ARTIFICIAL INTELLIGENCE BASED PREDICTIVE MODEL TO PREVENT DIALYSIS IN PATIENTS WITH EARLY KIDNEY DISEASE

PROJECT OBJECTIVE:

Chronic kidney disease (CKD) is a significant public health concern affecting people worldwide. According to a survey by National Institute of Health – India, there are approximately 1,75,000 patients who undergo dialysis all over India and this data was collected in 2018. The numbers would have risen up exponentially by now. When it comes to world statistics, according to International Society of Nephrology, between 5.3 to 10.5 million patients require dialysis or a transplant in order to survive chronic kidney disease. And over 2,22,000 patients are waiting for kidney transplant in India. Total population of people who suffer from chronic kidney disease in India is nearly 140 million i.e., around ten percent of the whole Indian population suffers this disease and all over the world it peaks by 800 million.

With right treatments, methodologies and resources, a huge ratio of the world population can be saved from reaching death. The resources are very limited and treatments and methodologies are narrowed down due to many anomalies. Some medicines work for certain patients and doesn't work for the other and the reasons are still unrecognisable. The huge problem and challenge to the doctors when it comes to treating CKD is to give out appropriate treatments to the right patient.

The ability to accurately predict disease progression or regression in CKD patients is crucial for personalized treatment planning, optimizing resource allocation, and improving patient outcomes. However, existing predictive models for CKD progression often rely on limited variables and have shown suboptimal accuracy. There is a need to leverage advanced machine learning techniques to analyse large and diverse datasets, identify relevant patterns and associations, and develop a predictive model that can accurately forecast the likelihood of CKD progression or regression in individual patients. Addressing this problem requires the development of a comprehensive predictive model for CKD progression that integrates a wide range of clinical, demographic, and laboratory variables. By providing healthcare providers with accurate predictions, they can make informed decisions, tailor treatment plans, and allocate resources more effectively.

The primary goal is to develop a powerful predictive model that leverages artificial intelligence and machine learning to accurately forecast CKD progression or regression, leading to better patient care, optimized resource utilization, and improved overall CKD management.

TEAM MEMBERS:

1. C.S. Kanimozhi Selvi – Head of the department – AI, KEC.
2. Kaushik B - 3rd year, AI&DS – KEC.
3. Bhavatharini N - 3rd year, AI&DS – KEC.
4. Hemanthh V V - 3rd year, AI&DS – KEC.
5. Hardik Raj B - 3rd year, AI&ML – KEC.

DATA POINTS:

Parameters collected/considered:

1. The most crucial factor is Glomerular Filtration rate (GFR), it is the rate of flow of plasma through the Bowman's capsule in the nephron (structural and functional unit of Kidney). It basically is proportional to the regression of the disease i.e., When GFR of a patient increase, it means that he/she is getting better.
2. Serum creatinine level: Number of milligrams of creatinine to a decilitre of blood. It is inversely proportional to the regression of the disease. More the creatinine level worser the patients get.
3. Haemoglobin Level.
4. Random Blood Sugar: It's the amount of glucose present in the blood in a given point of a day.
5. Total Count: It represents the amount of WBC, RBC and platelets.
6. Urea: It is the nitrogenous waste that is secreted by the body. It gets ejected from the body in the form of urine.
7. Pus cells: Can be a warning of any sort of infections.
8. Epithelial cells: constitute the fundamental cells in the kidney.
9. Age
10. Diabetes: Yes/No
11. Hypertension: Yes/No
12. Family history: Yes/No
13. Previous Kidney disease: Yes/No
14. Cardio Vascular disease: Yes/No
15. Urinary tract infection: Yes/No
16. Renal stone/Kidney stone: Yes/No
17. Kidney injury history: Yes/No
18. Type of worker: either manual or field.
19. Diet: Vegetarian or non-vegetarian.
20. Water intake in Litres
21. Height in centimetres
22. Weight in kilograms
23. BP: Systolic and Diastolic pressure.
24. Gender: Male/Female
25. Medications
26. Injections
27. Urine Albumin level

APPROACHES:

To treat this anomaly, machine learning algorithms can be used where we identify some key notable patterns between different patients and their treatment methods(medication). Data analysis can also be done to extract some valuable insights. For example, which medicine is been taken by the greatest number of patients who are getting better or regressed. With machine learning we can evaluate certain insights, for example, how much Serum creatinine level impacts the stage of CKD that a particular patient is in. With these insights and the huge data, we can derive an AI model that helps with the doctors in analysing the patient's present condition with respect to the previous ones, assisting which medicine can be given to the patient to achieve an improvement, and what are the other factors of a patient that has to tuned in order to get him/her out of the danger zone. Deep learning algorithms such as Time series models can be used to build a trajectory for the future and how healthy the patient will become. Risk factors can also be derived, patients can be classified as either high risk or low risk based on certain diligent criterions. All these will be done in the surveillance and guidance of a SME i.e., a subject matter expert which is, a Nephrologist.

PROPOSED METHODOLOGY:

Data Preprocessing:

Prepare the dataset by collecting all relevant features, including historical patient data (GFR values, lab results, medications, demographics).
Clean the data, handle missing values, and perform feature engineering to extract relevant features.

Data Analysis:

Conduct exploratory data analysis (EDA) to understand the patterns and trends in the data. Identify key features and their relationships to disease progression.

1. Frequency vs development in health

   Overview: There exists an assumption of Low-risk patients being highly visiting patients and High-risk patients being those who does not attend the checkup sessions quite often. To prove that assumption as to check the credibility of it, exploratory data analysis has been done.

   Purpose: If found to be proven, this fact can help the CKD patients a lot, and doctors would advise the patients to attend their checkups more often and how far it can impact their diagnosis and prevention of disease progressing into dialysis stage.

   Methodology: The assumption was proven by analysis methods by finding the ratio between number of visits and total number of months, and it clearly proves our assumption that Low risk patients often visit the hospital when compared to those who are at high risk.

First the patient's first visited date is taken and it is subtracted with the current date (today's date) to count the number of days the particular patient has been a CKD patient. From that, number of months is calculated by dividing the number of days by 30. (assuming each month has 30 days) So now we have got the number of months the patient has been as CKD patient under the hospital records. Now we calculate the number of visits i.e., the total number of times the patients had checked up with the hospital. Now the ratio of both the values gives us the frequency in the range (0,1). Now inverting the found value will give out the number of months per visits i.e., how many months does it take for a patient to hit the hospital or how often he/she does a checkup and multiplying the found value by 30 will give number of days per visit.

2. Stages vs Glomerular Filtration Rate:

Patients were classified as Low risk and high risk on basis of two different methodologies, one is just with GFR threshold barrier that is when the difference between the initial GFR and the latest GFR is greater than or equal to 15. Another one is blending in the idea of stages deviations.
Overview:
The stage labeling function serves to categorize Glomerular Filtration Rate (GFR) values into different stages of Chronic Kidney Disease (CKD) based on established clinical guidelines. This documentation outlines the purpose, methodology, and application of the stage labeling function.

Purpose:
The stage labelling function aims to categorize GFR values into distinct stages of CKD to facilitate clinical assessment, treatment planning, and disease management. By assigning each GFR value to a specific stage, healthcare providers can monitor disease progression, assess treatment effectiveness, and make informed decisions regarding patient care.

Methodology: Defining CKD Stages:

The function defines five stages of CKD based on GFR values, as per clinical guidelines:

Stage 1: GFR ≥ 90
Stage 2: $60 \leq$ GFR ≤ 89
Stage 3: $30 \leq$ GFR ≤ 59
Stage 4: $15 \leq$ GFR ≤ 29
Stage 5: GFR <=; 15
Stage Label Assignment:
The function evaluates each GFR value and assigns it to the corresponding CKD stage based on predefined criteria using conditional statements.

Interpretation:

The Stage column in the Data frame represents the CKD stage corresponding to each GFR value. This information can be utilized for clinical assessment, treatment planning, and monitoring disease progression in CKD patients.

Risk Classification Documentation

Overview:

The risk classification process aims to categorize patients based on their risk of disease progression or adverse outcomes in chronic kidney disease (CKD). This documentation outlines the purpose, methodology, and application of the risk classification process.

Purpose:

The primary purpose of the risk classification process is to identify patients who are at high risk of disease progression or adverse outcomes in CKD. By categorizing patients based on predefined risk thresholds and clinical criteria, healthcare providers can prioritize interventions, allocate resources, and tailor treatment plans to mitigate risks and improve patient outcomes.

Methodology:

Risk Threshold Definition:

The risk threshold (risk_threshold) is a predefined value that determines the threshold for identifying high-risk patients. This threshold value is set based on clinical judgment, expert recommendations, or established guidelines.

Patient Grouping:

The patient dataset is grouped by patient name (grouped) to analyse each patient's longitudinal data over multiple visits.

Risk Assessment Criteria: The risk classification process evaluates each patient's risk based on the following criteria:

Presence of Stage 4 or Stage 5 CKD: Counts the number of visits where the patient is categorized as Stage 4 or Stage 5 CKD.

Difference in Glomerular Filtration Rate (GFR): Calculates the absolute difference in GFR values between the first and last visits for each patient.

Risk Classification:

Patients are classified as high risk if they meet one of the following conditions:

The combined counts of Stage 4 and Stage 5 CKD exceed the specified risk threshold (risk_threshold) as a percentage of total visits.

The difference in GFR values between the first and last visits exceeds a predefined threshold value (e.g., 15 units).

Patients not meeting these criteria are classified as low risk.

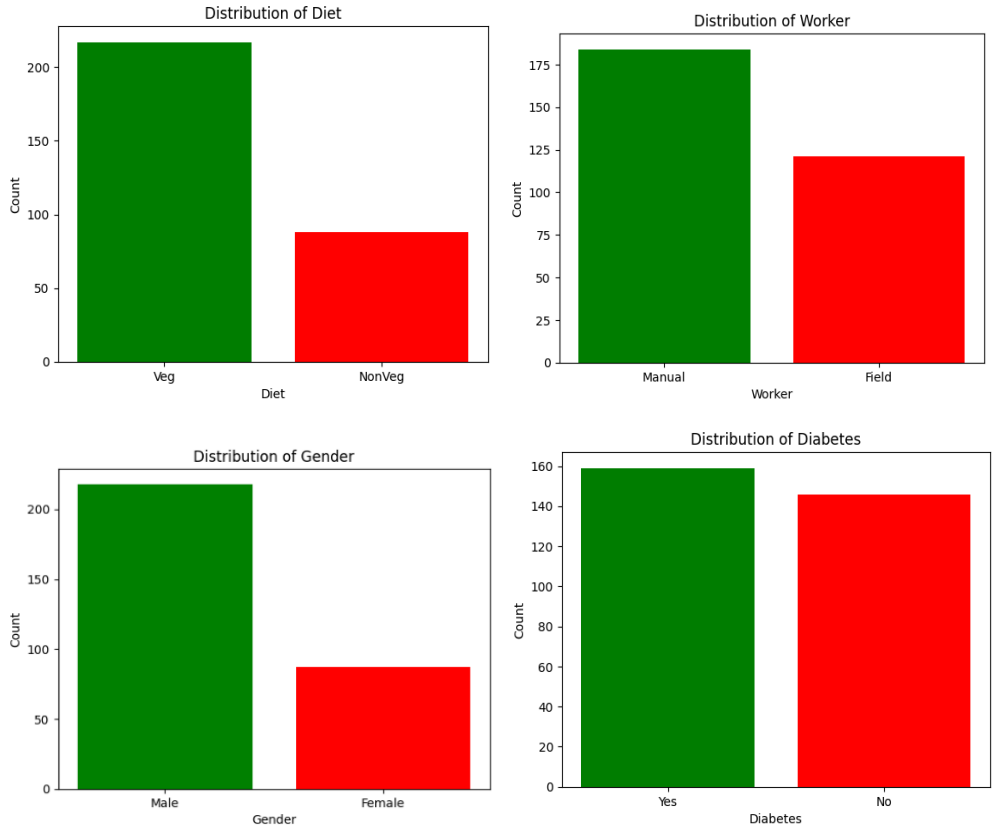
Application:

The risk classification process is applied as follows:

Data Input: Patient data, including GFR values and CKD stage classifications, is inputted into the risk classification process.

Risk Assessment: The risk classification process evaluates each patient's data based on predefined risk criteria.
Output: The output of the risk classification process is a categorical assignment of risk factors (high risk or low risk) for each patient.
Interpretation:
The risk classification process provides healthcare providers with valuable insights into patient's risk profiles for disease progression or adverse outcomes in CKD. By categorizing patients based on their risk factors, healthcare providers can implement targeted interventions, closely monitor high-risk
patients, and optimize patient management strategies to improve outcomes.

