have a tendency to classify most instances as the majority class (e.g., no heart disease), resulting in a high accuracy score that mask poor performance in detecting cases of heart disease. Therefore, it becomes imperative to preprocess the data effectively to ensure that the model is sensitive to both classes. compare several strategies for balancing the dataset, including under sampling the majority class, oversampling the minority class using synthetic data generation techniques like SMOTE, and hybrid approaches. By achieving a balanced distribution of classes, we aim to improve the model's ability to generalize to both instances of heart disease presence and absence. the dataset and its characteristics, the challenges posed by imbalanced data, the methodologies employed for balancing the data, the experimental setup, presents and analyzes the results. Finally, future directions for refining heart disease prediction models in the presence of imbalanced data.

To tackle this issue, we implement the Random Under Sampler method, which helps balance the dataset by reducing the number of instances in the majority class. This approach ensures that our machine learning models are trained on a more equitable representation of both classes, leading to more equitable and accurate predictions.

After achieving a balanced dataset, we embark on the training phase, employing supervised machine learning algorithms. This stage is pivotal as it forms the backbone of our predictive system. We systematically compare the performance of various algorithms to identify the one that offers the highest accuracy. The chosen algorithm will serve as a foundation for future applications, providing healthcare professionals with a robust tool for the early and accurate detection of heart disease.

In this article delves into the development of an advanced system for disease prediction in healthcare using machine learning. With a primary focus on heart disease, we utilize a

combination of classification algorithms, feature selection methods, and data balancing techniques to create an efficient and precise diagnostic tool. Through rigorous evaluation and comparison of machine learning algorithms, our objective is to pave the way for improved disease prediction in the healthcare sector, ultimately benefiting patients and healthcare providers worldwide.

LITERARURE SURVY

- **Estes et al., 2018 [1]:** "Modeling NAFLD disease burden in multiple countries". This study focuses on modeling the disease burden of non-alcoholic fatty liver disease (NAFLD) in several countries over a specific period. It provides insights into the projected prevalence and impact of NAFLD, highlighting the need for effective prevention and management strategies.
- **Drożdż et al., 2022 [2]:** "Risk factors for cardiovascular disease in patients with MAFLD using a machine learning approach". This research employs machine learning to identify risk factors for cardiovascular disease in individuals with MAFLD. It emphasizes the importance of early risk assessment and intervention in such patients.
- **Murthy and Meenakshi, 2014 [3]:** "Dimensionality reduction for early prediction of coronary heart disease using a neuro-genetic approach". The study explores dimensionality reduction techniques for the early prediction of coronary heart disease. It utilizes a neuro-genetic approach to enhance predictive accuracy.
- **Benjamin et al., 2019 [4]:** "Heart disease and stroke statistics report from the American Heart Association". This report provides comprehensive statistics on heart disease and stroke, including prevalence, risk factors, and mortality rates. It serves as a valuable reference for understanding the epidemiology of cardiovascular diseases.
- **Shorewala, 2021 [5]:** "Early detection of coronary heart disease using ensemble techniques". The study investigates the use of ensemble machine learning techniques for the early detection of coronary heart disease. It highlights the potential of ensemble models in improving predictive accuracy.
- **Mozaffarian et al., 2015 [6]:** "Heart disease and stroke statistics report from the American Heart Association". Similar to reference [4], this report presents updated statistics on heart disease and stroke, providing a comprehensive overview of the prevalence and impact of these diseases.
- **Maiga et al., 2019 [7]:** "Comparison of machine learning models in the prediction of cardiovascular disease using health record data". The study compares various machine learning models for predicting cardiovascular disease based on health record data. It offers insights into the performance of different algorithms in disease prediction tasks.

PROBLEM STATEMENT

Heart disease is a widespread and life-threatening ailment that necessitates timely and precise diagnosis. Traditional cardiology diagnostics are labor-intensive, potentially error-prone, and struggle to unveil subtle patterns within extensive datasets. This research endeavour's to develop a machine learning-based system for accurate heart disease prediction, mitigating these challenges. Key issues encompass data imbalance, necessitating strategies like the Random Under Sampler, and the identification of relevant attributes through feature selection. The study aims to assess and compare diverse classification algorithms to determine the most accurate model. By doing so, it aspires to equip healthcare professionals with a powerful tool for early heart disease detection, thereby enhancing patient care and outcomes globally.

LIMITATIONS

- **Data Quality:** The accuracy of predictions heavily relies on the quality and comprehensiveness of the available medical data. Inaccurate or incomplete data can compromise the reliability of the predictive model.
- **Generalization:** The developed predictive model may exhibit limitations in its ability to generalize to diverse patient populations, as it relies on existing datasets, which may not fully represent all demographic and clinical variations.
- **Ethical Considerations:** The use of patient health data for research purposes raises ethical concerns regarding data privacy and consent. Adhering to ethical guidelines and obtaining necessary permissions is paramount.
- **Algorithm Bias:** Machine learning algorithms can inherit biases present in the training data, potentially leading to biased predictions. Efforts should be made to mitigate and monitor algorithmic biases.
- **Model Interpretability:** Some complex machine learning algorithms may lack transparency, making it challenging to interpret the reasons behind specific predictions, which can be a concern in clinical settings.
- **Resource Requirements:** Developing and implementing machine learning models may demand significant computational resources and expertise, which may not be readily available in all healthcare institutions.
- **Validation:** Ensuring the validity and robustness of the predictive model requires extensive validation on independent datasets and in real-world clinical settings, which may present logistical challenges.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

- **Dynamic Nature of Healthcare:** Healthcare practices and patient demographics evolve over time. The predictive model may require periodic updates to remain effective and relevant.

