International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

Implementation of Decision Tree Algorithm for Prediction of Rheumatoid Arthritis Disease

Viswanatha V[1], Ramachandra A.C[2] and Krupa J[3]
[1]Assistant Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA
[2]Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA
[3]Student, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA

[1]Corresponding Author: viswas779@gmail.com

Received: 01-07-2023 Revised: 16-07-2023 Accepted: 30-07-2023

ABSTRACT

A very challenging and difficult part in finding cure for Rheumatoid Arthritis (RA) disease is the same as finding cure for autoimmune disease. In case of autoimmune diseases, it is very difficult to find the cause. RA is also one of the autoimmune diseases that causes inflammation in joints. It does not only limit to joints but also can spread or occur in various parts of body like skin, lungs and many others. Many features like age, sex, previous history of certain diseases affect the occurrence of RA. The proposed ML model uses a decision tree algorithm to analyse a comprehensive set of clinical and laboratory variables collected from RA patients. These variables include age, ID, time, treatment, gender, and other baseline features. The ML model employs feature selection techniques to identify the most relevant predictors that contribute to disease progression. By training the decision tree algorithm on a large dataset of RA patients, the model generates an accurate predictive model for assessing disease severity and progression. The model is able to detect RA disease with almost 80% accuracy with the given training dataset. RA can have many other symptoms and characteristics. This data was trained on a basic dataset available from the Kaggle website. The dataset is in CSV format and has 6 features. The model was trained using a decision tree algorithm, which categorizes the data into different categories and then checks for specific data when needed in Jupyter notebook. The study explored detecting basic symptoms and characteristics of RA disease. The Decision tree model has given accuracy of 86.5% and 74.4% on training and testing respectively. So Accuracy of model is pretty good for the dataset used.

Keywords— Rheumatoid Arthritis, Autoimmune Disease, Decision Tree, Machine Learning

I. INTRODUCTION

The joints are predominantly affected by the chronic inflammatory illness known as rheumatoid arthritis (RA). RA is a type of autoimmune illness that is brought on by a confluence of genetic and environmental variables. These elements alter our own antigens, rendering them invisible to the immune system. Antigen-presenting cells take these altered antigens up and transport them to the lymph nodes. There, T helper cells get the instructions they need to stop multiplying and start differentiating into plasma cells. Specific antibodies made by these plasma cells are directed against self-antigens. Joints are referred to as "arthr" and inflammation as "itis". The term "rheumatism," which refers to diseases of the musculoskeletal system, is where the name "rheumatoid" comes from. Chronic joint inflammation results from the immune system's erroneous attack on its own cells, which occurs in RA. The fibrous joint capsule, articular cartilage, and synovial membrane that carries synovial fluid, the joint lubricant, make up the area between the two joints.

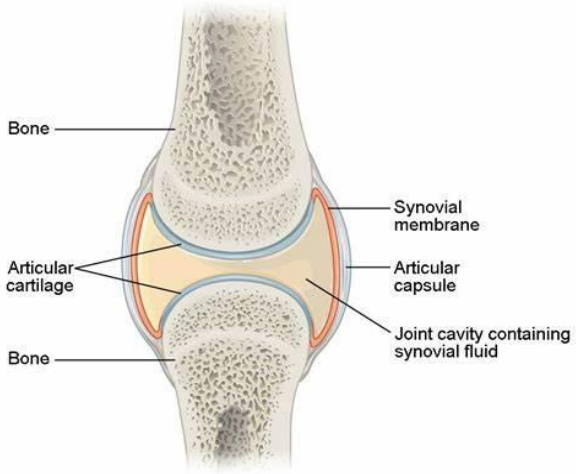

Figure 1: Synovial fluid present between membranes

A genetic and environmental interplay that alters our own antigens causes RA, an autoimmune disease. Our immune cells cannot identify and detect it as alien bodies as a result of this change. Antigen presenting cells transported to the lymph nodes take up these antigens. A signal is then given to B cells telling them to cease multiplying and differentiating into plasma

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

cells, which then manufacture specific antibodies against self-antigens, by T helper cells. T-helper cells and antibodies enter the bloodstream and travel to the joints in RA. Here, T-helper cells begin secreting cytokines and enlisting the creation of macrophages, which in turn boosts the quantity of inflammatory cells of various types. This results in the proliferation of synovial cells and the development of the PANNUS, which is made up of fibroblasts, myofibroblasts, and inflammatory cells and is depicted in figure 2b. Over time, this PANNUS can destroy bone in addition to degenerating cartilage. Once the cartilage is gone, the bones might rub against one another. Additionally, the production of osteoclasts and an increase in inflammatory cells cause the destruction of bones, as seen in figure 2a. Antibodies also penetrate the shared space in the meanwhile. Rheumatoid factor, an IgM antibody, is one of the antibodies. Immune complexes are created once they begin to attach to antigen. They induce damage and inflammation by activating the complement system. Finally, as illustrated in figure 2, persistent inflammation results in angiogenesis, which is the development of new blood vessels, which produces even more inflammatory cells.

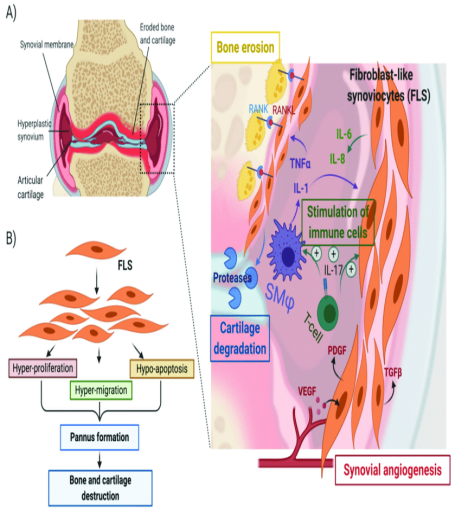

Figure 2: (a) shows bone erosion (b) shows PANNUS formation 2. Shows angiogenesis process

These are not only restricted to joints; through circulation, they can also damage the blood vessels, liver, brain, skeletal muscle, skin, and lungs. They are frequently symmetrical and are brought on by many joints. Consider the case of hands. All of these will result in a number of indications and symptoms that are present in the dataset that is being employed in the model.

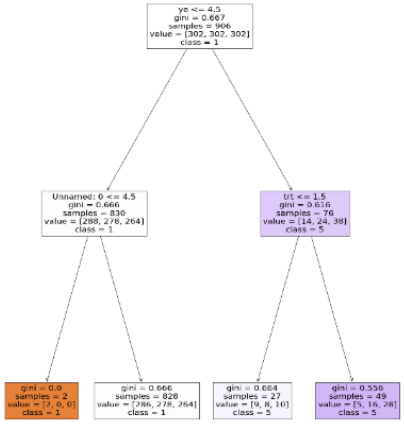

Figure 3: shows the decision tree algorithm representation and splitting into specific data

The decision tree method is used in the created model to forecast the sickness. The supervised machine learning technique known as the decision tree algorithm has a hierarchy of tree nodes and numerous nodes. The root node of the network is a representation of the total population that will be split. Splitting is the process of dividing a node. Decision nodes are produced when the formed sub node splits once more. The leaf node is the node that does not divide. Pruning is the process of removing a decision node's child nodes. To avoid data overfitting, this step is crucial. Data is initially extracted from the dataset, after which it is subjected to rules and separated accordingly. This is repeated until all of the data points have been examined and knowledge has been gathered. If there are no more decision groups to separate, as shown in figure 3, the splitting terminates. The decision tree is simple to perceive and understand. However, they have the drawback of being extremely susceptible to overfitting. To increase model accuracy and decrease overfitting, we employ a variety of regularization and normalizing techniques in the Jupyter notebook. The software tools that are most helpful for scientific computing and code execution are Jupyter notebook and anaconda. Together, these tools function as extraordinary software. Numerous other apps, like Data Spell, Jupyter Lab, Spyder, Glue Vis, and others, are included with the Anaconda package. In this application, the model is currently executed and trained using a Jupyter notebook using a specific dataset. The Jupyter notebook program is user-friendly and works with the majority of operating systems. Excellent tool for data exploration, algorithm prototyping, and sharing findings. They can be combined to be used for a variety of data research applications.