Conclusion:
The risk classification process is an essential component of CKD management, enabling healthcare providers to identify patients at high risk of disease progression or adverse outcomes. By stratifying patients based on their risk profiles, healthcare providers can implement proactive interventions and personalized treatment plans to optimize patient care and improve long-term outcomes in CKD.



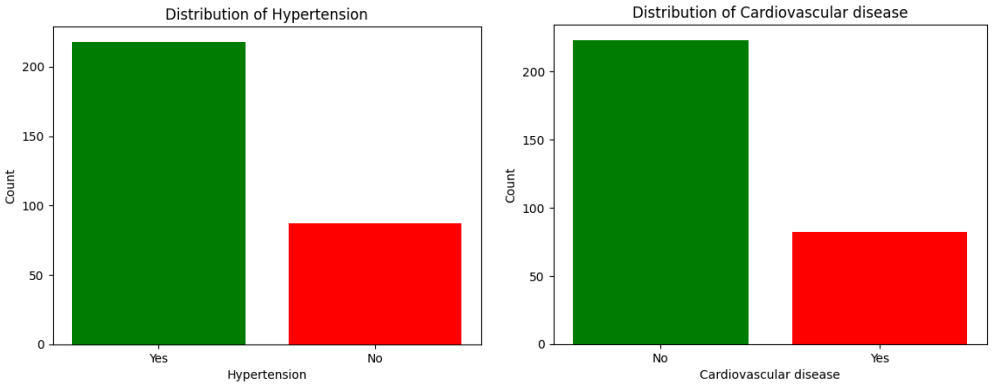


Fig 1. The representation of some important attributes in graphical form (Bar plots)

3. Mean GFR deviation analysis:
   The provided Python script conducts an in-depth analysis to determine the average mean deviation of Glomerular Filtration Rate (GFR) for each patient in the dataset. The script iterates through patient records, organizing data chronologically based on hospital visit dates and computing yearly deviations in GFR values. By calculating the mean deviation for each patient across multiple years, the script generates comprehensive dataframe summarizing the average mean deviation for all patients.

   Threshold Value Calculation:
   Additionally, the script computes the threshold value for GFR deviation using percentile analysis. Percentiles are statistical measures that divide a dataset into equal parts, with each part representing a certain percentage of the data. In this context, the 95th percentile represents the value below which 95% of the GFR deviations fall. By quantifying the 95th percentile of the average mean deviation values, the script establishes a critical benchmark for evaluating the severity of GFR deviations among patients.

   Percentile Values:
   Moreover, the script calculates percentile values for a broader understanding of the distribution of GFR deviations within the dataset. Percentiles such as the 25th, 50th (median), 75th, and 90^{th} percentiles provide insights into the spread and central tendency of GFR deviations. These percentile values serve as valuable reference points for assessing the variability and severity of GFR deviations among patients, guiding healthcare providers in identifying individuals requiring closer monitoring or intervention based on their deviation from established norms.

   Indeed, the analysis reveals that the 95th percentile value for Glomerular Filtration Rate (GFR) deviation among chronic kidney disease (CKD) patients is 4.4. This signifies that 95% of the patients in the dataset exhibit a GFR deviation of 4.4 or less, serving as a critical threshold for evaluating the severity of GFR deviations.

Furthermore, percentile values such as the 25th, 50th, 75th, and 90th percentiles provide additional insights into the distribution of GFR deviations, offering a comprehensive understanding of the variability and central tendency within the patient population.
Armed with this knowledge, healthcare providers can effectively tailor treatment strategies and interventions to address GFR deviations and optimize CKD management, ultimately improving patient outcomes.

4. Exploratory data analysis:

All the important key features of all the existing unique patients in the database has been extracted and projected to the Medical experts/nephrologists in the application as a separate module called "Analysis" module. This module contains the analysis report as in the form of a table and it represents some of the key values. It comprises of some attributes like:

1. Gender,
2. Age,
3. Diabetes (Yes/No),
4. Hypertension (Yes/No)

And some more important statistical attributes such as

5. GFR First (GFR value noted down in the first ever visit),
6. GFR last (GFR value noted down in the most latest visit),
7. SCR First (Serum creatinine rate noted down in the first ever visit),
8. SCR Last(Serum creatinine rate noted down in the most latest visit),
9. Change in GFR (GFR last – GFR first)

And regarding the frequency of visits:

10. Total number of visits.
11. Months (number of months the patient has been as CKD patient under the hospital records)
12. Frequency of visits.

Purpose:

The doctor should know about the basic details about the patient with respect to other major diseases like diabetes and hypertension which are very diligent when it comes to CKD diagnosis and medicine recommendations

And with the help of GFR and serum creatinine values provided the doctor can note down the trajectory in which the patient is travelling and how far he/she has come in his/her way of diagnosis.

And the doctor can also easily filter out Low-risk and high-risk patients which is a major part of recommendations and suggestions.

Patient ID	Age	Gender	Diabetes	Hypertension	Serum Creatinine First	GFR first	Months(Number of months from his/her first visited date)	Total Number of visits	Frequency (Count/Number of Visits)	Serum Creatinine Latest Visit	GFR last	Change in GFR(GFR last - GFR first)
1	23	1	0	0	1.4	95	14	6	0.429	1.2	65	-30
2	71	1	0	1	3.4	29	169	44	0.26	2.9	14	-15
3	58	1	0	1	4.4	55	26	15	0.577	4.6	38	-17
4	69	1	0	1	1.7	26	123	24	0.195	1.5	46	20
5	58	0	1	0	2.2	17	91	50	0.549	2.5	7	-10
6	59	0	1	1	1.9	31	71	23	0.324	1.7	36	5
7	25	0	0	1	2.1	30	132	39	0.295	3.1	27	-3
8	52	1	1	1	3	35	90	23	0.256	4.3	30	-5
9	35	1	0	1	2	44	41	13	0.317	1.8	30	-14
10	80	1	0	1	4.1	37	28	10	0.357	3.9	43	6
11	62	0	1	1	1.7	14	8	9	1.125	1.8	13	-1
12	67	1	1	1	2.1	8	10	7	0.7	2.3	13	5
13	55	1	1	1	14.1	22	13	5	0.385	2.4	23	1
14	38	1	1	1	3	43	77	28	0.364	4.4	39	-4
15	52	1	1	1	1.8	16	69	23	0.333	3	13	-3

TABLE: Exploratory data analysis

Note: This is a sample data for the first 15 patient in the existing database and it represents a static dataset whereas in the application the data is generated dynamically with the help of a pipeline that invokes whenever a new data is added hence attesting that the analysis would alter dynamically with each data entry.

Feature Selection:

Choose relevant features that are likely to influence disease progression. These include age, serum creatinine levels, GFR, comorbidities, and more.

To identify the most impactful feature out of all the attributes collected, Feature selection was done using five different approaches namely Mutual Info classification, Feature analysis using Random Forest, Lasso Regression, ANOVA, LightGBM. The best features are selected and decided by intersecting all the top features from all the five approaches.

Mutual Information measures how much information about the target variable is gained by knowing the value of a particular feature. Higher Mutual Information values indicate a stronger association between the feature and the target variable. Mutual Information is a non-negative measure, where higher values indicate more information shared between the

feature and the target. Random Forests can assign importance scores to each feature in the dataset. These scores indicate the contribution of each feature to the predictive performance of the model. Higher importance scores suggest that a feature is more influential in making accurate predictions. Lasso Regression, a variant of linear regression, can be used for feature analysis and selection by introducing a regularization term that encourages sparsity in the model. The regularization term in Lasso Regression is the absolute value of the coefficients, and it tends to shrink less important features toward zero, effectively excluding them from the model. Analysis of Variance (ANOVA) is a statistical technique used to analyse the differences among group means in a sample. While ANOVA itself doesn't directly provide feature importance scores like some machine learning algorithms, it can be used in the context of feature selection or understanding the significance of different features in explaining variability in a response variable. The basic idea behind feature importance in LightGBM is that during training, the algorithm evaluates the effectiveness of each feature in making splits that improve the model's performance. Features that contribute more to the reduction in loss or impurity are considered more important.

Top 10 features are as follows:

Table 1. Top ten features which impacts changes in GFR

S.NO	FEATURE
1	Haemoglobin Level
2	Urea
3	Serum Creatinine level
4	Age
5	BMI
6	Gender
7	Random Blood Sugar
8	Total Count
9	Diabetes
10	Total Water Intake

Data Splitting:

Split your dataset into a training set and a testing set. The training set is used to build the simulation model, and the testing set is used to evaluate the model's performance.

DATA SCIENCE TECHNOLOGIES:

Initially right after the data is collected, it is undergone a vast preprocess. Preprocessing the data includes noise removals and treating the missing values. It is done either manually or through getting the data into user defined algorithms. Data visualisations such as pie plots, bar plots and many such visuals are created in Python. For more polished visuals, a business intelligence tool called Power BI is used. Many valuable and game changing insights were gathered from them. Many methodologies in the domain of data science were used such as Aggregation (aggregation refers to the process of combining and summarizing data from

multiple sources or at a lower level of granularity to create a higher-level view. This can involve grouping and computing summary statistics or other aggregative measures to gain insights into patterns and trends within the data. Involves finding sum, mean, count, grouping, minimum and maximum, mode, median, standard deviation etc…), interpolation (Interpolation is a method used in mathematics and data analysis to estimate values that fall between known data points. It involves estimating the value of a function or data point within the range of known values based on the available data.), Label encoding (It involves converting categorical data, which consists of labels or categories, into numerical values. This transformation is particularly useful when working with machine learning algorithms that require numerical input. In label encoding, each unique category or label is assigned a unique integer or numerical value. The assignment is usually done in a way that preserves the ordinal relationship between the categories, if any), Grouping (refers to the process of grouping a dataset based on certain criteria, usually categorical variables, and then applying some form of aggregation or analysis within each group.) were used in pre-processing and descriptive analysis of the data. And the results from the descriptive analysis play a vital role in predictive and prescriptive analyses.

MACHINE LEARNING TECHNOLOGIES:

After the data is been pre-processed and all set to build a model, various machine learning algorithms are tried on with various different customisation and tunings in tons of different approaches, for more insights to be gained, the data was even augmented in two different manners, using Data Augmentation and Transfer learning and subjected to various test runs.

Seven different yet unique models have been tried on, namely Decision tree, Random Forest, Multi Linear Regression, Neural network regressor, Support vector regressor, KNN regressor and Light Gradient Boosting regressor. The above seven are some of the most significantly used Benchmark algorithms in the field of machine learning and deep learning. These algorithms were made to test run with Normal collected raw pre-processed data and Augmented data, so by far fourteen different approaches were made in order to identify the best working model for our use case.

DECISION TREE REGRESSOR:

A Decision Tree Regressor is a supervised machine learning algorithm used for regression tasks. It constructs a tree-like model by partitioning the feature space based on decision rules inferred from the data. At each node, it selects the feature and threshold that maximize the homogeneity of the target variable. Predictions are made by traversing the tree from the root to a leaf node, where the average of target values determines the predicted value.

SUPPORT VECTOR REGRESSOR:

A Support Vector Regressor (SVR) is a supervised machine learning algorithm used for regression tasks. It works by finding a hyperplane in a high-dimensional feature space that best fits the training data while minimizing the error margin. SVR is effective for handling non-linear relationships in data and is particularly useful when dealing with small to medium-sized datasets. The algorithm aims to maximize the margin between the hyperplane and the closest data points, known as support vectors, to generalize well to unseen data.

LIGHT GBM REGRESSOR:

Gradient Boosting Regressor is a supervised machine learning algorithm used for regression tasks. It builds an ensemble of decision trees sequentially, where each subsequent tree corrects the errors of the previous ones. The algorithm minimizes a loss function, typically the mean squared error, by adding decision trees that approximate the negative gradient of the loss function. Gradient Boosting Regressor is known for its high predictive accuracy and robustness against overfitting, making it popular for various regression problems across different domains.

MULTILINEAR REGRESSION:

Multilinear regression, also known as multiple linear regression, is a statistical method used to model the relationship between multiple independent variables (features) and a single dependent variable (target). It extends the concept of simple linear regression to scenarios where there are two or more independent variables influencing the dependent variable. The model assumes a linear relationship between the independent variables and the dependent variable, with coefficients representing the strength and direction of these relationships. Multilinear regression is commonly used for prediction and inference in various fields such as economics, social sciences, and engineering, where multiple factors affect an outcome of interest.