INPUT & OUTPUTS

Inputs:

- **Medical Data:** The primary input to the predictive system is a set of medical data for each patient. This data typically includes various features and diagnostic indicators, such as age, gender, blood pressure, cholesterol levels, electrocardiogram (ECG) readings, and other relevant clinical measurements.
- **Balanced Dataset:** To address data imbalance, a balanced dataset obtained through techniques like RandomUnderSampler serves as input. This dataset ensures equitable representation of both classes: patients with heart disease and those without.

Outputs:

- **Heart Disease Prediction:** The primary output of the system is a binary prediction for each patient, indicating whether they are likely to have heart disease (positive) or not (negative). This prediction is based on the analysis of the patient's medical data and the trained machine learning model.
- **Visualization:** In some implementations, the system may generate visualizations such as ROC curves, confusion matrices, and feature importance plots to facilitate interpretation and decision-making by healthcare professionals.
- **Clinical Recommendations:** In a clinical context, the system may provide recommendations or alerts to healthcare providers based on the predicted outcomes. For example, it may suggest further diagnostic tests or interventions for patients at high risk of heart disease.

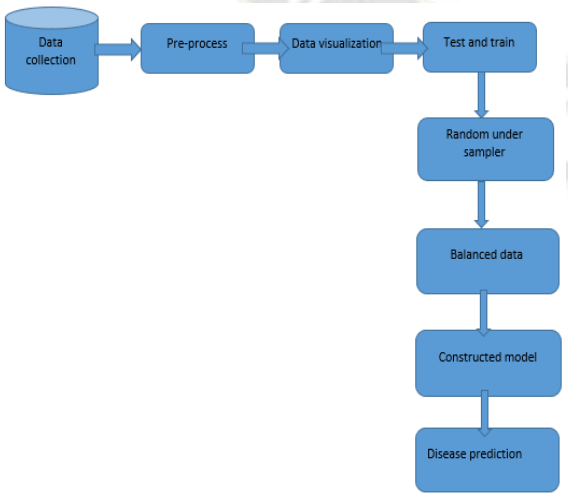

Fig 1: Proposed Block Diagram

II. METHODOLOGY

Remote mobile health monitoring has garnered significant attention as a potential solution for healthcare. However, it encounters numerous challenges at every stage, including data aggregation, maintenance, integration, analysis, and pattern interpretation, when dealing with the complexities of healthcare big data (HBD). The existing system also grapples with issues related to the complexity of data analysis and scalability within parallel computing models. Furthermore, it falls short in delivering accurate predictions for heart disease.

Proposed System

The proposed system aims to revolutionize heart disease prediction in healthcare through advanced data analytics and machine learning. Central to its design is the enhancement of diagnostic accuracy for heart conditions, leveraging the scalability of big healthcare data. This system stands out for its efficiency, significantly reducing the time typically required for data processing. It is meticulously engineered to ensure high-performance management and maintenance of heart disease-related data. A key innovation in this system is the integration of the Random Under Sampler technique, which adeptly balances datasets. This is particularly crucial in healthcare, where data often exhibits imbalance, with fewer instances of diseased patients compared to healthy ones. By rectifying this imbalance, the system enhances the reliability and precision of predictive models. Overall, this approach is not just about harnessing technology for data analysis; it's a stride towards more accurate, timely, and efficient heart disease diagnosis, ultimately aiming to improve patient care and outcomes.

Here's a detailed description of the block diagram:

- **Data Collection**: This stage involves collecting relevant medical data such as patient demographics, vital signs, lab results, medical history, and diagnostic tests.
- **Data Preprocessing**: In this step, data is cleaned by handling missing values, normalizing or scaling numerical features, encoding categorical variables, and removing outliers to ensure data quality and prepare it for model training.
- **Feature Selection/Engineering**: This involves selecting important features that significantly contribute to the prediction of heart disease and may include creating new features or transforming existing ones to enhance model performance.
- **Splitting the Data**: The dataset is divided into training, validation, and test sets, with the training set used for model training, the validation set for tuning

hyperparameters, and the test set for evaluating the model's performance.

- **Convert Imbalanced Data to Balanced Data (Using Random Under Sampling Technique)**: This crucial step uses Random Under Sampling to balance the dataset, addressing the issue of class imbalance common in medical datasets.
- **Model Selection**: Appropriate machine learning algorithms for classification tasks are chosen, such as Decision Trees, Random Forests, Support Vector Machines, and Naïve Bayes.
- **Model Training**: The selected models are trained on the training dataset using the chosen algorithms and hyperparameters, allowing the model to learn patterns and relationships within the data.
- **Model Evaluation**: The models are assessed using the validation set to choose the best-performing model.
- **Model Validation**: The final model is validated on the test set to assess its real-world performance and generalization to unseen data, confirming its reliability and effectiveness.