II. LITERATURE REVIEW

RA may be a systemic immune system illness that fundamentally influences the joints. It appears shifting predominance around the world, with higher

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

rates in industrialized nations due to natural and hereditary components, socioeconomics, and under-reporting. The seriousness of RA has declined over the past three decades, ascribed to made strides treatment approaches and illness administration. RA predominance, in any case, has been expanding. Chance components for RA incorporate modifiable way of life factors and non-modifiable components like hereditary qualities and sex. Understanding the common history of RA and its advancement in particular populaces may help in actualizing focused on anticipation methodologies. Assist inquire about is required to improve information and present viable avoidance measures for this debilitating illness.**[1]** Significant advancements in our understanding of the pathogenesis of RA (RA) have led to notable improvements in treatment modalities and overall quality of life. The introduction of targeted biologic and synthetic disease-modifying anti-rheumatic drugs (DMARDs) has transformed clinical outcomes in a remarkable manner. However, it is important to acknowledge that RA remains a chronic condition without a definitive cure. There are still unmet needs, including the challenge of partial or non-response to treatment in a substantial number of patients, the inability to achieve immune homeostasis or drug-free remission, and the lack of effective tissue repair mechanisms. RA is now recognized as the culmination of a prolonged prodromal phase characterized by systemic immune dysregulation, likely originating from mucosal surfaces, followed by a symptomatic clinical phase. The inflammatory and immune responses primarily manifest in the synovium, resulting in pain, joint damage, and a range of associated comorbidities. In this review, we aim to provide insights into recently elucidated immunologic mechanisms that underlie the breach of tolerance, chronic synovitis, and remission in RA. **[2]** RA (RA) is a chronic inflammatory autoimmune disorder characterized by symmetrical joint involvement, persistent pain, tenderness, and joint destruction. The majority of RA patients exhibit autoantibody production, and the contribution of immune cells, as well as other cell types like synovial fibroblasts, in the pathogenesis of the disease is well-established. Genetic factors, particularly within the HLA locus, have significant associations with RA, while non-HLA genetic variants confer relatively modest risk. Distinct genetic profiles exist between autoantibody-positive and autoantibody-negative RA, with specific alleles of HLA-DRB1, such as HLA-DRB1*04 and *10, demonstrating the highest risk association. Conversely, HLA-DRB1*13 alleles exert a protective effect against autoantibody-positive RA. Mechanistic understanding of the precise role of these genetic variations in RA susceptibility is an ongoing area of research, focusing on immune receptor binding, T cell activation, and cell signalling pathways. Gene-gene and gene-environment interactions further contribute to the overall risk of RA. Currently, more than 150 candidate loci with polymorphisms associated with RA, particularly seropositive disease, have been identified, and ongoing investigations in diverse populations hold promise for future discoveries. These advancements will contribute to a comprehensive integration of genetic, epigenetic, transcriptomic, and proteomic data, enriching our understanding of RA pathogenesis.**[3]** Rheumatoid joint pain (RA) could be a persistent and systemic immune system illness characterized by synovial tissue hyperplasia, arrangement of pannus, cartilage debasement, and systemic complications. Noteworthy strides have been made in explaining the pathogenic instruments including autoreactive CD4+ T cells, B cells, macrophages, provocative cytokines, chemokines, and autoantibodies that contribute to the pathogenesis of RA. Be that as it may, in spite of these outstanding progressions, there remains a significant sum of information to be revealed. This comprehensive survey presents an upgraded and in-depth understanding of the fundamental pathogenesis of RA, advertising profitable bits of knowledge into potential novel helpful targets for future investigation.**[4]** RA is classified among the various types of arthritis, exceeding a count of 100. This chronic autoimmune condition affects the synovial joint lining, resulting in significant joint deformities and functional impairment. With an estimated prevalence of approximately 0.5%, RA impacts individuals across different age groups and genders, with a higher incidence among elderly individuals and women. Substantial advancements have been made in recent decades, unravelling the pathogenesis of RA, facilitating early detection, and introducing novel therapeutic options.

While non-steroidal anti-inflammatory drugs (NSAIDs), corticosteroids, and disease-modifying antirheumatic drugs (DMARDs) remain commonly employed treatments, their efficacy varies among patients. Consequently, there is a pressing need for innovative solutions to enhance disease outcomes. This comprehensive review delves into recent discoveries pertaining to various classes of RA therapies, encompassing traditional and contemporary drug therapies, as well as emerging avenues such as phyto-cannabinoid and cell- and RNA-based therapies. A deeper comprehension of their mechanisms and pathways holds the potential to identify precise targets for inflammation, cartilage damage, and mitigate adverse effects in the context of arthritis.**[5]** Recent research has revealed a heightened risk of renal damage and concomitant cardiovascular complications in individuals diagnosed with RA (RA), significantly impacting their prognosis. However, the precise prevalence of chronic kidney disease (CKD) within the Uzbekistan cohort of RA patients remains to be accurately determined. The prognostic implications of renal impairment in RA have garnered considerable attention among researchers in recent years. Several clinical presentations of kidney involvement in the pathological process of RA have been observed in the majority of patients. These include glomerulonephritis, amyloidosis, vasculitis, as well as