NEURAL NETWORK REGRESSOR:

A Neural Network Regressor is a type of artificial neural network (ANN) used for regression tasks. It consists of interconnected nodes (neurons) organized in layers, including an input layer, one or more hidden layers, and an output layer. Each neuron applies an activation function to the weighted sum of its inputs to produce an output. During training, the neural network learns the optimal weights and biases that minimize a predefined loss function, typically the mean squared error, between predicted and actual target values. Neural Network Regressors are capable of capturing complex non-linear relationships in data and are widely used in various fields, including finance, healthcare, and image processing, for regression tasks involving continuous target variables.

RANDOM FOREST REGRESSOR:

A Random Forest Regressor is a machine learning algorithm used for regression tasks. It is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction of the individual trees for regression tasks. Each decision tree is trained on a random subset of the training data and a random subset of the features. Random Forest Regressor combines the predictions of multiple decision trees to improve predictive accuracy and reduce overfitting. It is known for its robustness, ability to handle high-dimensional data, and resistance to overfitting compared to individual decision trees. Random Forest Regressor is widely used in various domains, including finance, healthcare, and bioinformatics, for regression tasks involving continuous target variables.

K NEAREST NEIGHBOUR REGRESSOR:

K-Nearest Neighbours (KNN) Regressor is a supervised machine learning algorithm used for regression tasks. Unlike KNN Classifier, which is used for classification tasks, the KNN Regressor predicts a continuous target variable based on the values of its nearest neighbours in the feature space. The key parameters involve Number of neighbours, distance metric, weights.

MSE:

MSE, or Mean Squared Error, is a measure of the average squared difference between predicted and actual values in a dataset. It is commonly used as a loss function in regression tasks to assess the performance of predictive models. Lower MSE values indicate better model accuracy.

RMSE:

RMSE stands for Root Mean Squared Error. It is a metric used to evaluate the performance of regression models, similar to Mean Squared Error (MSE), but it provides the measure of the standard deviation of the residuals. RMSE is calculated by taking the square root of the average of the squared differences between the predicted values and the actual values in a dataset. It is commonly used to assess the accuracy of predictive models, with lower RMSE values indicating better model performance.

Table 2. Results of Test automations for Predictor Module

SNO	ALGORITHM	MEAN ABSOLUTE ERROR
1	DECISION TREE REGRESSOR	8.61133962
2	SUPPORT VECTOR REGRESSOR	10.13738169
3	LIGHT GBM REGRESSOR	8.796332021
4	MULTILINEAR REGRESSION	9.828309051
5	KNN REGRESSOR	9.515449775
6	NEURAL NETWORK REGRESSOR	10.30775364
7	DECISION TREE REGRESSOR AUGMENTED	8.702529111
8	RANDOM FOREST REGRESSOR AUGMENTED	7.404983983
9	SUPPORT VECTOR REGRESSOR AUGMENTED	10.1177329
10	LIGHT GBM REGRESSOR AUGMENTED	8.663401425
11	MULTILINEAR REGRESSION AUGMENTED	8.659412644
12	KNN REGRESSOR AUGMENTED	9.238786563
13	NEURAL NETWORK REGRESSOR AUGMENTED	9.881999047
14	RANDOM FOREST REGRESSOR	7.094858465

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	5.805439	121
1	30-40	8.629586	234
2	40-50	7.184369	676
3	50-60	7.395435	1220
4	60-70	6.788107	1507
5	70-90+	6.890408	655

Fig 2.1. Mean absolute Error and number of training points with respect to age category by Random Forest

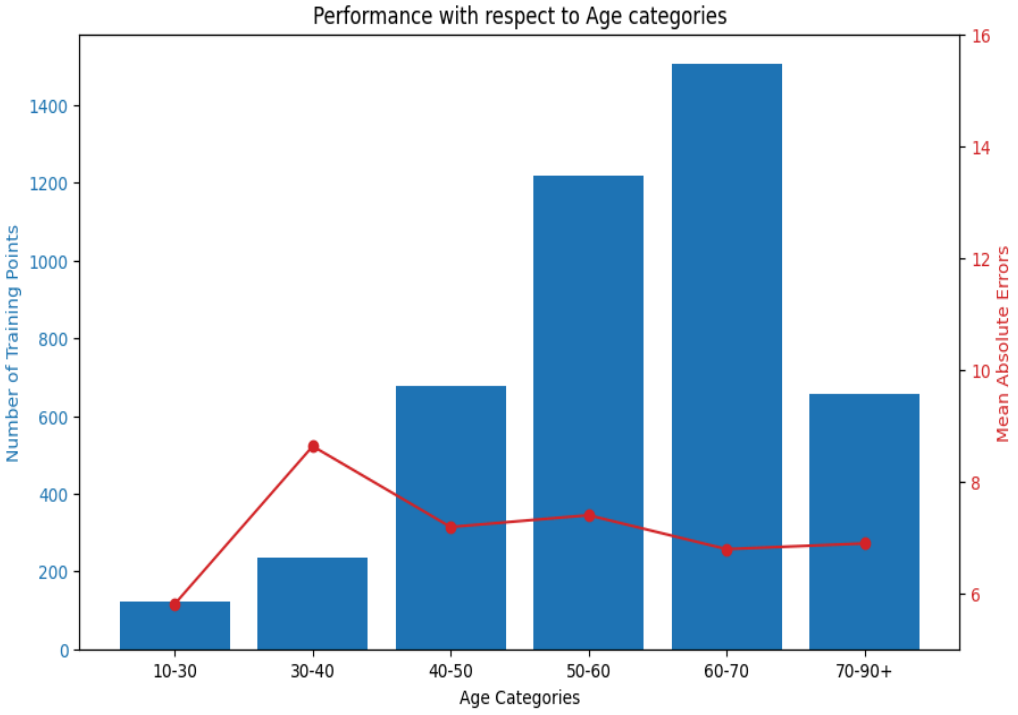


Fig 2.2. The above data in fig 2 is represented in a bar and line plot form

Fig 3.1, 3.2: Decision tree regressor:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	7.317544	121
1	30-40	10.212008	234
2	40-50	9.254814	676
3	50-60	7.879249	1220
4	60-70	8.700487	1507
5	70-90+	9.237670	655

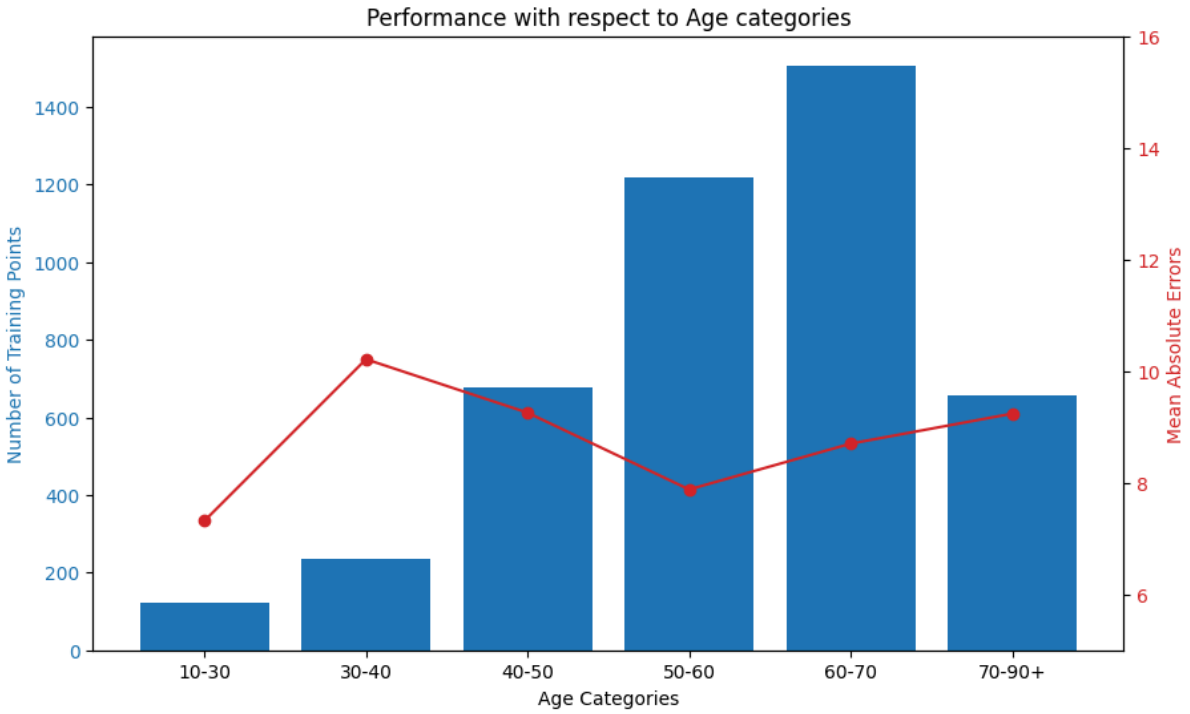


Fig 4.1, 4.2: Support Vector regressor:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	15.240071	121
1	30-40	11.335178	234
2	40-50	10.161501	676
3	50-60	10.082920	1220
4	60-70	9.478051	1507
5	70-90+	10.019018	655

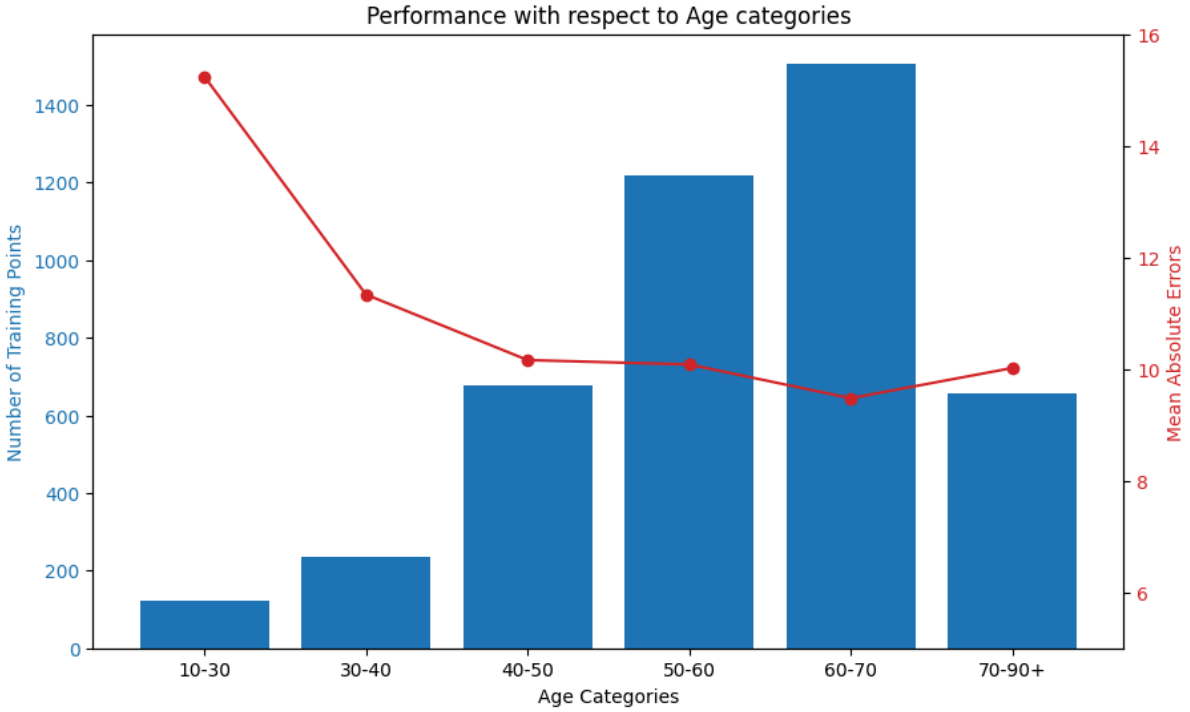


Fig 5.1, 5.2: Light GBM regressor:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	6.604891	121
1	30-40	7.137198	234
2	40-50	8.753790	676
3	50-60	9.705291	1220
4	60-70	8.727158	1507
5	70-90+	8.020608	655