Challenges:

- **Data Imbalance:** The dataset used for heart disease prediction typically exhibits a significant class imbalance, where the number of patients with heart disease is substantially smaller than those without. This imbalance can lead to biased models that prioritize the majority class, affecting the system's overall accuracy.
- **Feature Selection:** Medical datasets often contain numerous variables, some of which may be irrelevant or redundant for heart disease prediction. Effective feature selection methods must be employed to identify the most informative attributes while discarding irrelevant ones.
- **Algorithm Selection:** The choice of machine learning algorithms is critical in achieving high predictive accuracy. The research needs to identify and compare various classification algorithms to determine the most suitable ones for heart disease prediction.
- **Data Preprocessing:** Proper data preprocessing techniques, such as normalization and handling missing values, are essential to ensure the quality and consistency of the dataset.

III. RESULTS & DISCUSSION

The study on predicting heart disease using machine learning presented in "below figures" encompasses a comprehensive approach, integrating various data analysis and visualization techniques. Initial exploratory data analysis is conducted on key variables such as the target outcome (heart disease presence), sex, and age, providing foundational insights into the dataset's demographic and health-related attributes. Notably, the study emphasizes the significance of cholesterol and blood pressure (trestbps) in relation to heart disease, highlighted through targeted visualizations like stripplots and scatter plots. Additionally, a meticulous examination of glucose levels against blood pressure is undertaken, underscoring their critical roles in heart health.

Histograms and Kernel Density Estimates are employed for detailed distribution analysis of specific attributes, such as 'oldpeak', enhancing the understanding of variable characteristics within the dataset. The issue of data imbalance, a common obstacle in machine learning applications, is addressed through visual representations like pie charts, illustrating the dataset's composition both before and after implementing balancing techniques. This methodical approach to balancing the data ensures the development of a more accurate and unbiased machine learning model. The study culminates with a correlation heatmap, elucidating the interdependencies between various health indicators, thereby aiding in the construction of a robust predictive model for heart disease using techniques like Support Vector Machine, Naïve Bayes, Random Forest Classifier, and Decision Tree.

Execution steps:

The execution steps for predicting disease in healthcare using machine learning, as detailed in your document, are as follows:

Data Collection: Collect relevant medical data, including patient demographics, vital signs, lab results, medical history, and diagnostic tests.

Feature Selection/Engineering: Select important features that significantly contribute to the prediction of heart disease. This may also involve creating new features or transforming existing ones to enhance model performance.

Data Preprocessing: Clean the data by handling missing values, normalizing or scaling numerical features, and encoding categorical features.

Splitting the Data: Divide the dataset into training, validation, and test sets. The training set is used for model training, the validation set for tuning hyperparameters, and the test set for evaluating the model's performance.

Converting Imbalanced Data to Balanced Data: Use the Random Under Sampling technique to balance the dataset, addressing the issue of class imbalance common in medical datasets.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

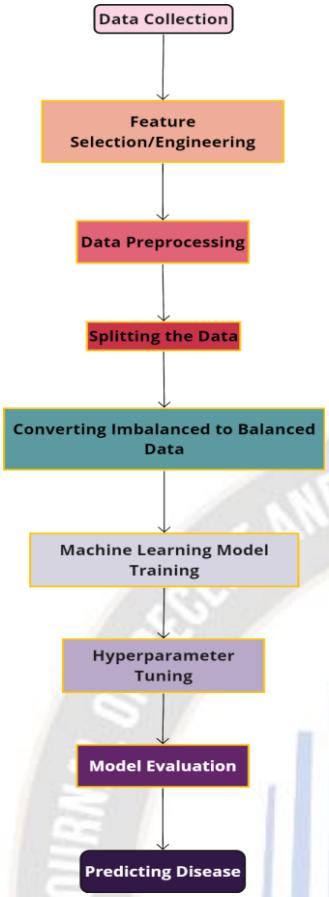

Fig 2: The execution steps for predicting disease in healthcare using machine learning

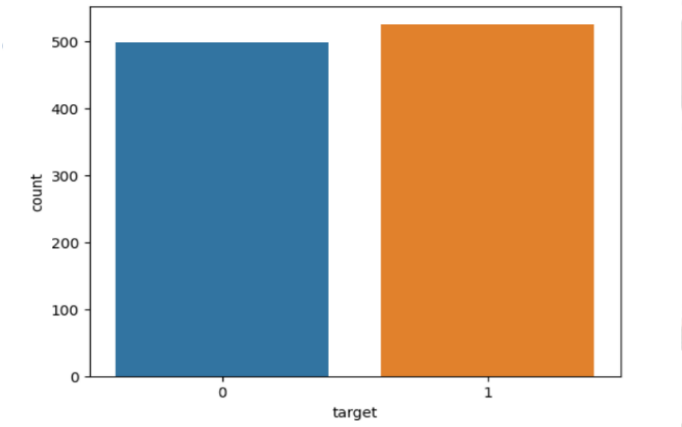

Fig 3: Exploratory Data Analysis for target (0,1-no. of person not effected and effected)

The "Exploratory Data Analysis for target" is likely a visualization that demonstrates the distribution of data points for the target variable in your study on heart disease prediction using machine learning. This target variable represents the key outcome that the study aims to predict – in this case, the presence or absence of heart disease.