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

iatrogenic forms such as analgesic tubulopathy and membranous nephropathy. Notably, due to practical reasons, the morphological verification of renal pathology may be delayed in real clinical settings. The early signs of functional renal impairment, particularly in cases of moderate severity, are not always discerned by clinicians, despite the potential rapid progression of CKD in RA, particularly in the elderly and those with concurrent cardiovascular conditions. The development of nephropathy in RA encompasses a complex multifactorial nature, presenting in various clinical and morphological manifestations. Thus, distinct clinical patterns of kidney damage are recognized in RA, including amyloidosis, glomerulonephritis, rheumatoid granulomatosis, rheumatoid renal vasculitis, and iatrogenic forms such as medicinal tubulointerstitial nephritis and membranous nephropathy. In practical clinical practice, the nosological diagnosis of kidney disease in RA is typically established upon the appearance of clinical and laboratory criteria, with proteinuria being a vital parameter. However, recent evidence indicates that renal dysfunction can develop even in the absence of notable proteinuria, highlighting the importance of detecting early signs of renal impairment. It is noteworthy that rheumatologists may not consistently pay attention to the initial manifestations of functional renal disorders, particularly in cases of moderate proteinuria. Yet, the rate of decline in kidney function in RA can be substantial, especially in older individuals and in the presence of cardiovascular comorbidities.**[6**] RA (RA) poses a significant medical and societal challenge, characterized by progressive connective tissue disorganization and profound immunopathological changes with auto aggressive features. Among inflammatory joint diseases, RA exhibits the highest prevalence. The disease's societal impact extends beyond its high prevalence, causing substantial financial burdens on society, patients, and their families due to significant disability rates and early onset of functional impairment. Despite the utilization of modern therapeutic approaches, RA demonstrates a relentless progression, resulting not only in substantial locomotor functional deficits but also in a reduced lifespan of patients by 4-10 years, surpassing the mortality rate of the general population. The prognosis is particularly unfavourable for RA patients with systemic manifestations, including generalized vasculitis, rheumatoid nodules, lymphadenopathy, and multi-organ involvement (such as lungs, heart, liver, and kidneys). While extra-articular manifestations of RA have been extensively studied, gastrointestinal (GI) involvement remains relatively understudied, despite the severity of intestinal amyloidosis, which affects approximately 11% of patients and is often concurrent with amyloidosis in other internal organs.**[7]** RA is characterized by chronic inflammation and progressive degradation of bone and cartilage in affected joints, resulting in significant bone loss. The pathological process involves an imbalance between osteoclastic bone resorption and impaired osteoblastic bone formation. Extensive research has demonstrated the pivotal role of osteoclasts in bone erosion, supported by the clinical efficacy of antibodies targeting RANKL, a key regulator of osteoclastogenesis. Synovial fibroblasts contribute to joint damage by activating both pro-inflammatory and tissue-destructive pathways. Recent advancements, including state-of-the-art techniques like single-cell RNA sequencing, have unveiled the heterogeneity of synovial fibroblasts and immune cell populations, including T cells and macrophages. Understanding the intricate interplay between immune cells and fibroblasts is crucial in elucidating the mechanisms underlying bone damage in RA. Specifically, the dysregulated balance between regulatory T cells and T helper 17 cells intensifies both inflammation and bone destruction by promoting RANKL expression on synovial fibroblasts. A comprehensive comprehension of the immune mechanisms governing joint damage and the intricate immune system-synovial fibroblast-bone interplay holds immense potential for identifying novel therapeutic targets in the management of RA. **[8]** RA imposes a substantial medical and societal burden due to its high morbidity and disability rates. Recent evidence suggests that RA exhibits a microenvironment akin to a tumour, known as the RA microenvironment (RAM). The RAM is a complex and intricate network comprising various extracellular matrix factors and diverse stromal cells. Conventional anti-inflammatory therapies face limitations in penetrating the RAM, offering only transient relief for RA symptoms. In this context, nanomaterials have emerged as a promising approach to overcome the challenges posed by the RAM. Leveraging their ability to target multiple pathogenic factors simultaneously, nanomaterials provide an avenue for precise and effective treatment beyond traditional modalities. This comprehensive review provides a systematic overview of the distinct features and constituents of the RAM, emphasizing the pivotal role of the vicious cycle involving reactive oxygen and nitrogen species (RONS) and inflammatory factors in driving RA progression. Furthermore, it comprehensively summarizes the treatment strategies and recent advancements in nanomaterial-based therapies tailored for RA. The review concludes by discussing the challenges associated with the clinical application of RAM-targeting nanomaterials and highlighting future research directions in the field of RA treatment.**[9]** In recent years, considerable attention has been dedicated to understanding the epigenetic dysregulation that contributes to the pathogenesis of autoimmune rheumatic diseases. RA is a heterogeneous disease, where a complex interplay of immunologic, genetic, and epigenetic factors shapes disease manifestations and progression. Within this realm, microRNAs (miRNAs) have emerged as pivotal regulators of immune cell development and function. Uncovering disease-associated miRNAs ushers in a new era of post-genomic exploration, offering potential avenues to modulate the

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

genetic impact of autoimmune diseases. Certain miRNAs hold promise as biomarkers for disease diagnosis, prognosis, treatment response, and other clinical applications. This review not only outlines the influence of miRNAs on immune and inflammatory responses in RA but also highlights their utility as diagnostic and prognostic biomarkers. While research on miRNAs is still evolving, investigating these novel biomarkers has the potential to advance the field of personalized medicine in RA treatment. Lastly, we explore the possibility of miRNA-based therapies in RA patients, leveraging significant strides made in the management of inflammatory arthritis.**[10**] RA (RA) is a prevalent autoimmune disorder, afflicting an estimated 0.5-1% of the global population. It is characterized by symmetrical peripheral polyarthritis, systemic inflammation, and diverse clinical manifestations. The disease exerts a significant burden on individuals and society, leading to joint destruction, profound disability, and an augmented risk of cardiovascular and pulmonary complications. RA encompasses two major subtypes based on the presence or absence of specific autoantibodies. During the preclinical phase, immune system activation, autoantibody production, and non-specific musculoskeletal symptoms emerge. Early initiation of treatment in RA has yielded noteworthy enhancements in disease outcomes and physical function. Targeted interventions during the preclinical stage hold promise for preventing or delaying disease onset. Multiple factors, encompassing genetics, environmental exposures, and lifestyle choices, contribute to disease risk. Identifying at-risk individuals and implementing timely preventive measures or pharmacological interventions are critical for mitigating the impact of RA. Nonetheless, formal treatment recommendations for individuals in the preclinical stage of RA remain elusive, necessitating further research to explore interventional strategies and incorporate valuable insights from patient perspectives.**[11]** RA (RA) is characterized by significant joint and bone damage resulting from an aberrant autoimmune response localized at the articular sites. The worldwide annual incidence and prevalence rates of RA are estimated at 3 cases per 10,000 population and 1%, respectively. Pathogenesis involves a complex interplay of genetic and environmental factors, including microbiota, smoking, and infectious agents. While conventional treatment approaches, mainly relying on Disease Modifying Anti Rheumatic Drugs (DMARDs) and Glucocorticoids (GC), remain the cornerstone of management, novel strategies incorporating biological DMARDs are being actively investigated. Personalized therapeutic approaches targeting specific disease pathways offer promising prospects, albeit the economic burden and potential side effects associated with these treatments call for early identification of inadequate responders to conventional DMARDs. Given the variable remission rates in RA, it is essential to evaluate current therapeutic options and explore potential personalized interventions. This comprehensive analysis presents valuable insights into RA treatment strategies, encompassing clinical trials exploring combination therapies and addressing specific subgroups such as seronegative patients with moderate to high disease activity. Additionally, the review highlights the need for further research into novel therapeutic modalities, such as gene therapy and mesenchymal stem cell therapy, to overcome the limitations and adverse events associated with current approaches.**[12]** The existing pharmacological interventions for managing RA encompass nonsteroidal anti-inflammatory drugs, disease-modifying antirheumatic drugs, and biologics aimed at alleviating disease symptoms. However, these therapies administered through conventional delivery systems are associated with drawbacks such as limited selectivity and potential adverse effects on non-target tissues. The emergence of microneedles-based transdermal drug delivery has garnered significant interest as a potential solution to overcome these limitations inherent in conventional formulations. Microneedles offer a promising approach for targeted drug delivery, enabling improved therapeutic efficacy while minimizing systemic side effects. This professional summary highlights the potential of microneedles-based transdermal drug delivery as an innovative strategy to enhance the treatment of RA, addressing the current challenges associated with conventional preparations.**[13]** RA is a systemic autoimmune disorder characterized by the accumulation of inflammatory and immune cells within inflamed joints. Extensive research efforts have focused on exploiting the unique features of RA to develop effective therapeutic strategies. One approach involves utilizing endogenous materials to engineer drug-loaded nanoparticles that can specifically target RA by binding to cell adhesion molecules or chemokines. Additionally, nanoparticles can be designed to respond to the microenvironmental cues associated with RA. These innovative approaches hold great promise in achieving targeted and responsive treatment for RA, potentially improving therapeutic outcomes and minimizing off-target effects.[14]

III. METHODOLOGY

The methodologies include the algorithm used, dataset used and flowchart of the data used and implemented. Below is the provided step by step explanation of the algorithm used.