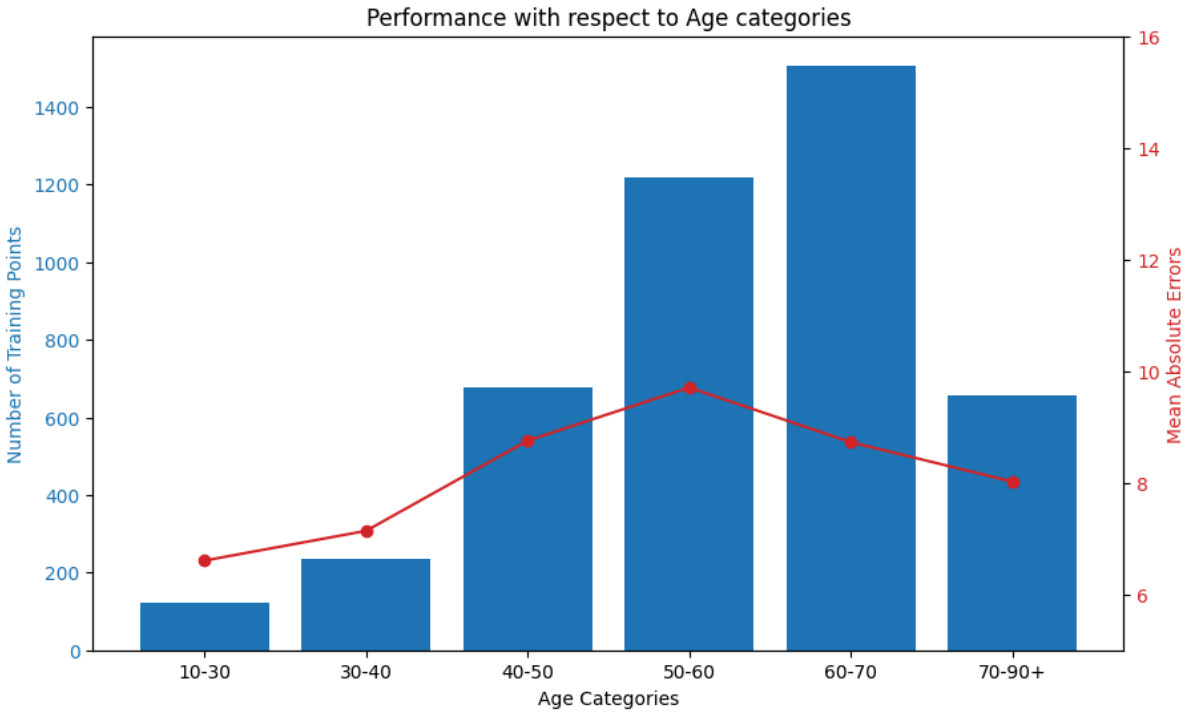


Fig 6.1, 6.2: Multi linear regressor:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	10.921361	121
1	30-40	8.329616	234
2	40-50	9.633857	676
3	50-60	10.530514	1220
4	60-70	9.655924	1507
5	70-90+	8.951133	655

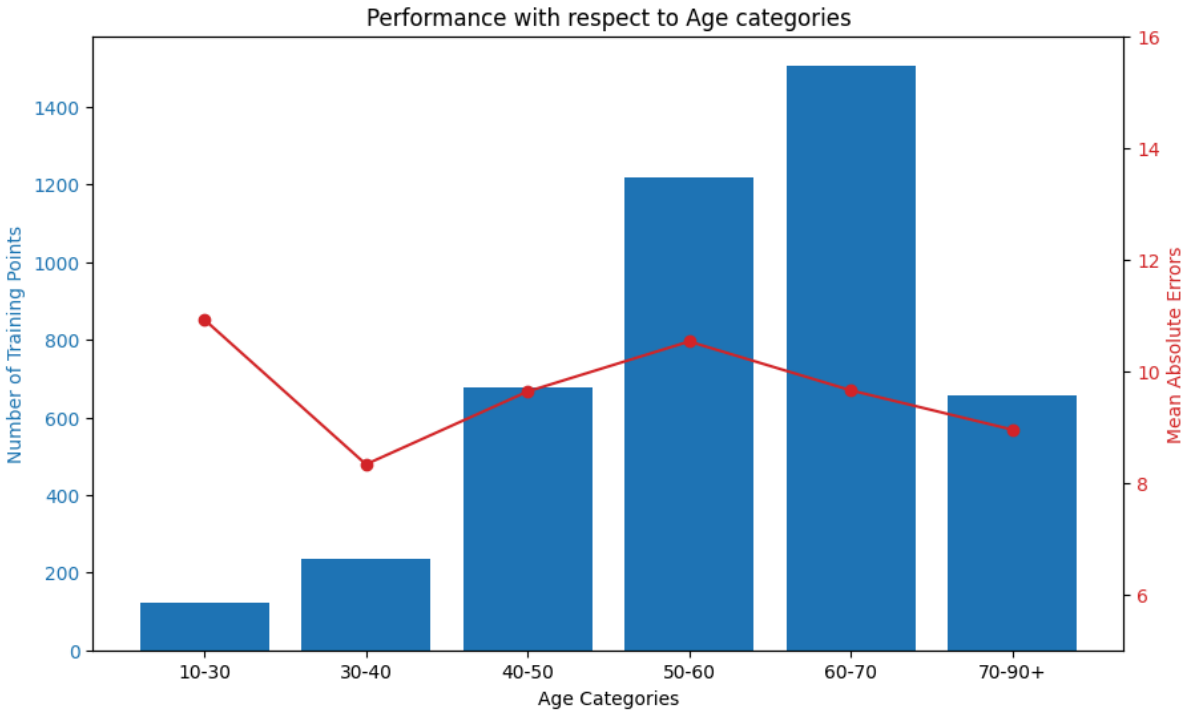


Fig 7.1, 7.2: K nearest neighbour regressor:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	15.535088	121
1	30-40	10.405383	234
2	40-50	9.013555	676
3	50-60	9.271518	1220
4	60-70	9.346360	1507
5	70-90+	8.961097	655

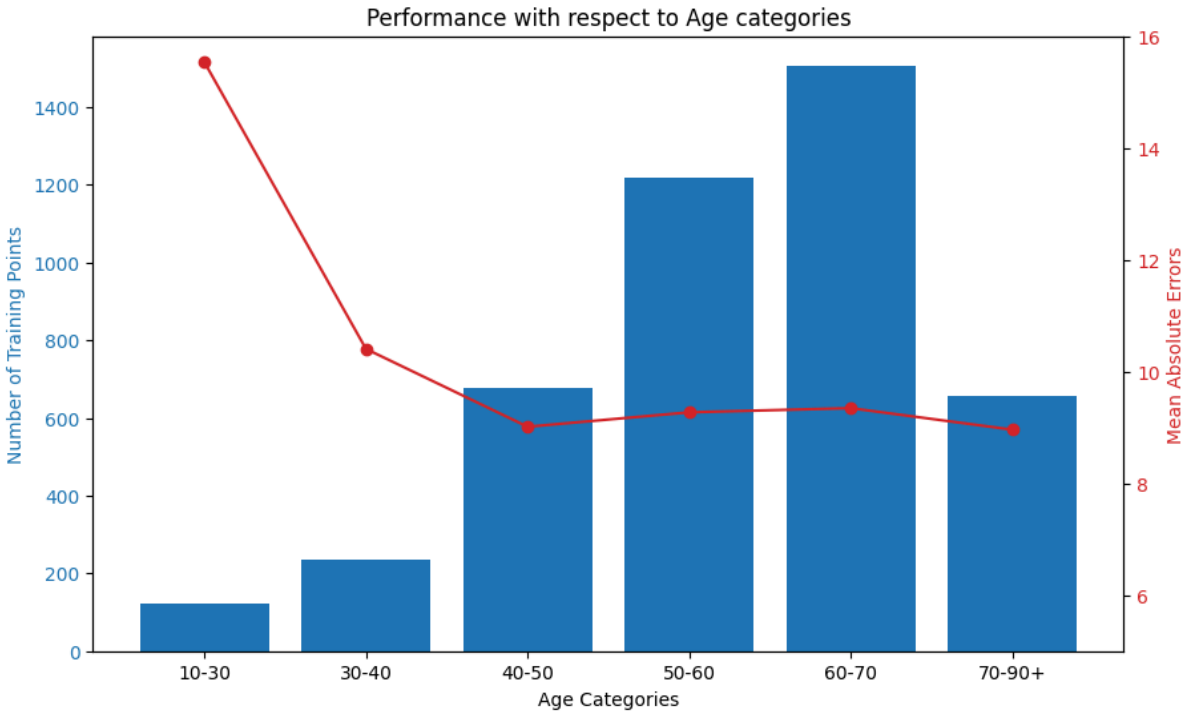


Fig 8.1, 8.2: Neural Network regressor:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	12.647415	121
1	30-40	9.949969	234
2	40-50	10.857242	676
3	50-60	11.100843	1220
4	60-70	9.366992	1507
5	70-90+	9.601566	655

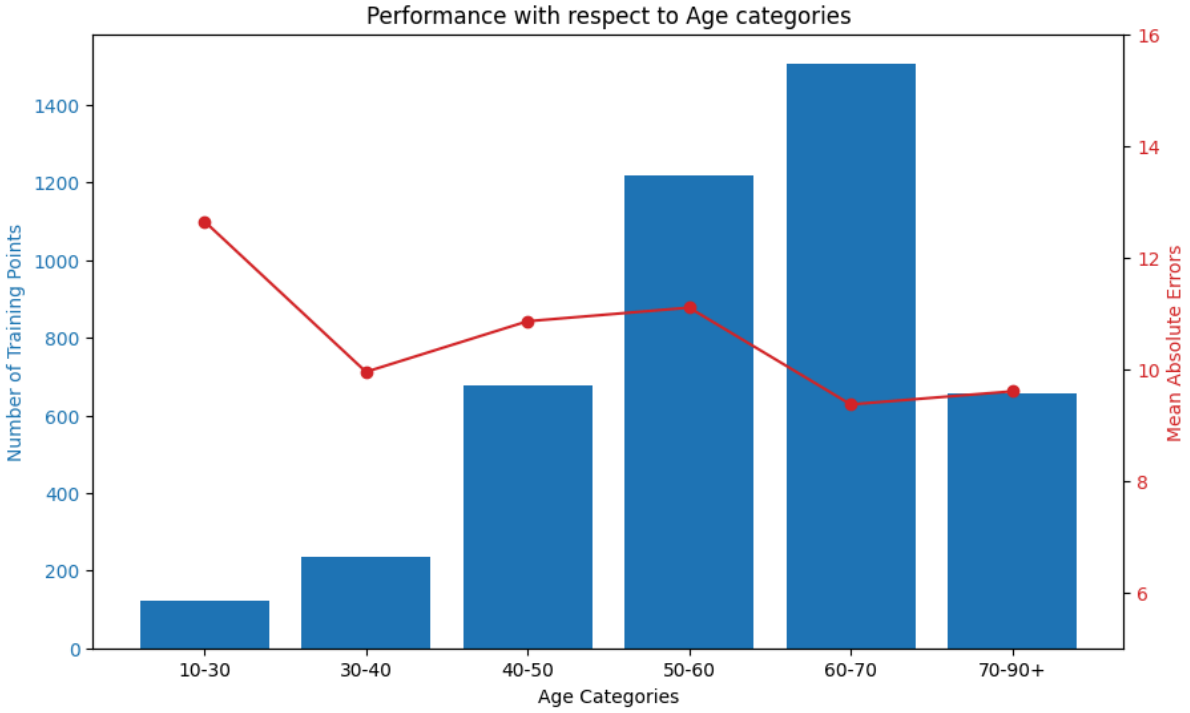


Fig 9.1, 9.2: Decision tree regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	5.689702	121
1	30-40	8.376660	234
2	40-50	9.589515	676
3	50-60	8.527111	1220
4	60-70	8.869691	1507
5	70-90+	8.626474	655