In such a figure, you would typically see a graphical representation, such as a bar graph or pie chart, showing the proportion of cases in the dataset that are diagnosed with heart disease versus those that are not. This kind of analysis is crucial for understanding the baseline characteristics of your dataset. It provides insight into the balance or imbalance between the two classes (diseased vs. non-diseased), which is essential for guiding the choice of machine learning models and for interpreting their performance accurately.

If the data is highly imbalanced (for instance, if a large majority of the cases are non-diseased), it might necessitate special techniques in machine learning modeling, such as oversampling the minority class or applying different metrics for evaluating model performance, to ensure that the predictive model does not become biased towards the majority class.

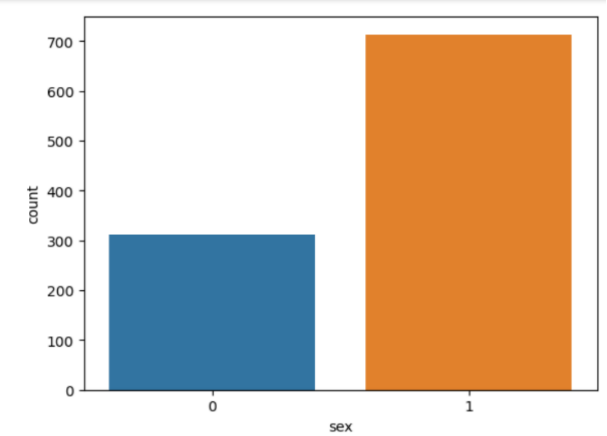

Fig 4: Exploratory Data Analysis for sex (0,1-male and female)

The "Exploratory Data Analysis for sex" is likely presents a visualization of how heart disease distribution varies between different genders. This type of analysis is crucial in understanding the impact of gender on heart disease prevalence and risk factors.

In such a visualization, you might typically see a bar graph, pie chart, or other graphical representation that displays the number or percentage of male and female subjects in the study, along with how many of them have been diagnosed with heart disease. This allows for a clear comparison of the prevalence of heart disease between genders.

Analyzing heart disease data by sex is important for several reasons:

- **Biological Differences**: There are known biological and physiological differences between males and females that can influence heart disease risk and symptoms. For instance, males and females may exhibit different symptoms or have different risk factors.
- **Tailored Healthcare**: Understanding these differences can help in developing gender-specific strategies for prevention, diagnosis, and treatment of heart disease.
- **Research Insights**: It provides valuable insights for further research, as certain risk factors or treatment

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

modalities may be more effective for one gender over the other.

- **Public Health Policies**: This data can inform public health policies and educational campaigns tailored to the specific needs of each gender.

Overall, this aspect of the exploratory data analysis helps in painting a comprehensive picture of the dataset and aids in developing more nuanced and effective machine learning models for predicting heart disease.

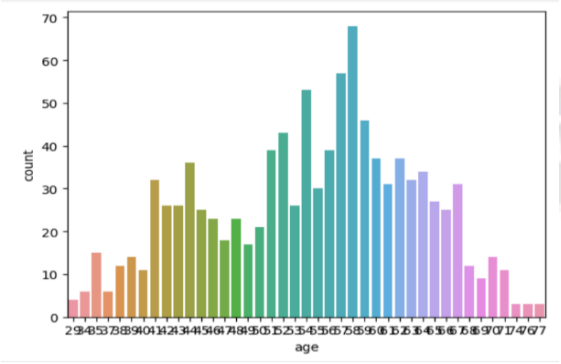

Fig 5: Graphical representation for age

The "Graphical representation for age" in your document is likely a visualization that illustrates the age distribution of the study's participants. This is a fundamental aspect of exploratory data analysis in medical research, especially in studies related to conditions like heart disease, where age is a significant risk factor.

In such a graph, you would typically see:

- **Age Distribution**: A histogram or bar graph displaying the number of participants in various age groups. This helps in understanding the age composition of the study population.
- **Age-Related Risk Analysis**: It might show the prevalence of heart disease across different age groups, highlighting any trends or patterns. For example, an increase in heart disease cases in older age groups would be expected.
- **Identification of Key Age Groups**: The graph can help identify age ranges that are more susceptible to heart disease, which is crucial for targeted prevention and treatment strategies.
- **Comparison with General Population**: The age distribution can also be compared with that of the general population to understand if the sample is representative, especially in terms of age.
- **Guidance for Machine Learning Models**: Understanding the age distribution is essential for developing and fine-tuning machine learning models. It informs the model about the significance of age as a feature in predicting heart disease.

Overall, this graphical representation provides valuable insights into how age factors into the risk and prevalence of heart disease, guiding both medical research and the development of predictive models.