Algorithm used: The decision tree algorithm is a widely used supervised learning technique employed for both classification and regression tasks. It constructs a structured model resembling a flowchart, driven by input features.

1. Tree Construction: The algorithm commences by considering the entire dataset as the root node, and selects the optimal feature for partitioning the data.

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

2. Feature Split: The chosen feature is utilized to divide the data into subsets, thereby creating branches or paths within the decision tree.
3. Recursive Splitting: The process of feature splitting is iteratively applied to each subset until a predefined stopping criterion is satisfied.
4. Leaf Node Assignment: Leaf nodes are assigned class labels or regression values based on the majority class or mean value of the target variable within each respective subset.
5. Prediction: To make predictions, the algorithm traverses the decision tree by evaluating feature values and ultimately reaching a leaf node to obtain the final prediction.
Advantages: Easy to comprehend and interpret - Accommodates numerical and categorical data - Handles missing values gracefully - Captures non-linear relationships effectively.

Limitations: Prone to overfitting, necessitating proper regularization techniques - Can be sensitive to changes in the dataset, leading to instability. Exhibits bias towards features with high cardinality or many levels In conclusion, decision trees offer versatility and transparency in model interpretation. However, caution must be exercised to address overfitting issues and effectively manage the algorithm's limitations.1. Import necessary libraries and module: from sklearn. tree import Decision tree classifier, import pandas as pd and few other libraries are imported. 2. Read the dataset: use necessary commands to read the dataset provided and collect the information from it. 3. Preprocess the data: here the data collected is pre-processed so that feature extraction can take place easily. 4. Load the data: The data pre-processed now will be loaded into the training model to train the model properly. 5. Splitting the data: The data is splitted into training and testing data in proportion of 80:20 out of 100. 6. It checks for the target: if present then goes to next step or else declares it as unsupervised model as it does not have specified output. 7. Checks for target data: we have used the decision tree classifier therefore uses discrete data. 8. Training model: after all the necessary adjustments in the code we train the model using decision tree algorithm. 9. Testing data: after training the model is tested and accuracy is obtained with necessary graphs. A small comparison plot is also included in algorithm. 10. Results: as last step the results are obtained and the model execution ends. Few additional steps of regularization imputation and many other commands are included in the code to improve the accuracy and prevent the model from overfitting and underfitting. This also increases the generalization of the model and allows it to predict the new data with more accuracy.

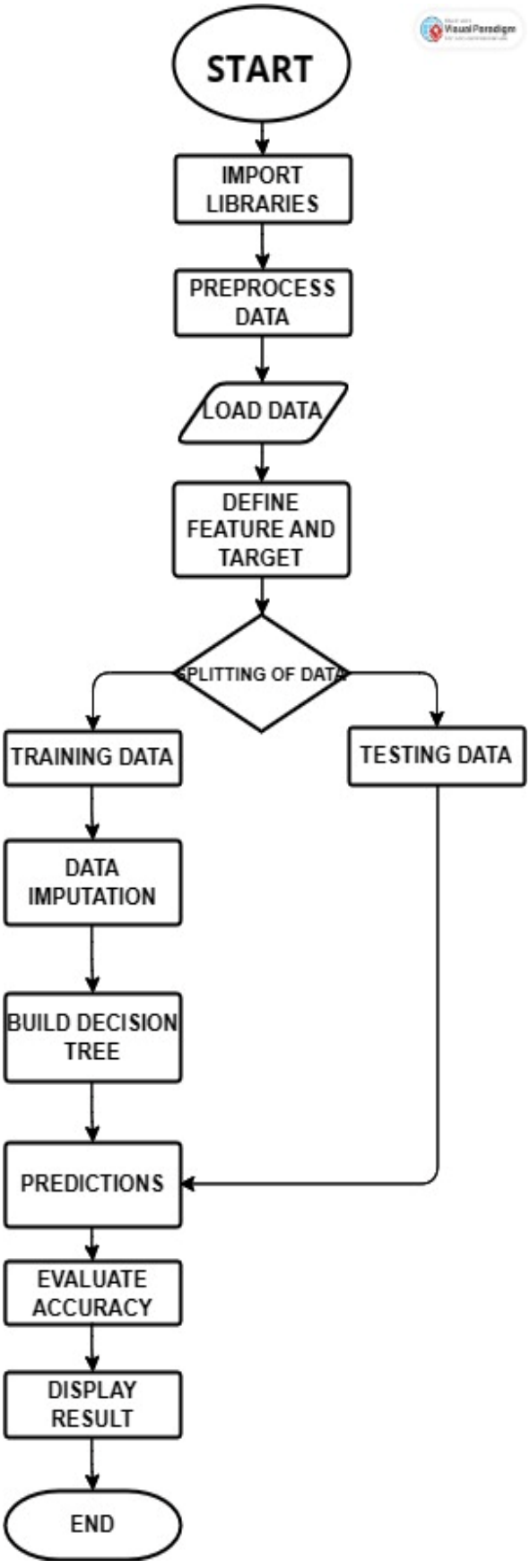

Figure 4: Flow chart of model training

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

Here is a brief explanation of the flowchart:
1. Start: The flowchart begins with the start symbol, indicating the beginning of the decision tree algorithm.
2. Load Dataset: The algorithm loads the dataset, which contains the input features and target variable.
3. Define Features and Target: The feature columns and target column are defined, specifying the variables to be used for training the decision tree.
4. Split Data: The dataset is split into training and testing sets using the train_test_split function, allocating a portion of the data for model evaluation.
5. Data Imputation: The Simple Imputer object is used to handle missing values in the dataset, replacing them with the mean value of the respective feature.
6. Build Decision Tree: The DecisionTreeClassifier object is created, representing the decision tree model. It is trained on the training data using the fit function.
7. Predictions: The trained decision tree is utilized to make predictions on the test set, using the predict function.
8. Evaluate Accuracy: The accuracy of the model is calculated by comparing the predicted values with the actual target values using the accuracy score function.
9. Display Results: The accuracy score is printed to the console, providing an assessment of the model's performance.
10. End: The flowchart concludes with the end symbol, indicating the completion of the decision tree algorithm. The flowchart shown in figure 4 provides a visual representation of the steps involved in training and evaluating the decision tree model, aiding in understanding the overall process and facilitating communication between different stakeholders.
Datasets: The dataset currently used in the model training is taken from the Kaggle website. Kaggle is a popular platform for data scientists and machine learning practitioners to discover and share datasets, as well as participate in data science competitions. It hosts a vast collection of datasets from various domains, allowing users to access and analyse real-world data. Kaggle datasets are typically provided in structured formats such as CSV, Excel, or SQL, and may contain a wide range of variables or features. These datasets cover diverse topics including healthcare, finance, social sciences, computer vision, natural language processing, and more. Users can explore and download datasets from Kaggle for their own analysis, model training, or research purposes. They can also contribute by uploading and sharing their own datasets with the Kaggle community. Kaggle datasets are accompanied by detailed descriptions, documentation, and often include pre-split train/test datasets to facilitate model development and evaluation. Additionally, many datasets come with sample code notebooks, known as kernels, which provide examples and insights on how to work with the data. By leveraging the vast collection of Kaggle datasets, data scientists can gain access to high-quality, real-world data to explore, analyse, and build predictive models or develop insights for a wide range of applications. The dataset contains basic clinical features of RA which includes id, sex, age, time, treatment, baseline and time. These features are utilized in the model training to predict the disease accurately.