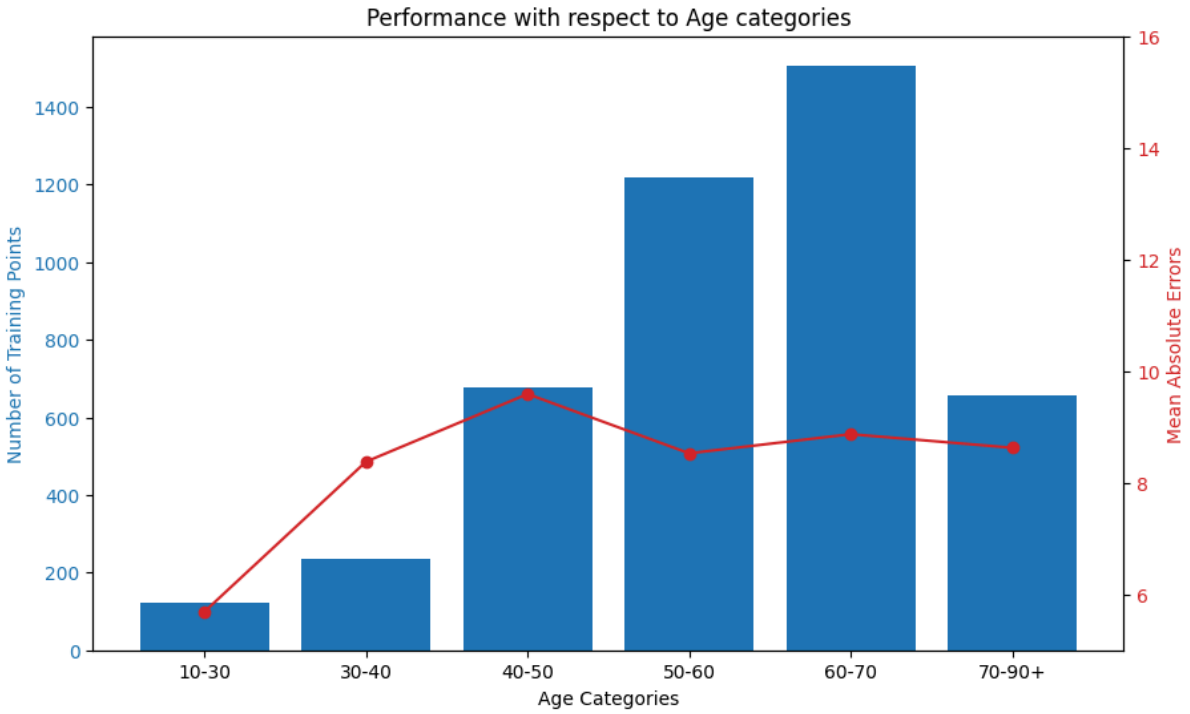


Fig 10.1, 10.2: Support Vector regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	15.320318	121
1	30-40	11.125023	234
2	40-50	9.971105	676
3	50-60	10.021201	1220
4	60-70	9.617543	1507
5	70-90+	9.909376	655

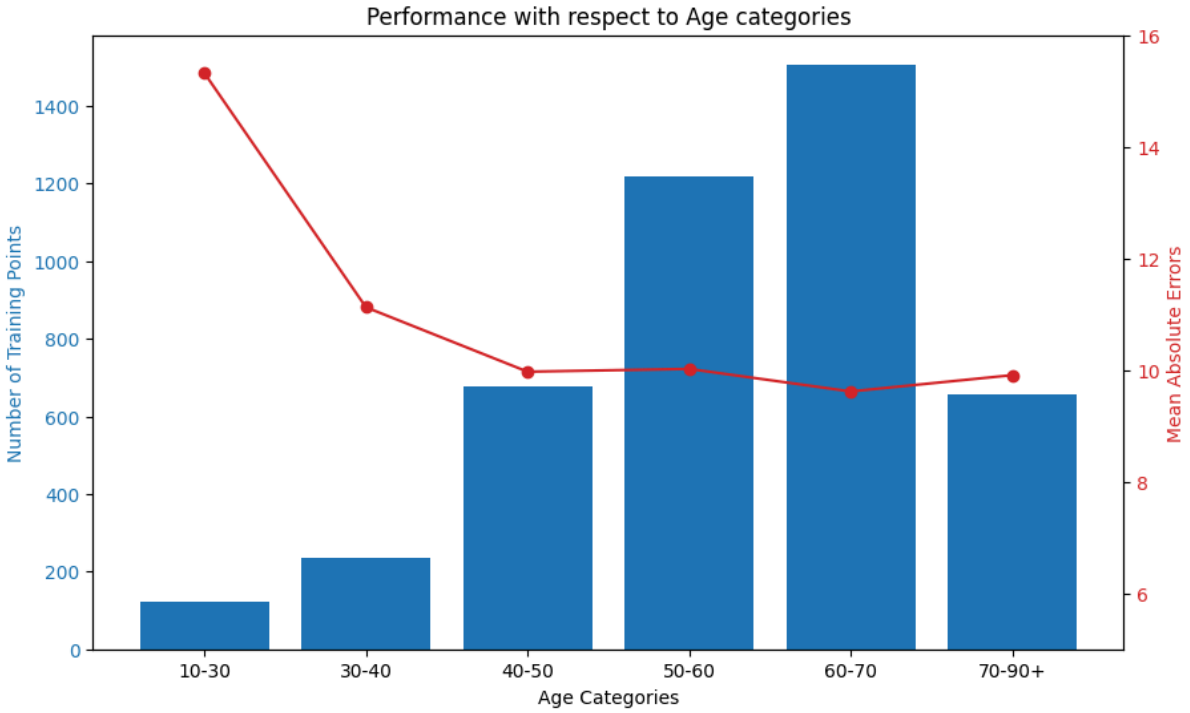


Fig 11.1, 11.2: Light GBM regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	5.183053	121
1	30-40	7.606538	234
2	40-50	8.128517	676
3	50-60	9.973631	1220
4	60-70	8.762432	1507
5	70-90+	7.149019	655

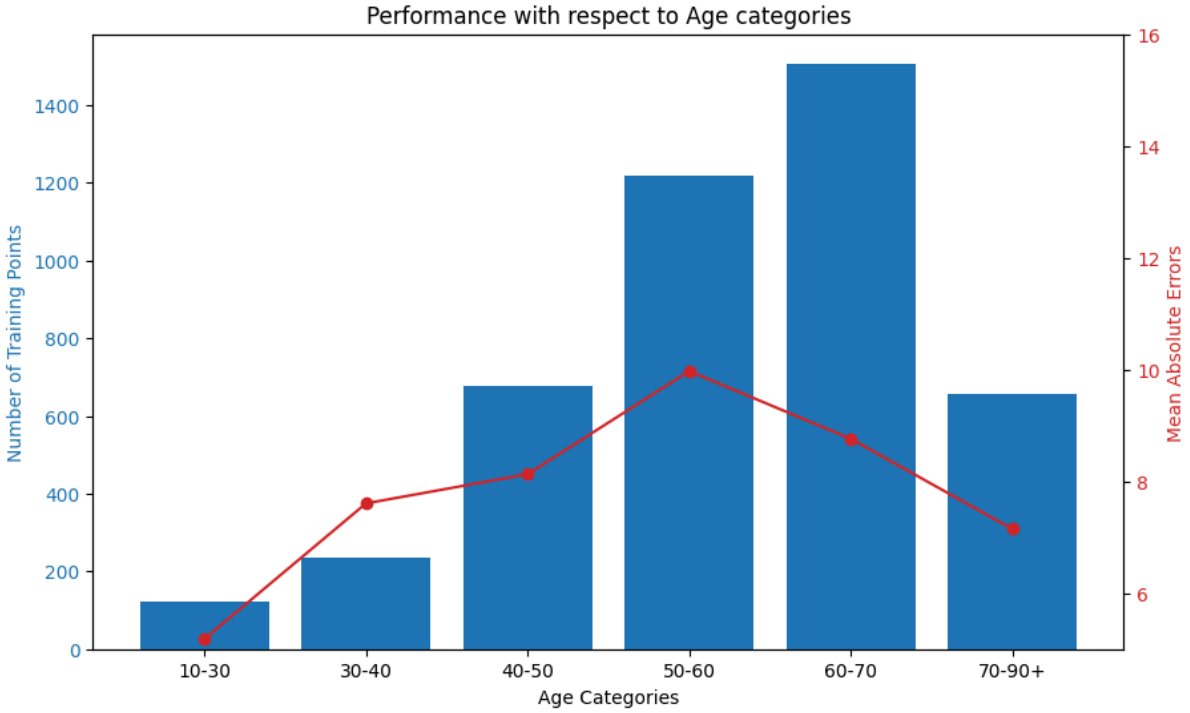


Fig 12.1, 12.2: Multilinear regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	5.079135	121
1	30-40	7.606538	234
2	40-50	8.128517	676
3	50-60	9.973631	1220
4	60-70	8.762432	1507
5	70-90+	7.149019	655

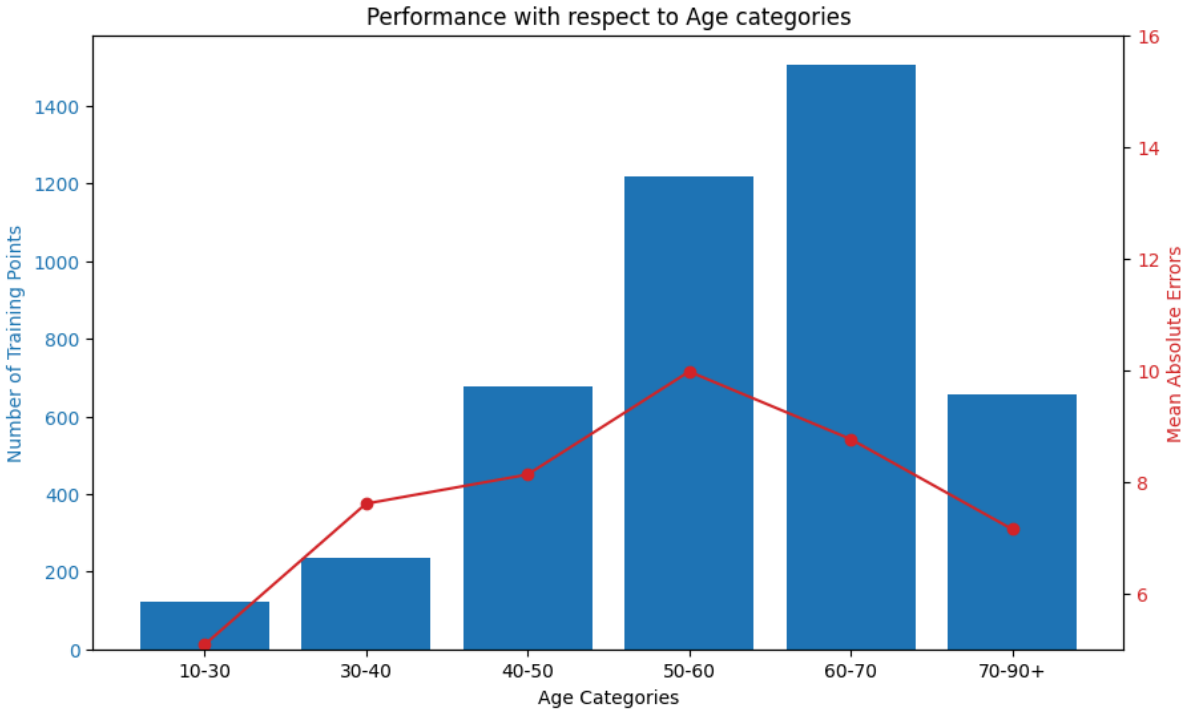


Fig 13.1, 13.2: KNN regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	14.867692	121
1	30-40	9.854093	234
2	40-50	8.014577	676
3	50-60	9.259668	1220
4	60-70	8.893953	1507
5	70-90+	9.721673	655