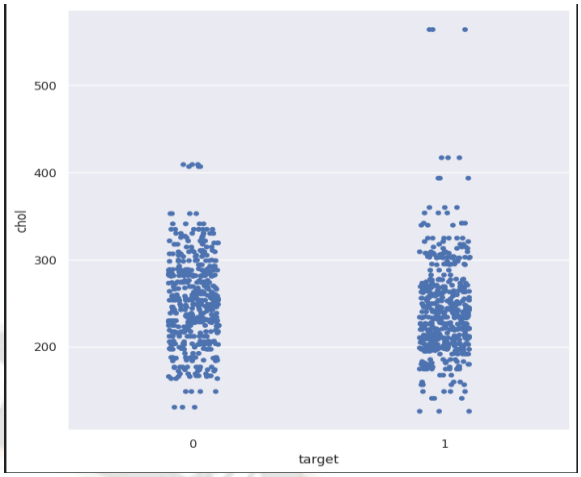

Fig 6: Stripplot for target and cholestral

The "Stripplot for target and cholesterol" is likely demonstrates the relationship between cholesterol levels and the presence of heart disease, which is the target variable in your study. A strip plot is an effective way to visualize the distribution of a continuous variable (like cholesterol levels) across different categories (in this case, the presence or absence of heart disease).

In a strip plot:

- **Cholesterol Levels**: The horizontal or vertical axis (depending on the plot orientation) represents cholesterol levels. Each point on the plot corresponds to an individual's cholesterol measurement.
- **Heart Disease Presence**: The other axis categorizes individuals based on whether they have heart disease or not. This binary classification allows for a clear visual comparison.
- **Data Distribution and Overlap**: The plot will scatter individual data points, providing a sense of how cholesterol levels are distributed within each category of the target variable. It's particularly useful for identifying any overlap or distinct clusters between groups.
- **Outliers and Trends**: You can identify any outliers or unusual observations in cholesterol levels among those with and without heart disease. It also helps in observing any apparent trends, such as higher cholesterol levels being more common in individuals with heart disease.
- **Informing Machine Learning Models**: This visual representation can inform the feature selection and model development in your machine learning analysis. For instance, if a clear pattern or trend is observed, cholesterol levels might be a significant predictive feature for heart disease.

Overall, the strip plot is a valuable tool in exploratory data analysis, offering clear insights into how a key risk factor like

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

cholesterol correlates with the likelihood of heart disease, thereby guiding further analysis and model development.

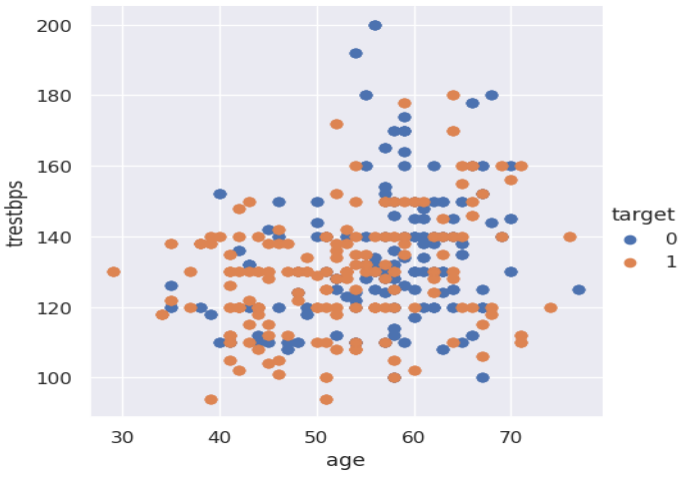

Fig 7: Implot for target and trest bps

The "Implot for target and trestbps" is likely showcases a scatter plot or similar visualization, illustrating the relationship between trestbps (which likely stands for 'resting blood pressure') and the target variable (presence or absence of heart disease). This kind of visualization is crucial in understanding how a specific medical measurement, like resting blood pressure, correlates with the risk of developing heart disease.

In an Implot (which might be a specific form of interactive or integrated plot):

- **Resting Blood Pressure (trestbps)**: This variable is usually plotted along one axis (either the horizontal or vertical axis), representing the range of resting blood pressure measurements observed in the study participants.
- **Heart Disease Presence**: The other axis typically categorizes individuals based on the target variable, which in this case is the presence or absence of heart disease.
- **Data Points Representation**: Each point on the plot corresponds to an individual participant, with their position determined by their resting blood pressure and heart disease status. This allows for a visual assessment of how these two variables are related.
- **Trend Identification**: You can identify patterns or trends, such as whether higher resting blood pressure is associated with a higher incidence of heart disease.
- **Outliers and Clusters**: The plot may reveal outliers (individual cases that stand apart from the general trend) or clusters (groups of data points that share similar characteristics).
- **Guidance for Machine Learning Analysis**: Insights from this plot are valuable for machine learning modeling, as they inform about the potential predictive power of resting blood pressure in determining heart disease risk.

Overall, the Implot for target and trestbps is a critical tool in exploratory data analysis, helping to uncover the relationship between a key health indicator (resting blood pressure) and the risk of heart disease, thereby guiding the development of predictive algorithms.