IV. RESULTS AND DISCUSSIONS

The machine learning model utilizes python code in Jupyter notebook to train and test the model. The model uses decision tree algorithm. The code starts from training the feature and target columns from dataset. Numpy is a Python library for efficient numerical computations, offering multidimensional arrays and mathematical functions. It is widely used in scientific computing and data analysis. Pandas is a powerful data manipulation library built on top of Numpy. It provides high-level data structures like Data Frames and Series, making data manipulation and analysis easier. Data Frames are two-dimensional tables with labelled rows and columns, while Series are one-dimensional labelled arrays. Pandas offers tools for data cleaning, preprocessing, merging, reshaping, and analysing structured data. It supports flexible indexing, filtering, and grouping operations, allowing easy extraction and manipulation of specific subsets of data. Numpy and Pandas are widely used in scientific computing, data analysis, and machine learning applications. Numpy provides efficient storage and manipulation of large arrays, while Pandas provides intuitive data manipulation capabilities. Both libraries integrate well with other Python data ecosystem tools like Matplotlib and Scikit-learn. Numpy and Pandas are essential for data manipulation, analysis, and preprocessing in Python. Together, they provide efficient and convenient tools for working with arrays and structured data. The first code snippet used in code in summary, the code performs data preprocessing, splits the data into training and test sets, trains two Decision Tree Classifiers with different parameters, and evaluates their accuracy in predicting the target variable. Here a simple imputer is included. The SimpleImputer is a class from the scikit-learn library that provides a simple strategy for handling missing values in a dataset. It is used to replace missing values with a chosen strategy, such as the mean, median, or most frequent value of the respective feature.

The output we get for the accuracy of model is as follows.

```
Accuracy on training set: 0.865
Accuracy on test set: 0.744
```

The dataset taken, contains 6 features where 5 are taken as feature columns and 1 as target column. The distribution of data is as shown in figure 5 . In summary, next code snippet utilizes a trained Decision Tree Classifier to generate a visual representation of the decision boundaries in the feature space as shown in figure 6. The decision boundaries are plotted as filled contour regions, and the training points from the dataset

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

are displayed with colours representing different features. This visualization helps in understanding how the classifier separates and categorizes the data points based on their feature values. Decision boundaries are lines or surfaces that separate different classes in a classification problem. They define the regions where a classifier assigns different labels. Decision boundaries can be linear or nonlinear, depending on the complexity of the problem and the algorithm used.

They are determined by learning from the training data and finding the optimal separation between classes. Decision boundaries are influenced by the features used for classification and can vary in complexity. Visualizing decision boundaries helps understand how a classifier distinguishes classes and identifies areas of uncertainty.

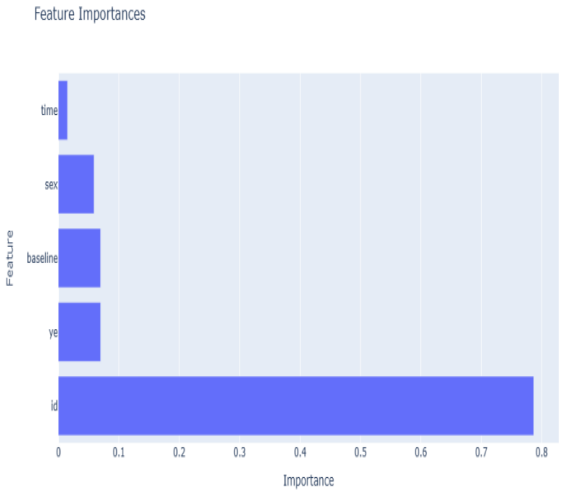

Figure 5: Feature importance of all the features

The complexity of decision boundaries reflects the relationships between features. They are crucial for evaluating classifiers and assessing their generalization ability. Decision boundaries aid in model interpretation, evaluation, and decision-making, providing a visual representation of class separation in the feature space.

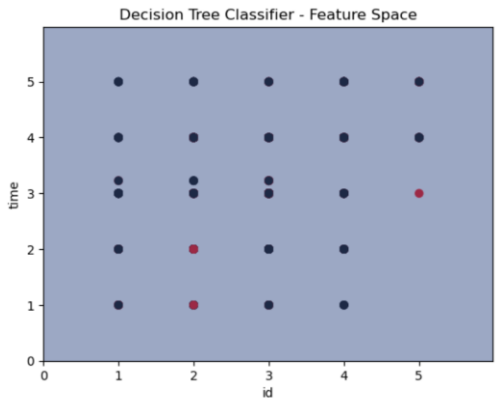

Figure 6: Decision boundaries in feature space [time=black colour, id=red colour]

In summary, next code loads a dataset using Pandas, separates the features and the target variable, and creates an instance of the DecisionTreeClassifier from scikit-learn. It then fits the classifier on the dataset and plots the decision tree using the tree. plot_tree () function. The resulting plot visualizes the decision tree structure, with each node representing a decision based on a specific feature, and the leaf nodes indicating the predicted classes. The plot is displayed using matplotlib. pyplot. show (). A decision tree plot is a graphical representation that illustrates the decision-making process of a decision tree classifier or regressor. It visually depicts the hierarchical structure of the decision tree, where nodes represent decision points and branches represent possible outcomes. The plot displays the feature names and their corresponding thresholds at each decision node. Terminal nodes, or leaves, signify the final decisions or predicted outcomes. The size and colour of the nodes can be utilized to convey the sample count or class distribution at each node. The plot offers a clear and intuitive representation of the decision tree's logic and decision boundaries. It aids in comprehending how the decision tree makes predictions based on various features and thresholds. The plot can be customized to include class labels, feature importance, and other pertinent information. It is valuable for interpreting and explaining the decision tree model to stakeholders or non-technical audiences. The decision tree plot allows for easy identification of significant features and their influence on the decision-making process. It assists in identifying regions of high predictive accuracy as well as areas where the model may struggle to make accurate predictions. The plot can be generated using various Python libraries, such as scikit-learn, matplotlib, or plotly. It helps identify potential overfitting or underfitting issues by visualizing the complexity of the decision tree. The plot serves as a valuable tool for model validation, comparison, and explanation. It facilitates effective communication between data scientists, analysts, and stakeholders by providing a visual representation of the decision tree's logic. The decision tree plot is a powerful visualization tool that enhances the interpretability and transparency of decision tree models. Plot has a max depth of 2 as shown in figure 7. Max depth is command used to reduce the complexity of model and improve model accuracy. Due to their readability and simplicity, decision trees are frequently employed in many different fields. They find use in a variety of fields, including credit scoring, disease diagnosis, fraud detection, loan approval, market segmentation, recommender systems, predictive maintenance, risk assessment, natural language processing, quality control, stock market forecasting, environmental analysis, intrusion detection, customer segmentation, disease risk forecasting, human resource management, and fault diagnosis. Making recommendations, classifying images, analysing sentiment, segmenting markets, predicting equipment failures, assessing risks, performing natural language

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

processing tasks, conducting quality control, predicting stock prices, evaluating environmental impacts, detecting network intrusions, segmenting customers, and predicting disease are all tasks that decision trees are useful for. Decision trees are effective tools in a variety of problem-solving contexts due to their adaptability and application.