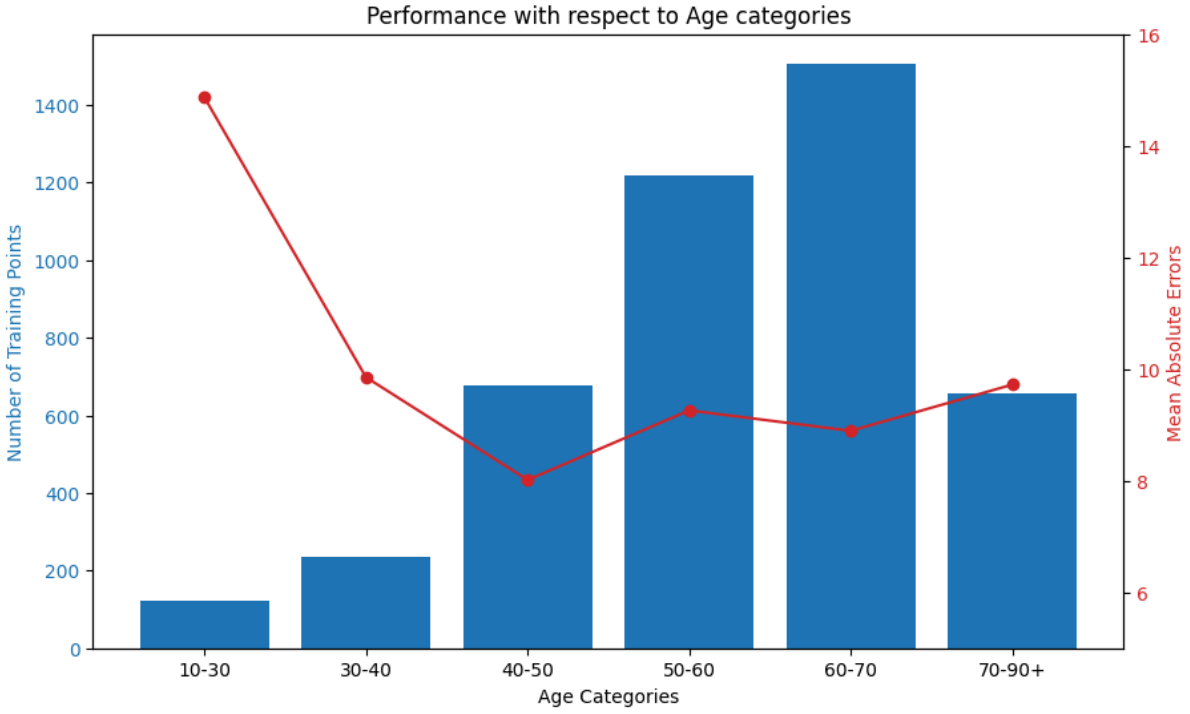


Fig 14.1, 14.2: Neural Network regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	11.878755	121
1	30-40	9.471411	234
2	40-50	9.671225	676
3	50-60	10.499345	1220
4	60-70	9.601203	1507
5	70-90+	8.170056	655

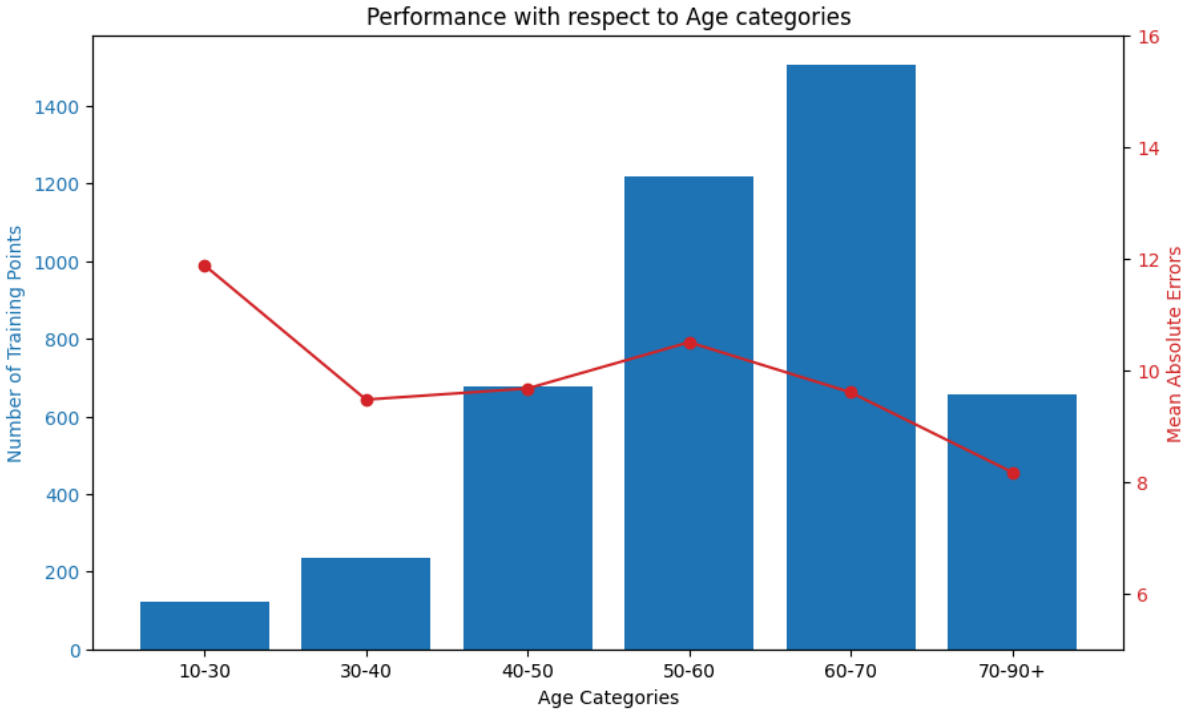
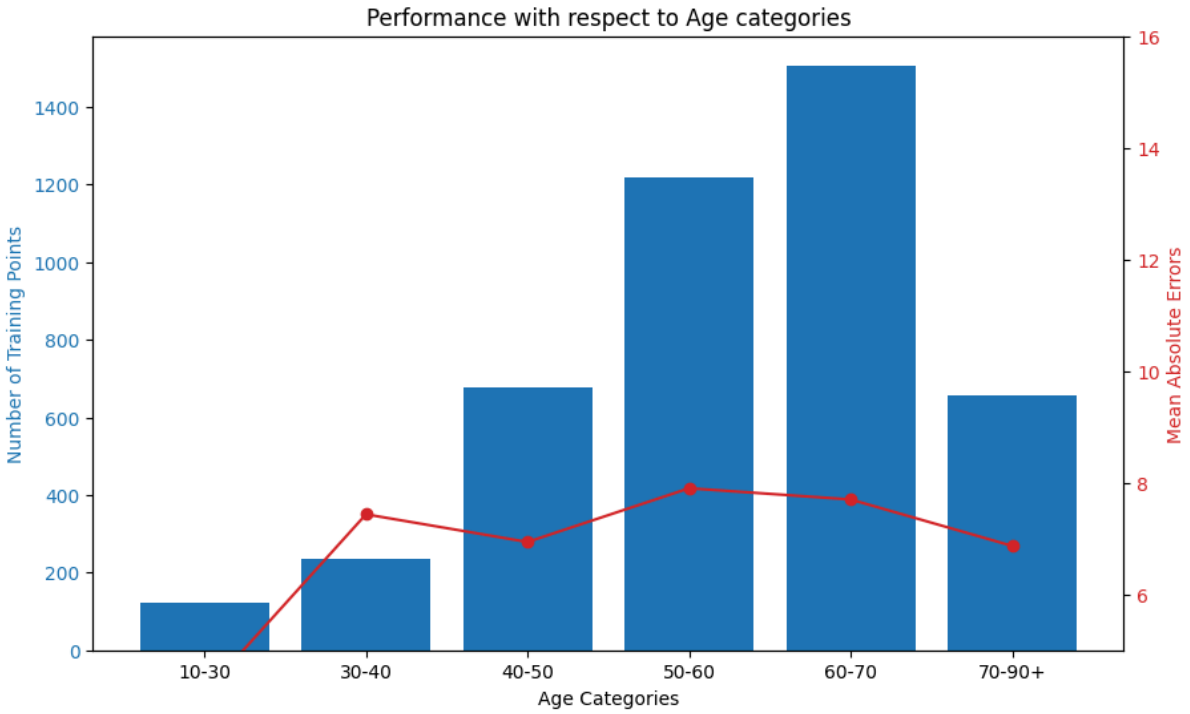


Fig 15.1, 15.2: Random Forest Regressor regressor with Augmented data:

	Age Category	Mean Absolute Error	Number of Training Points
0	10-30	4.229351	121
1	30-40	7.434238	234
2	40-50	6.938209	676
3	50-60	7.900823	1220
4	60-70	7.699894	1507
5	70-90+	6.868051	655



The Random Forest Regressor proves to be an impressive tool in accurately predicting the Glomerular Filtration Rate (GFR) of new patients. Its precision and accuracy are unparalleled, making it a formidable choice for this task. One key indicator of its strong performance is its ability to minimize the Mean Absolute Error (MAE), which measures the average magnitude of prediction errors and highlights the model's accuracy.

Upon closer examination of the MAE across different age categories, interesting patterns in the deviation emerge. As age categories increase, there is a noticeable increase in data anomalies, suggesting a complex interplay of variables that do not follow a straightforward pattern. However, the Random Forest Regressor effectively compensates for these deviations by utilizing an augmented dataset, ensuring that the MAE remains consistent and remarkably stable. This showcases the model's adaptability and resilience in accommodating the intricacies introduced by age-related factors.

The compensation for increased data anomalies, combined with the expansion of the dataset, contributes to maintaining a stable MAE with a slight reduction. This reduction indicates an improved predictive capacity, highlighting the Random Forest Regressor's ability to enhance its accuracy as the dataset grows and account for age-related nuances. In essence, this model not only demonstrates precision in predicting GFR but also exhibits an impressive ability to navigate and mitigate the challenges posed by age-related variations. This solidifies its reputation as a reliable and robust predictive tool in medical applications.

DEEP LEARNING TECHNOLOGIES:

The project dives into the realm of deep learning for forecasting the Glomerular filtration for the existing patients for the consecutive years. This process takes the advantage of existing Time Series Prediction models like LSTM (Long short-term Memory), ARIMA (Auto Regressive Integrated Moving Average) and S-ARIMA.

The two approaches that were made are namely,

- Transfer learning
- Data Augmentation

Transfer Learning:

Transfer learning is a machine learning technique where a model trained on one task is adapted for a related but different task. It involves leveraging knowledge gained from solving one problem to solve a different, but related, problem. By transferring learned representations or knowledge from one domain to another, transfer learning can significantly reduce the amount of labelled data and training time required for the target task. This approach is particularly useful when labelled data is limited or when training from scratch is impractical. Transfer learning has been successfully applied across various domains, including computer vision, natural language processing, and healthcare, to improve model performance and efficiency.

Data Augmentation:

Data augmentation is a technique used to artificially increase the diversity and quantity of training data by applying various transformations to the existing dataset. These transformations include but are not limited to rotation, flipping, cropping, scaling, translation, color jittering, and adding noise. By augmenting the dataset with modified versions of the original data, data augmentation helps improve the robustness, generalization, and performance of machine learning models. This technique is particularly beneficial in scenarios where the available dataset is limited, as it allows for the generation of additional training samples without collecting new data. Data augmentation is widely used in computer vision, natural language processing, and other machine learning tasks to enhance model training and performance.

MODULES:

1. Predictor: Predicting the Glomerular filtration rate for a newly coming patient for the next consecutive years (3)
2. Forecaster: Predicting the Glomerular filtration rate for an existing patient for the next consecutive years (3)
3. Suggester: Suggesting important biological value to be changed for a newly incoming patient in order to improve Glomerular Filtration Rate.
4. Recommender: Suggesting important biological changes to be made and the medicines that needed to be prescribed.

PREDICTOR: GFR PREDICTION FOR THE NEXT THREE CONSECUTIVE YEARS FOR NEWLY INCOMING PATIENTS:

Objective: The primary goal of this project is to develop a predictive model for estimating the Glomerular Filtration Rate (GFR) of dialysis patients for the next 3 years. The model incorporates various patient-related features to enhance the accuracy of predictions.

Model Training:

In the model training phase, several crucial steps were undertaken to ensure the effectiveness of the predictive model. Initially, the dataset, consisting of the mentioned features, was collected comprehensively. Following data collection, a meticulous data preprocessing stage was executed. This involved handling missing values through imputation, addressing outliers, and standardizing or normalizing numerical features to ensure uniform scaling.

Several regression algorithms, including linear regression, polynomial regression, multi linear regression, support vector regression, decision tree regressor, random forest regressor, neural network regressor and light GBM boost regressor were employed during model training. Model evaluation metrics such as Mean Squared Error (MSE) and R-squared were used to assess the performance of each model. MSE quantifies the average squared difference between the predicted values and the actual values in a dataset. In the context of regression analysis, MSE is used to assess how well a model is able to predict continuous outcomes and it is theoretically unbounded and takes any non-negative value that is ranging from zero to infinity. R squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It

ranges from 0 to 1, For example: an R-squared value of 0.75 means that 75% of the variability in the dependent variable is explained by the independent variables, Random Forest Regressor is selected as the best model for predictor.

Post-model training, an additional step was taken to refine prediction accuracy. The inputs provided to the model underwent further processing based on the age of the patients. Specifically, consecutive differences in the numerical data were calculated for each patient's age category. These differences were then added to the corresponding inputs in the dataset.

This approach of taking consecutive differences and incorporating them into the dataset aligns with the goal of capturing age-related trends and variations in the input features. By adapting the model inputs based on age-specific differences, the predictive model becomes more adept at capturing subtle changes in health parameters associated with different age groups.