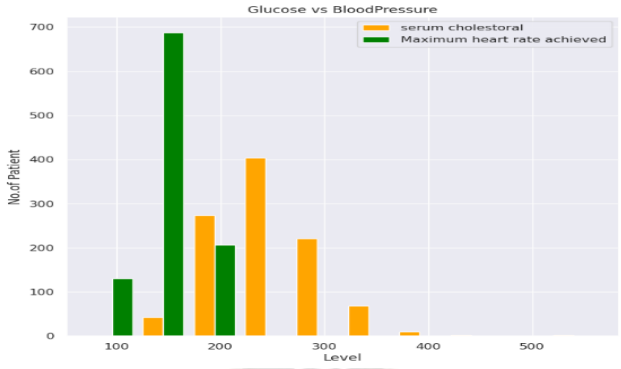

Fig 8: Glucose VS Blood Pressure

The "Glucose VS Blood Pressure" is likely a comparison between glucose levels and blood pressure, two crucial factors in heart health. This comparison is essential in understanding how these two health indicators interact and correlate with each other, and their combined impact on heart disease risk.

In such a visualization, you might typically expect:

- **Bivariate Analysis**: The graph would typically plot blood pressure on one axis (either systolic or diastolic) and glucose levels on the other. Each point on the graph represents an individual's measurements for both variables.
- **Pattern Identification**: This type of visualization helps in identifying patterns or correlations between glucose levels and blood pressure. For example, it might show whether higher glucose levels are associated with higher blood pressure.
- **Clusters and Trends**: The graph can reveal clusters of data points indicating common combinations of glucose and blood pressure values. It can also show trends, such as a linear or non-linear relationship between these variables.
- **Risk Factor Analysis**: Since both high glucose levels (indicative of potential diabetes) and high blood pressure are risk factors for heart disease, this comparison is crucial in identifying individuals who might be at higher risk.
- **Informing Predictive Models**: Insights gained from this comparison are invaluable for predictive modeling in healthcare. They can inform the development of machine learning models by highlighting the importance of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 10 September 2023 **Revised:** 20 October 2023 **Accepted:** 30 October 2023*

considering the interaction between different health indicators.

Overall, the comparison between glucose levels and blood pressure provides a comprehensive view of individual health statuses and helps in understanding the complex interplay of different risk factors for heart disease.

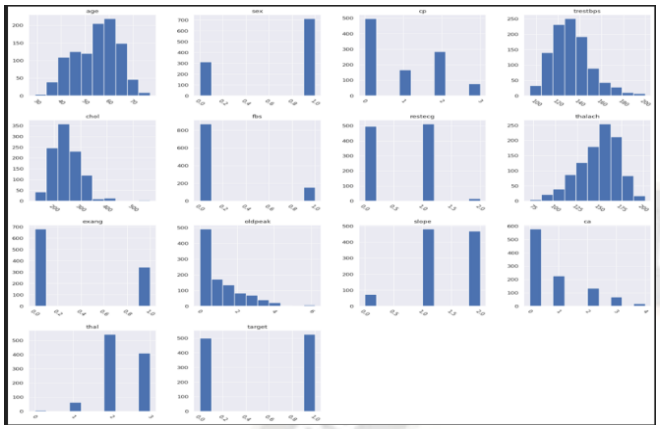

Fig 9: Histogram plotting for attributes.

The "Histogram plotting for attributes" is likely displays the distribution of various attributes or features that are used in the machine learning model for predicting heart disease. Histograms are a fundamental tool in exploratory data analysis, providing a visual representation of the distribution of numerical data.

In such histograms, you can expect:

- **Distribution of Individual Attributes**: Each histogram represents the distribution of a single attribute, such as age, cholesterol levels, blood pressure, glucose levels, etc. It shows how many participants fall into different ranges or bins of that attribute.
- **Identification of Skewness or Normality**: Histograms can reveal whether the data for an attribute is normally distributed, skewed to the left or right, or follows another type of distribution. This is important for understanding the underlying characteristics of the data and for selecting appropriate statistical methods and machine learning algorithms.
- **Detection of Outliers**: The shape and spread of the histogram can help in identifying outliers - data points that are significantly higher or lower than the rest of the data. Outliers can significantly impact the performance of machine learning models.
- **Insights into Data Range and Variability**: Histograms provide insights into the range and variability of each attribute, showing how spread out the data points are.
- **Guiding Data Preprocessing**: Understanding the distribution of each attribute is critical for data preprocessing steps such as normalization, standardization, or handling missing values, which are essential for effective machine learning modeling.
- **Multivariate Analysis Preparation**: By analyzing the distribution of each variable individually, you can better prepare for multivariate analysis, where the interactions between different attributes are considered.

Overall, histogram plotting for attributes is an essential step in preparing and understanding your dataset, ensuring that the machine learning models you develop are based on well-understood and appropriately processed data.

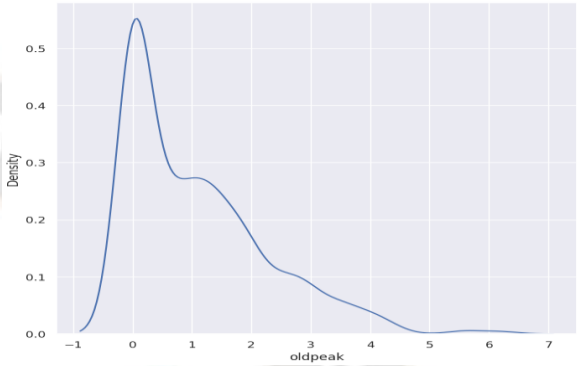

Fig 10: Kernel Density Estimate for oldpeak

The "Kernel Density Estimate for oldpeak" represents a density plot for the 'oldpeak' variable, which is likely a specific measurement or attribute within your dataset, particularly in the context of heart disease research. A Kernel Density Estimate (KDE) is a way to estimate the probability density function of a continuous random variable.