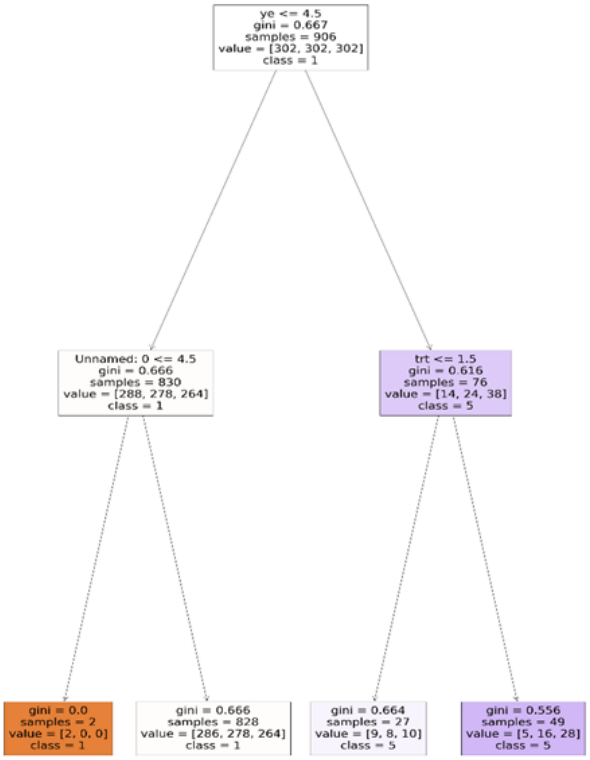

Figure 7: Decision tree plot

In order to understand the behaviour of the features with respect to each other scatter plots are plotted. The code uses the Pandas and Plotly libraries to create a subplot with 6-line plots representing different features of a dataset. The dataset is loaded using Pandas, and a subplot with 6 rows and 1 column is created using Plotly's `make subplots () `. Line plots for each feature are added to the subplot using the `Scatter` trace type from Plotly's graph objects module. The x-values for the line plots are the index of the dataset, and the y-values are the corresponding feature values. The subplot's layout is customized using the `update layout () ` function, specifying the height, width, and title. Finally, the plot is displayed using the `show () ` function. This visualization allows for the examination of the logarithmic trends in the dataset as shown in figure 8. The use of logarithmic trends in data analysis and visualization is to better understand and represent exponential or multiplicative relationships between variables. By plotting data on a logarithmic scale, it compresses large ranges of values and magnifies smaller changes, making it easier to observe patterns and trends. In the provided code, the logarithmic trends of different features in the dataset are visualized using line plots, allowing for a clearer understanding of the relationships and patterns present in the data.

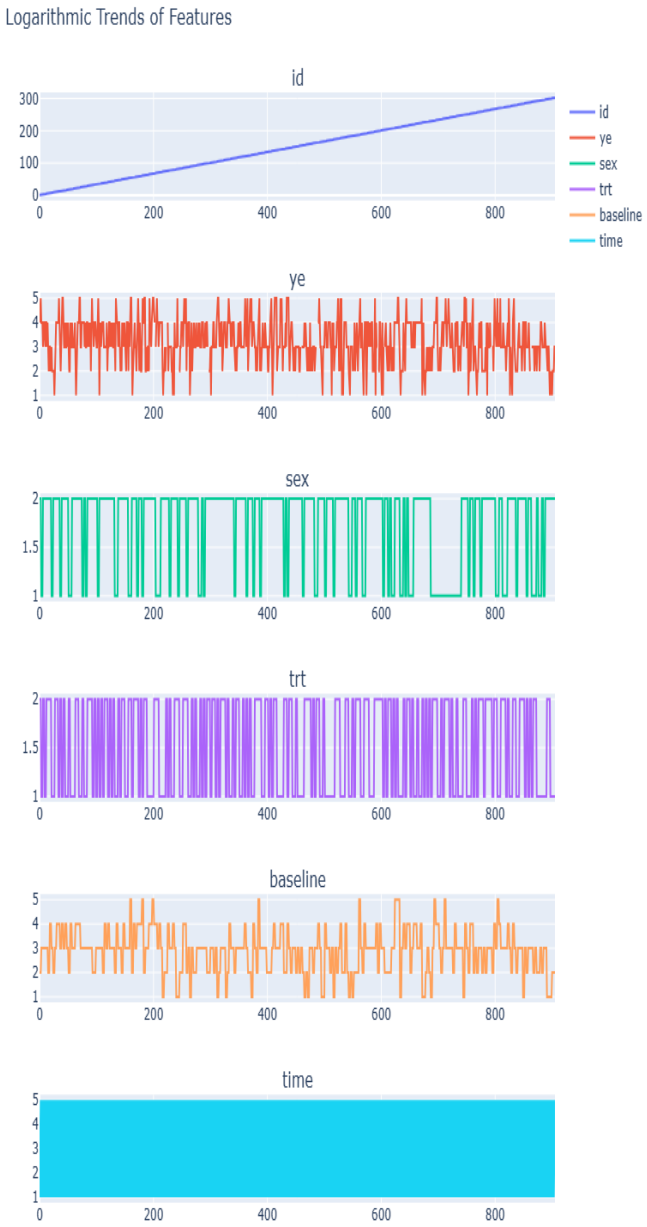

Figure 8: Scatter plot of all features

The model can be trained using any algorithms. But specific applications require specific algorithms for the model give maximum accuracy. Therefore, a small comparison between logistic regression and regression tree were made using necessary libraries and commands. The output is as shown in figure 9.

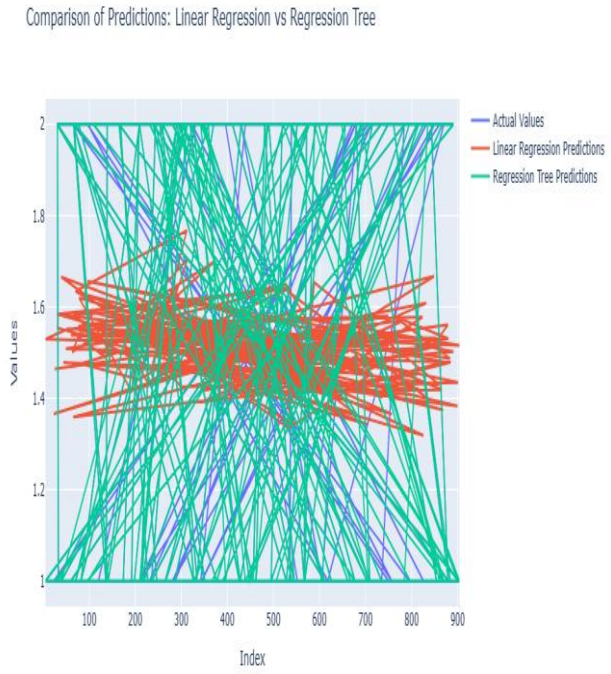

Figure 9: Interactive graph for comparison between two algorithms for same dataset

```
Linear Regression Mean Squared Error: 0.24145570497458857
Regression Tree Mean Squared Error: 0.16483516483516483
```