The integration of the Random Forest Regressor with age-specific processing, incorporating consecutive differences, ensures a refined and nuanced prediction of Glomerular Filtration Rate (GFR) for dialysis patients. This sophisticated approach enhances the model's adaptability to age-related health dynamics.

Why Regression Over Time-Series Model:

Regression models are preferred over time-series models in this context for several reasons:

1. Flexibility:
    - Regression models are more flexible and can accommodate a variety of input features, making them suitable for datasets with diverse variables like patient demographics, lab results, and vital signs.
2. Interpretability:
    - Regression models provide clear interpretability of the impact of each feature on the predicted outcome. This transparency is crucial in a healthcare setting were understanding the factors influencing predictions is vital.
3. Handling Multivariate Inputs:
    - The dataset includes various numerical and categorical features beyond the time dimension, and regression models are well-suited for handling such multivariate inputs.
4. Non-Linear Relationships:
    - Regression models, especially ensemble models like Random Forest, can capture complex, non-linear relationships in the data, offering a more nuanced understanding of the factors affecting GFR.
5. Generalization:
    - Regression models are often more capable of generalizing well to new data, which is valuable for predicting GFR in diverse patient populations.

While time-series models are specialized for temporal data, the nature of this dataset, with a mix of temporal and non-temporal features, makes regression more appropriate for comprehensive prediction. The additional step of incorporating consecutive differences enriches the model's ability to adapt to temporal changes, combining the strengths of both approaches.

Validation: The process of validating is automated using an appropriate set of codes in order to check whether the model chosen as the best one (Random Forest Regressor) is giving accurate results than the others in discussion. Already existing patient are taken as the input for this automation process as we have the actual GFR trajectory in our raw data. The patient's first visit is taken and all the features, the values attributed and their first visited date are taken as their inputs for the algorithm. The devised model predicts the GFR for the next consecutive three years. Now we have the predicted GFR and the actual GFR for the patients. That's when we arrived with the results in Table 2 in the above sections. To even delve into the errors analysis, we check the number of patients whose predictions are more accurate and which models produces more patients with more precision. There are totally 225 patients with at least one record for three consecutive years (there are 225 patients who has been visiting the hospital for at least three years) those records are taken and subjected to analysis. It was based on how many predictions done by the model were closer to their actual GFR. If a patient has all three predictions nearer to their actual, that says the model has done a great job in learning the patterns with respect to that patient.

Table 4. Number of Patients with respect to number of records with deviation more than threshold (threshold = 15)

Algorithm used	Three records with deviation>10	Two records with deviation>10	One record with deviation>10	No record with deviation>10
Random Forest	9	21	52	143
Decision Tree	14	26	54	131
Support Vector	27	33	65	100
Light GBM	16	29	67	113
Linear Regression	21	35	67	102
KNN regressor	21	28	70	106
Neural Network Regressor	21	46	69	89
Random Forest with augmentation	12	19	44	150
Decision Tree with augmentation	16	22	56	131
Support vector with augmentation	26	34	62	103
Light GBM with augmentation	20	22	65	118
Linear regression with augmentation	20	22	65	118
KNN regressor with augmentation	16	34	53	122
Neural Network Regressor with augmentation	16	34	53	122

Table 4. Number of Patients with respect to number of records with deviation more than threshold (threshold = 15)

Algorithm used	Three record with deviation>15	Two record with deviation>15	One record with deviation>15	No record with deviation>15
Random Forest	4	6	26	190
Decision Tree	7	13	34	171
Support Vector	13	14	48	150
Light GBM	7	14	32	172
Linear Regression	12	15	36	162
KNN regressor	12	12	30	171
Neural Network Regressor	12	24	44	145
Random Forest with augmentation	6	6	26	187
Decision Tree with augmentation	8	14	37	166
Support vector with augmentation	13	15	47	150
Light GBM with augmentation	7	14	28	176
Linear regression with augmentation	7	14	28	176

KNN regressor with augmentation	10	14	29	172
Neural Network Regressor with augmentation	10	14	29	172

This table shows that there are 190 patients whose predictions by random forest regressor sounds good and under the threshold suggested and there are only 4 patient whose predictions does not sound credible. So, from the above table, it justifies our decision of taking Random Forest Regressor without Augmentation as the best working model for our project.

To identify why that happened their GFR and Serum Creatinine rate are taken into consideration and analysis has been made and the results are as follows.

Table 5. Mean deviation of GFR and SCR for patients with respect to their deviation in prediction (from Random Forest)

Attribute	Three records with deviation>15	Two records with Deviation>15	One record with deviation > 15	No records with deviation>15
GFR	4.00198	2.795604	0.77446	0.024611
SCR	0.103026	0.17278	0.00169	0.019155

Each cell in the above table holds the value of Mean deviation of attributes GFR and SCR. As it is clearly visible deviation of actual GFR is causing prediction in GFR. More the deviation in actual GFR, less accurate the predicted GFR. The anomaly in the actual data points resists with the model's ability to predict future GFR value correctly. It's clearly seen that actual data deviation is lesser for those patients who predicter GFR is more accurate, justifying our theory.

Conclusion:

In conclusion, the Random Forest Regressor, identified through rigorous evaluation, is complemented by the incorporation of age-specific differences in the model inputs. This advanced processing technique enhances the accuracy and relevance of GFR predictions by capturing age-related variations in health parameters. Continuous monitoring and potential refinements, considering age-specific trends, contribute to the ongoing efficacy of the predictive model for the next three years.

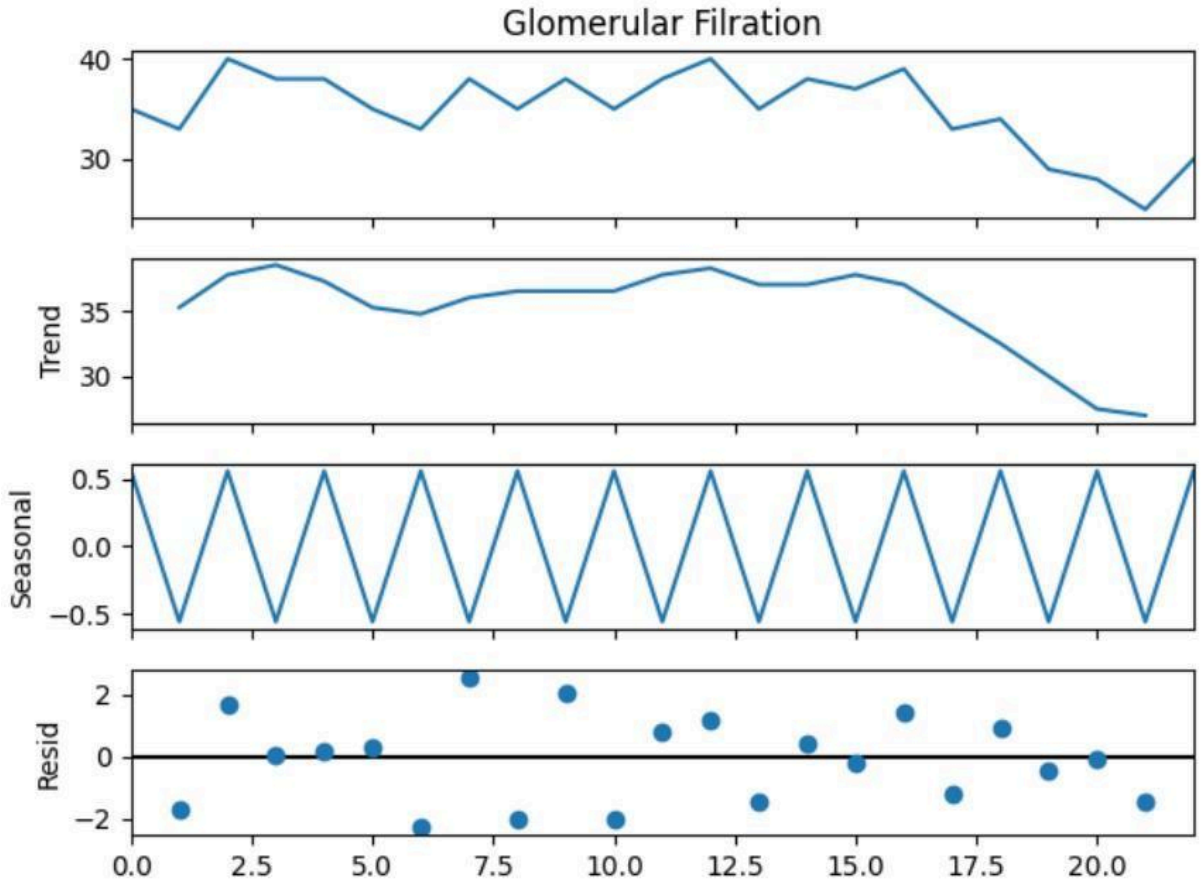
FORECASTER: GFR PREDICTION FOR THE NEXT THREE CONSECUTIVE YEARS FOR EXISTING PATIENTS:

The FORECASTER module serves as the cornerstone of this thesis, focusing on predicting Glomerular Filtration Rate (GFR) values for chronic kidney disease (CKD) patients over a span of three years. Through the integration of advanced machine learning techniques, this module holds paramount importance in enabling personalized treatment planning tailored to individual patient needs. By accurately forecasting GFR values, healthcare providers can make informed decisions, optimize resource allocation, and enhance overall management of CKD. The module's ability to predict GFR values empowers clinicians to proactively address disease progression, thereby improving patient outcomes and quality of care. Through its pivotal role, this module contributes significantly to the advancement of personalized medicine in the field of nephrology, ultimately striving towards better management and treatment of CKD on a patient-specific level.

Explanation of Time Series Model Selection

Before proceeding with time series modelling, several approaches were considered to analyse the dataset. However, the decision to use time series modelling was based on specific characteristics observed in the data, as well as the results obtained from preliminary analysis techniques. Here's an explanation of the reasons for choosing time series modelling:

Identification of Seasonal Patterns: The dataset under consideration exhibited clear seasonal patterns, indicating periodic fluctuations in the Glomerular Filtration Rate (GFR) values of chronic kidney disease (CKD) patients. By employing the seasonal decomposition technique using seasonal_decompose, the data was decomposed into trend, seasonal, and residual components. This analysis revealed distinct seasonal variations, highlighting the presence of recurring patterns over time.



Assessment of Stationarity:

To further assess the suitability of time series modelling, the Augmented Dickey-Fuller (ADF) test was conducted. The ADF test is a statistical method used to determine whether a given time series dataset is stationary or non-stationary. In a stationary time series, the statistical properties such as mean and variance remain constant over time, whereas in a non-stationary time series, these properties exhibit trends or fluctuations.

Interpretation of ADF Test Results:

The ADF test provides a p-value, which is used to assess the stationarity of the time series data. A small p-value (typically less than a predefined significance level, such as 0.05) suggests that the null hypothesis of non-stationarity can be rejected, indicating that the data is stationary. Conversely, a larger p-value suggests that the null hypothesis cannot be rejected, indicating non-stationarity.

Decision for Time Series Analysis:

In the case where the p-value falls within the range associated with non-stationary data, it suggests the presence of trends or fluctuations in the dataset. Given that the dataset exhibited clear seasonal patterns and non-stationary behaviour based on the ADF test results, time series modelling was deemed appropriate. Time series models, such as SARIMA, are

well-suited for capturing and forecasting data with seasonal trends and fluctuations, making them an ideal choice for analysing the CKD patient data in this scenario. In summary, the decision to use time series modelling was justified by the presence of seasonal patterns observed in the dataset and confirmed by the results of the ADF test, indicating non-stationary behaviour. This approach enables effective analysis and forecasting of GFR values for CKD patients over time, facilitating personalized treatment planning and resource allocation.

a. ARIMA Model:

The ARIMA (AutoRegressive Integrated Moving Average) model is a widely used time series forecasting method known for its effectiveness in capturing temporal trends and seasonality. In this project, the ARIMA model was employed to forecast Glomerular Filtration Rate (GFR) values for chronic kidney disease (CKD) patients over the next three years.

Methodology: The ARIMA model is composed of three main components: AutoRegressive (AR), Integrated (I), and Moving Average (MA). The AR component models the linear relationship between an observation and a certain number of lagged observations. The I component involves differencing the time series data to make it stationary. The MA component models the relationship between an observation and a residual error from a moving average model applied to lagged observations.

Usage in the Project: The ARIMA model was trained using both actual and augmented datasets comprising clinical, demographic, and laboratory variables of CKD patients. Transfer learning techniques were explored to enhance the model's predictive accuracy by leveraging pre-trained models. The model was evaluated based on metrics such as Mean Squared Error (MSE) and accuracy, showcasing promising results in capturing temporal trends and seasonality in GFR values.