In a KDE for oldpeak, you can expect:

- **Continuous Probability Distribution**: Unlike histograms, a KDE provides a smooth curve representing the data's distribution. This can give a more accurate picture of the underlying probability density function of the 'oldpeak' variable.
- **Understanding 'oldpeak' Variable**: The 'oldpeak' typically refers to the ST depression induced by exercise relative to rest, a measurement used in stress tests for heart disease. The KDE plot will show how this measurement is distributed among your study participants.
- **Identification of Skewness and Modes**: The plot can reveal whether the data is skewed in any particular direction or if it has multiple modes (peaks). Multiple peaks might suggest different subgroups within the population.
- **Insights into Data Characteristics**: It provides insights into the central tendency, dispersion, and tail behavior of the 'oldpeak' data, which are important for understanding its role in the context of heart disease.
- **Guidance for Machine Learning Models**: The distribution of 'oldpeak' as seen in the KDE can influence how this variable is treated in machine learning models.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

For example, if the data is highly skewed, it might require transformation before being used in the model.

- **Comparison with Other Variables**: In some cases, KDE plots are used in conjunction with other variables to understand relationships and correlations.

Overall, the Kernel Density Estimate for oldpeak is a valuable tool in your exploratory data analysis, providing a nuanced view of this specific measurement and its implications in the context of predicting heart disease.

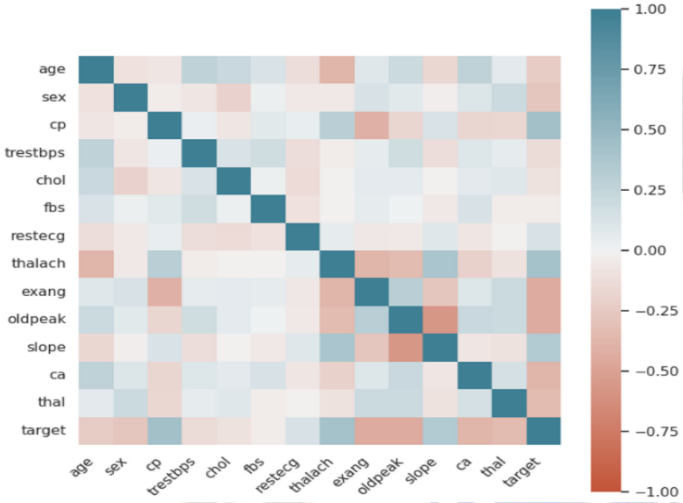

Fig 11: HeatMap for Correlation between attributes

The "HeatMap for Correlation between attributes" is likely a visualization that illustrates the correlation between various attributes in your dataset, which is crucial for understanding the relationships between different health indicators in the context of heart disease prediction.

In a heatmap for correlation, you can expect:

- **Matrix of Correlations**: The heatmap typically displays a matrix where each cell represents the correlation coefficient between two attributes. The attributes are listed along both the horizontal and vertical axes.
- **Color-Coding**: The strength and direction of the correlation are indicated by colors. For instance, a strong positive correlation might be shown in one color (like a deep red), a strong negative correlation in another (like deep blue), and no correlation in a neutral color (like white or light gray).
- **Identification of Strong Relationships**: This visualization is particularly useful for quickly identifying which pairs of attributes have strong positive or negative correlations. This can highlight potential relationships, such as between cholesterol levels and heart disease, or age and blood pressure.
- **Guidance for Feature Selection**: Understanding correlations is crucial for feature selection in machine learning models. Highly correlated features can lead to multicollinearity, which might require addressing during the modeling process.
- **Insights into Health Indicators**: For heart disease prediction, correlations can reveal how different health indicators like cholesterol levels, blood pressure, and glucose levels interrelate, offering insights into the multifactorial nature of heart disease.
- **Predictive Model Refinement**: The correlation heatmap can help in refining predictive models by identifying which variables might be most important or which combinations of variables should be explored further.

Overall, the heatmap for correlation between attributes is a powerful tool in exploratory data analysis, providing a comprehensive overview of how different health indicators interact with each other, which is essential in building effective and accurate machine learning models for heart disease prediction.

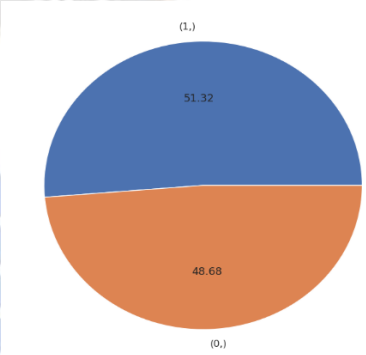

Fig 12: Pie chart showing imbalanced data.

The "Pie chart showing imbalanced data" in your document likely illustrates the imbalance in your dataset, a common challenge in machine learning, especially in medical datasets like those used for predicting heart disease. A pie chart is an effective way to visualize the proportion of different classes in a dataset, making it immediately apparent if there is an imbalance.