By looking at the outputs we can say that for this application regression tree is the best fit. The code analyses the "arthritis.csv" dataset by performing various data analysis and visualization tasks. Firstly, the dataset is loaded using pandas. Then, violin plots [figure 10] are created to visually explore the distribution of features in relation to the target variable. A correlation heatmap [figure 11] is generated to investigate the relationships between features and the target variable. The data is prepared for training by splitting it into training and testing sets. Missing values are handled through data imputation using a simple imputer. A decision tree classifier is trained on the training data. Predictions are made on the test set, and the accuracy score is calculated to evaluate the classifier's performance. A confusion matrix [figure 12] is created to visualize the classification results. Finally, a table visualization is generated to provide an overview of the dataset. Overall, this code provides valuable insights into feature distributions, correlations, and the effectiveness of the decision tree classifier for predicting the target variable. The violin plot is a visualization technique used to show.

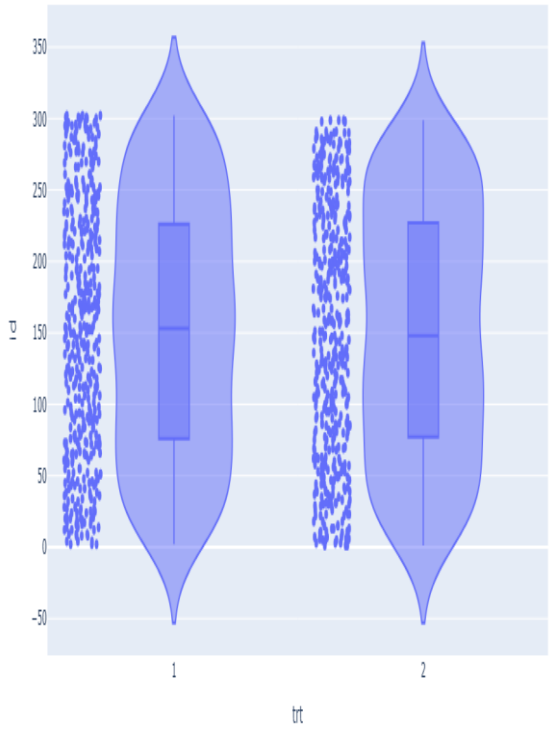

Figure 10 (a): Violin plot between id and trt

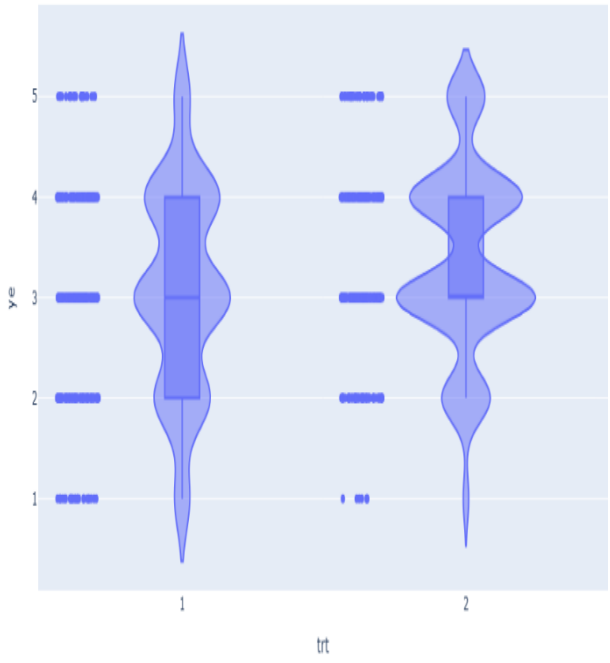

Figure 10 (b): Violin plot between ye and trt

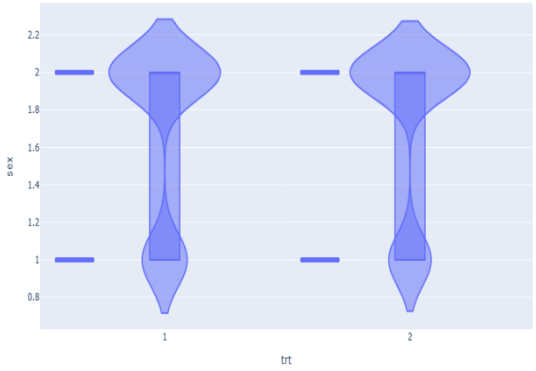

Figure 10 (c): Violin plot between sex and trt

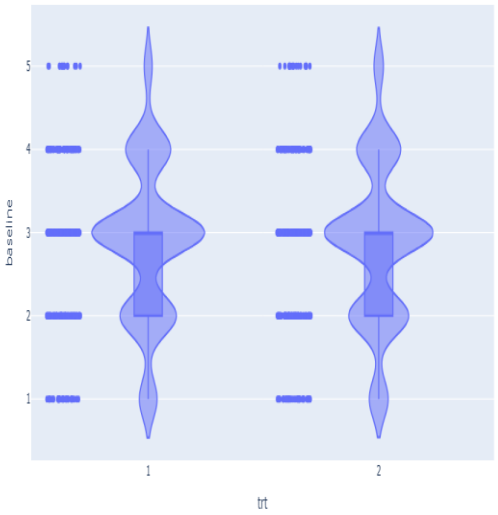

Figure 10 (d): Violin plot between baseline and trt

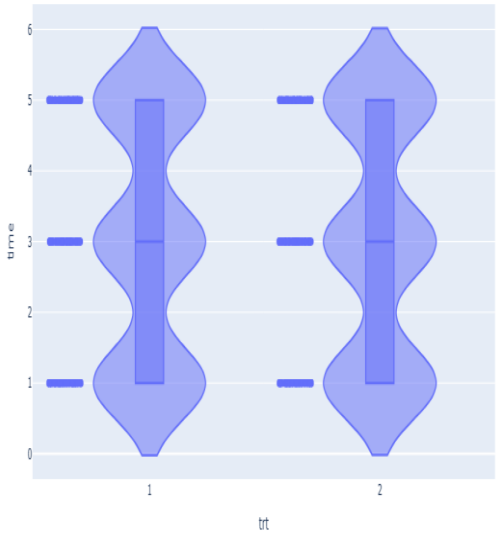

Figure 10 (e): Violin plot between time and trt

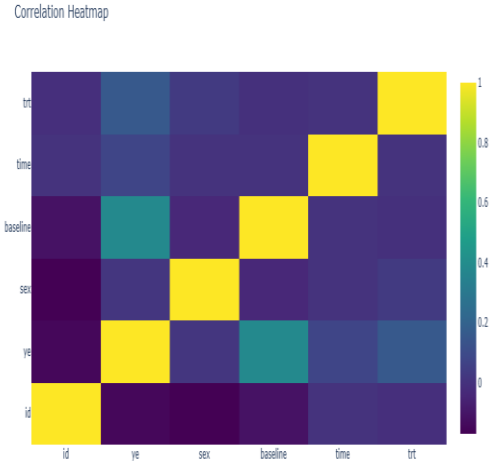

Figure 11: Correlation Heatmap of all the features

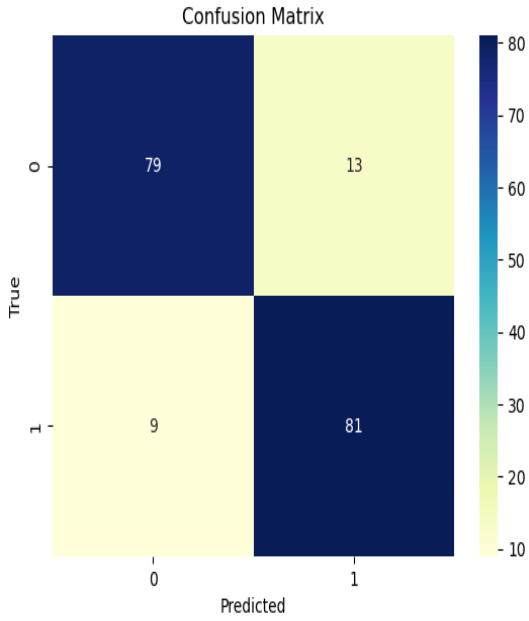

Figure 12: Confusion matrix of true and predicted values

A classification report summarizes the performance of a classification model on a dataset, providing an evaluation of its predictive accuracy for each class. It includes metrics such as precision, recall, F1-score, and support for each class. Precision measures the accuracy of positive predictions, while recall evaluates the model's ability to find positive instances. The F1-score combines precision and recall into a balanced measure of performance. Support indicates the number of instances for each class in the dataset. The classification report as shown in table -I helps assess a model's strengths and weaknesses in classifying different classes, making it useful for comparing algorithms or different configurations. Accurate diagnostic tools,

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

personalized treatment approaches, and better disease management**.**

Table -I: Classification report

Name	Precision	recall	f1-Score	support
1	0.90	0.86	0.88	92
2	0.86	0.90	0.88	90
accuracy			0.88	182
macro avg	0.88	0.88	0.88	182
weighted avg	0.88	0.88	0.88	182

V. CONCLUSION AND FUTURE SCOPE

Researchers and practitioners can gain insights into the dataset, understand feature importance, and build predictive models for RA. The future scope of this includes enhancing feature engineering, exploring different machine learning algorithms, optimizing model hyperparameters, and implementing ensemble methods. Integration of biomarkers and genomic data can provide deeper insights and improve predictions. The developed ML model can be integrated into clinical decision support systems to assist healthcare professionals in making informed decisions. Real-time monitoring and prediction methods can enable early intervention and personalized treatment strategies. Further research can focus on patient stratification, personalized medicine, and developing interpretable models. The application of machine learning techniques to RA holds potential for improving patient care, treatment outcomes, and our understanding of the disease.