In such a pie chart, you can expect:

- **Proportional Representation**: The pie chart will visually represent the proportions of different classes in your dataset. For instance, if you're predicting heart disease, the classes might be 'disease present' and 'disease absent'.
- **Visualizing Imbalance**: The chart will clearly show if one class significantly outweighs the other. In many medical datasets, the number of negative cases (no disease) often exceeds the number of positive cases (disease present), leading to an imbalanced dataset.
- **Importance of Addressing Imbalance**: Visualizing this imbalance is crucial as it impacts the performance of machine learning models. Models trained on imbalanced data can develop a bias towards the majority class, reducing their effectiveness in accurately predicting the minority class.

- **Guiding Data Preprocessing Steps**: Recognizing data imbalance is the first step in applying techniques to address it, such as oversampling the minority class, undersampling the majority class, or using specialized algorithms designed to handle imbalanced data.
- **Impact on Model Evaluation**: The pie chart can also inform how you evaluate your machine learning models. In cases of imbalance, traditional accuracy might not be the best metric, and you might need to consider other metrics like precision, recall, or the F1 score.

Overall, the pie chart showing the imbalanced data is a clear and straightforward way to highlight one of the key challenges in your dataset, guiding the necessary steps in data preprocessing and model evaluation to develop a robust and effective predictive model for heart disease.

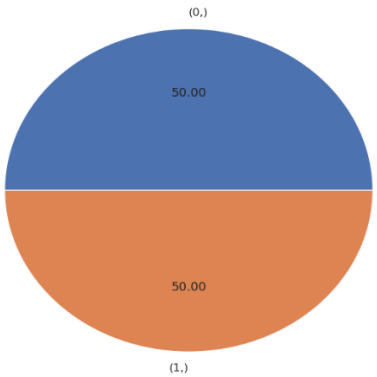

Fig 13: Piechat showing balanced data.

The "Pie chart showing balanced data" is likely illustrates the dataset's composition after applying techniques to balance it. Balancing the data is a crucial step in preparing for machine learning, especially in fields like healthcare, where imbalanced data can lead to biased models that do not accurately predict outcomes for the minority class.

In this pie chart, you would typically see:

- **Equal or Near-Equal Proportions**: After balancing, the pie chart should show equal or nearly equal proportions of the different classes in your dataset. For instance, in a heart disease prediction model, the classes might be 'disease present' and 'disease absent'.
- **Visualization of Balanced Classes**: The chart provides a clear visual representation of the now-balanced classes, contrasting the initial imbalanced state of the data.
- **Techniques Used for Balancing**: The balanced state reflects the application of techniques like oversampling the minority class, undersampling the majority class, or employing synthetic data generation methods like SMOTE (Synthetic Minority Over-sampling Technique).
- **Importance for Model Training**: Balanced data ensures that the machine learning model does not develop a bias towards the majority class, leading to more accurate and generalizable predictions.
- **Guidance on Model Evaluation**: With balanced data, traditional accuracy becomes a more relevant metric for model evaluation. However, other metrics like precision, recall, and F1-score remain important for a comprehensive assessment of model performance.

Overall, the pie chart showing balanced data is an essential visualization in your exploratory data analysis, indicating the effectiveness of your data preprocessing steps in creating a dataset that allows for the training of unbiased and effective machine learning models, particularly critical in the domain of heart disease prediction.

IV. CONCLUSION

In conclusion, the utilization of machine learning techniques for disease prediction in healthcare, with a specific focus on heart disease, represents a significant stride towards enhanced patient care and diagnostic accuracy. Heart disease, a prevalent and complex health issue, demands timely and precise identification for improved outcomes.

Addressing the challenge of imbalanced datasets, a common hurdle in healthcare, is pivotal in mitigating biased predictions. The application of the Random Under Sampler method ensures a balanced representation of data, facilitating fairer and more reliable predictions.

Through rigorous evaluation and comparison of supervised machine learning algorithms, this project strives to identify the most accurate predictive model for future applications in healthcare. Ultimately, this research aims to empower healthcare professionals with a robust tool for the early and accurate detection of heart disease, offering hope for improved patient care and outcomes on a global scale.

REFERENCE

[1] Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. *J. Hepatol.* 2018, *69*, 896–904.

[2] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* 2022, *21*, 240.

[3] Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits,

Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332.

[4] Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. *Circulation* 2019, *139*, e56–e528.

[5] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* 2021, *26*, 100655.

[6] Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. *Circulation* 2015, *131*, e29–e322.

[7] Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

[8] Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. *J. Occup. Health* 2016, *58*, 216–219.

[9] Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. *Procedia Comput. Sci.* 2016, *85*, 962–969.

[10] Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *Int. J. Comput. Appl.* 2011, *17*, 43–48.

[11] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* 2019, *7*, 81542–81554.

[12] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. *Eur. J. Mol. Clin. Med.* 2020, *7*, 1638–1645.

[13] Breiman, L. Random forests. *Mach. Learn.* 2001, *45*, 5–32.

[14] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

[15] Gietzelt, M.; Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. *Comput. Methods Programs Biomed.* 2013, *111*, 62–71.

[16] K, V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. *Int. J. Comput. Appl.* 2011, *19*, 6–12.

[17] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

[18] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* 2020, *1*, 345.

[19] Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl.* 2019, *10*, 261–268.

[20] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol.* 2020, *11*, 49–62.