ABBREVIATIONS

RA: Rheumatoid Arthritis
ML: Machine Learning
DMARD's: Disease-Modifying Antirheumatic Drugs
HLA: Human Leukocyte antigens
CKD: Chronic Kidney Disease
RAM: RA Microenvironment

REFERENCES

[1] Finckh, Axel. et al. (2022). Global epidemiology of RA. .*Nature Reviews Rheumatology, 18*(10), 591-602.

[2] Alivernini, Stefano, Gary S. Firestein & Iain B. McInnes. (2022). The pathogenesis of RA. *Immunity*, *55*(12), 2255-2270.

[3] Padyukov, Leonid. (2022). Genetics of RA. *Seminars in immunopathology*, *44*(1). Berlin/Heidelberg: Springer Berlin Heidelberg,.

[4] Jang, Sunhee, Eui-Jong Kwon & Jennifer Jooha Lee. (2022). RA: Pathogenic roles of diverse immune cells. *International Journal Of Molecular Sciences*, *23(*2), 905.

[5] Mrid, Reda Ben, et al. (2022). Anti-rheumatoid drugs advancements: New insights into the molecular treatment of RA. *Biomedicine & Pharmacotherapy, 151*, 113126.

[6] Totlibayevich, Yarmatov Suvon, et al. (2022). Risk factors for kidney damage in RA. *Texas Journal of Medical Science*, *13*, 79-84.

[7] Rustamovich, Toirov Doston, et al. (2022). Gastrointestinal conditions in RA patients. *Texas Journal of Medical Science*, *15*, 68-72.

[8] Komatsu, Noriko & Hiroshi Takayanagi. (2022). Mechanisms of joint destruction in RA—immune cell–fibroblast–bone interactions. *Nature Reviews Rheumatology, 18*(7), 415-429.

[9] Zhu, Yan, et al. (2022). RA microenvironment insights into treatment effect of nanomaterials. *Nano Today, 42*, 101358.

[10] Kmiołek, Tomasz & Agnieszka Paradowska-Gorycka. (2022). miRNAs as biomarkers and possible therapeutic strategies in RA. *Cells 11*(3), 452.

[11] Frazzei, Giulia, et al. (2022). Prevention of RA: A systematic literature review of preventive strategies in at-risk individuals. *Autoimmunity Reviews*, 103217.

[12] Prasad, Peeyush, et al. (2023). RA: Advances in treatment strategies. *Molecular and Cellular Biochemistry*, *478*(1), 69-88.

[13] Gorantla, Srividya, et al. (2022). Emerging trends in microneedle-based drug delivery strategies for the treatment of RA. *Expert Opinion on Drug Delivery*, *19*(4), 395-407.

[14] Li, Chunhong, et al. (2022). Recent progress in therapeutic strategies and biomimetic nanomedicines for RA treatment. *Expert Opinion on Drug Delivery*, *19(*8), 883-898.

[15] C, R., V. V, K. K, S. H & P. S. E. (2022). In-cabin radar monitoring system: detection and localization of people inside vehicle using vital sign sensing algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication*, *10(*8), 104-9. DOI: 10.17762/ijritcc.v10i8.5682.

[16] V. V, R. A. C, S. B. M, A. Kumari P, V. S. Reddy R & S. Murthy R. (2022). Custom hardware and software integration: bluetooth based wireless thermal printer for restaurant and hospital management. *IEEE 2nd Mysore Sub Section International Conference (MysuruCon),*

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.3

Mysuru, India, pp. 1-5. DOI: 10.1109/MysuruCon55714.2022.9972714.

[17] V. V, R. A. C, V. S. R. R, A. K. P, S. M. R & S. B. M, Implementation of IoT in agriculture: A scientific approach for smart irrigation. *IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India*, pp. 1-6. DOI: 10.1109/MysuruCon55714.2022.9972734.

[18] Viswanatha, V. & R. Venkata Siva Reddy. (2017). Digital control of buck converter using arduino microcontroller for low power applications. *International Conference On Smart Technologies For Smart Nation (SmartTechCon). IEEE*.

[19] Viswanatha, V., Venkata Siva Reddy & R., Rajeswari. (2020). Research on state space modeling, stability analysis and pid/pidn control of dc–dc converter for digital implementation. In: *Sengodan, T., Murugappan, M., Misra, S. (eds) Advances in Electrical and Computer Technologies. Lecture Notes in Electrical Engineering, 672*. Springer, Singapore. https://doi.org/10.1007/978-981-15-5558-9_106.