Key Points:

ARIMA is effective for time series forecasting tasks due to its ability to capture both short-term and long-term dependencies in the data. The model's performance heavily depends on the stationarity of the time series data and the appropriate selection of model parameters. Transfer learning techniques can be beneficial in improving the model's accuracy by leveraging knowledge from pre-trained models.

b. SARIMA Model:

The SARIMA (Seasonal AutoRegressive Integrated Moving Average) model is an extension of the ARIMA model that incorporates seasonality into the time series forecasting process. In this project, the SARIMA model was utilized to forecast GFR values for CKD patients, considering seasonal fluctuations over time.

Methodology: The SARIMA model extends the ARIMA model by introducing additional parameters to capture seasonal patterns in the time series data. It comprises the same components as ARIMA (AR, I, MA) along with seasonal parameters that account for periodic

fluctuations. This model is particularly effective for capturing and forecasting data with clear seasonal patterns.

Usage in the Project: The SARIMA model was trained using both actual and augmented datasets, similar to the ARIMA model. Transfer learning techniques were also explored to enhance its predictive accuracy. The model's performance was evaluated based on metrics such as MSE and accuracy, demonstrating improved accuracy compared to the ARIMA model, especially in capturing long-term trends and seasonal fluctuations in GFR values.

Key Points:

SARIMA is specifically designed to handle time series data with clear seasonal patterns, making it suitable for forecasting tasks where seasonality plays a significant role. Proper identification and modelling of seasonal parameters are crucial for the accuracy of the SARIMA model. Transfer learning techniques can be effectively applied to SARIMA to leverage pre-trained models and improve predictive performance.

c. LSTM Model:

The LSTM (Long Short-Term Memory) model is a type of recurrent neural network (RNN) known for its ability to capture long-term dependencies in sequential data. In this project, the LSTM model was explored for its potential in forecasting GFR values for CKD patients over time.

Methodology: LSTM is designed to overcome the limitations of traditional RNNs in capturing long-term dependencies by introducing memory cells and gates that regulate the flow of information. This architecture allows LSTM to effectively capture temporal patterns and dependencies in sequential data.

Usage in the Project: The LSTM model was trained using both actual and augmented datasets, similar to the ARIMA and SARIMA models. However, transfer learning techniques were not applied to LSTM due to its unsuitability for this specific task. The model's performance was evaluated based on metrics such as MSE and accuracy, revealing suboptimal results compared to the ARIMA and SARIMA models.

Key Points:

LSTM is particularly effective for sequence prediction tasks where capturing long-term dependencies is essential. Despite its capabilities in capturing temporal patterns, LSTM may not be the most suitable model for all time series forecasting tasks, as demonstrated in this project. Transfer learning techniques may not always be applicable or beneficial for LSTM models, depending on the nature of the dataset and prediction task.

ARIMA Model with Actual Dataset:

Mean Absolute Error (MAE): 4.756482597904378

Mean Squared Error (MSE): 116.98127855639113

Root Mean Squared Error (RMSE): 10.815788392733612

Total Number of rows: 651

Count of rows with GFR Deviation greater than 15: 36

Count of rows with GFR Deviation less than 15: 615

Percentage of GFR Deviation greater than 15: 5.52%

Unique Patient ID's with GFR Deviation:

[1. 4. 14. 53. 68. 79. 82. 92. 95. 106. 120. 181. 193. 200. 217. 236. 243. 244. 250. 260. 273. 280. 283. 291. 296. 299. 301. 303. 304. 305.] (count = 30)

ARIMA Model with Augmented Dataset:

Mean Absolute Error (MAE): 3.8686089552607092

Mean Squared Error (MSE): 92.7311190733481

Root Mean Squared Error (RMSE): 9.629699843367295

Total Number of rows: 676

Count of rows with GFR Deviation greater than 15: 23

Count of rows with GFR Deviation less than 15: 653

Percentage of GFR Deviation greater than 15: 3.40%

Unique Patient ID's with GFR Deviation:

[1. 4. 14. 53. 68. 95. 106. 120. 181. 193. 200. 217. 250. 260. 280. 283. 296. 303. 304.]

(count = 19)

ARIMA Model with Transfer Learning and Actual Dataset:

Mean Absolute Error (MAE): 20.51940725630521

Mean Squared Error (MSE): 551.440597358899

Root Mean Squared Error (RMSE): 23.482772352490645

Total Number of rows: 199

Count of rows with GFR Deviation greater than 15: 129

Count of rows with GFR Deviation less than 15: 70

Percentage of GFR Deviation greater than 15: 64.82%

Unique Patient ID's with GFR Deviation:

[1. 10. 11. 12. 13. 19. 23. 29. 30. 32. 41. 46. 50. 52. 55. 57. 63. 64. 69. 71. 79. 82. 84. 92. 96. 97. 104. 112. 119. 120. 123. 126. 127. 128. 131. 134. 140. 144. 145. 147. 150. 158. 161. 162. 173. 182. 194. 197. 198. 200. 201. 206. 209. 210. 213. 214. 224. 225. 232. 235. 236. 243. 244. 246. 248. 253. 263. 264. 271. 274. 280. 285. 286. 287. 291. 297. 300. 302. 303. 304. 305.] (count = 81)

ARIMA Model with Transfer Learning and Augmented Dataset:

Mean Absolute Error (MAE): 20.76718560746568

Mean Squared Error (MSE): 556.5599947267012

Root Mean Squared Error (RMSE): 23.591523789842427

Total Number of rows: 224

Count of rows with GFR Deviation greater than 15: 152

Count of rows with GFR Deviation less than 15: 72

Percentage of GFR Deviation greater than 15: 67.85%

Unique Patient ID's with GFR Deviation:

[1. 10. 11. 12. 13. 19. 23. 29. 30. 32. 41. 46. 50. 52. 55. 57. 63. 64. 69. 71. 78. 79. 80. 82. 84. 92. 96. 97. 104. 112. 118. 119. 120. 123. 126. 127. 128. 131. 134. 140. 144. 145. 147. 150. 153. 158. 159. 161. 162. 165. 173. 177. 182. 194. 195. 197. 198. 199. 200. 201. 205. 206. 209. 210. 213. 214. 219. 224. 225. 232. 234. 235. 236. 241. 243. 244. 245. 246. 248. 253. 261. 263. 264. 271. 274. 279. 280. 284. 285. 286. 287. 291. 297. 299. 300. 302. 303. 304. 305. 308.] (count = 100)

SARIMA Model with Actual Dataset:

Mean Absolute Error (MAE): 5.219886676253568

Mean Squared Error (MSE): 408.31928003315073

Root Mean Squared Error (RMSE): 20.206911689645963

Total Number of rows: 961

Count of rows with GFR Deviation greater than 15: 33

Count of rows with GFR Deviation less than 15: 928

Percentage of GFR Deviation greater than 15: 3.43%

Unique Patient ID's with GFR Deviation:

[4 14 21 27 55 61 70 79 82 93 107 120 166 182 183 192 202 203 220 230 238 280 291 299 301 302 303 306] (count = 28)

SARIMA Model with Augmented Dataset:

Mean Absolute Error (MAE): 4.374410896571458

Mean Squared Error (MSE): 298.8424361849611

Root Mean Squared Error (RMSE): 17.287059790055714

Total Number of rows: 1228

Count of rows with GFR Deviation greater than 15: 27

Count of rows with GFR Deviation less than 15: 1201

Percentage of GFR Deviation greater than 15: 2.19%

Unique Patient ID's with GFR Deviation:

[1. 4. 14. 21. 27. 61. 82. 93. 107. 120. 166. 183. 202. 220. 230. 238. 242. 303. 306.]

(count = 19)

SARIMA Model with Transfer Learning and Actual Dataset:

Mean Absolute Error (MAE): 21.300495816470114

Mean Squared Error (MSE): 588.3339491936352

Root Mean Squared Error (RMSE): 24.255596244859355

Total Number of rows: 199

Count of rows with GFR Deviation greater than 15: 135

Count of rows with GFR Deviation less than 15: 64

Percentage of GFR Deviation greater than 15: 67.83%

Unique Patient ID's with GFR Deviation:

[1. 10. 11. 12. 13. 19. 23. 29. 30. 32. 41. 45. 46. 50. 52. 55. 57. 63. 64. 69. 70. 71. 79. 82. 84. 89. 92. 96. 97. 104. 112. 119. 120. 123. 126. 127. 128. 131. 133. 134. 140. 144.

145. 147. 150. 158. 161. 162. 173. 182. 194. 197. 198. 200. 201. 206. 209. 210. 212. 213. 214. 224. 225. 232. 235. 236. 243. 244. 246. 248. 253. 263. 264. 271. 274. 280. 285. 286. 287. 291. 297. 299. 300. 302. 303. 304. 305.] (count = 87)

SARIMA Model with Transfer Learning and Augmented Dataset:

Mean Absolute Error (MAE): 21.566982231959237

Mean Squared Error (MSE): 593.6191883226851

Root Mean Squared Error (RMSE): 24.364301515181698

Total Number of rows: 224

Count of rows with GFR Deviation greater than 15: 157

Count of rows with GFR Deviation less than 15: 67

Percentage of GFR Deviation greater than 15: 70.08%

Unique Patient ID's with GFR Deviation:

[1. 10. 11. 12. 13. 19. 23. 29. 30. 32. 41. 45. 46. 50. 52. 55. 57. 63. 64. 69. 70. 71. 78. 79. 80. 82. 84. 89. 92. 96. 97. 104. 112. 118. 119. 120. 123. 126. 127. 128. 131. 133. 134. 140. 144. 145. 147. 150. 153. 158. 159. 161. 162. 165. 173. 177. 182. 194. 195. 197. 198. 199. 200. 201. 205. 206. 209. 210. 212. 213. 214. 219. 224. 225. 232. 234. 235. 236. 241. 243. 244. 245. 246. 248. 253. 261. 263. 264. 271. 274. 279. 280. 284. 285. 286. 287. 291. 297. 299. 300. 302. 303. 304. 305. 308.] (count = 105)

LSTM Model with Actual Dataset:

Mean Absolute Error (MAE): 3.3646318514205653

Mean Squared Error (MSE): 36.879715723120626

Root Mean Squared Error (RMSE): 6.072867174829417

Total Number of rows: 631

Count of rows with GFR Deviation greater than 15: 17

Count of rows with GFR Deviation less than 15: 614

Percentage of GFR Deviation greater than 15: 2.69%

Unique Patient ID's with GFR Deviation:

[1. 4. 14. 56. 68. 92. 106. 120. 181. 200. 203. 230. 236. 260. 303. 304. 305.] (count = 17)

LSTM Model with Augmented Dataset:

Mean Absolute Error (MAE): 2.8669777985982154

Mean Squared Error (MSE): 29.402866173568587

Root Mean Squared Error (RMSE): 5.422440979260963

Total Number of rows: 676

Count of rows with GFR Deviation greater than 15: 15

Count of rows with GFR Deviation less than 15: 661

Percentage of GFR Deviation greater than 15: 2.22%

Unique Patient ID's with GFR Deviation:

[1. 4. 14. 56. 68. 106. 120. 181. 193. 200. 236. 260. 280. 303.] (count = 14)

Table 3. Test results for Module Forecaster

SNO	ALGORITHM	MEAN ABSOLUTE ERROR	ROOT MEAN SQUARED ERROR
1	ARIMA Model with Actual Dataset	4.756	10.81
2	ARIMA Model with Augmented Dataset	3.868	9.62
3	ARIMA Model with Transfer Learning and Actual Dataset	20.519	23.48
4	ARIMA Model with Transfer Learning and Augmented Dataset	20.767	23.59
5	SARIMA Model with Actual Dataset	5.219	20.2
6	**SARIMA Model with Augmented Dataset**	3.342	7.26
7	SARIMA Model with Transfer Learning and Actual Dataset	21.3	24.25
8	SARIMA Model with Transfer Learning and Augmented Dataset	21.566	24.36
9	LSTM Model with Actual Dataset	8.364	6.07
10	LSTM Model with Augmented Dataset	12.866	5.42

Out of all the ten different approaches made with the help of above three models, it was found that **S-ARIMA with the Augmented dataset** acts as the most accurate model with least error metrics, and it is implemented in the final project.

Validation:

The validation process was automated and results were noted down:

Patient ID	Actual GFR	Predicted GFR

115	34.0, 34.0, 36.0	38.714135204713614, 35.00668718915387, 38.12783991634884
44	30.666666666666668, 30.0, 27.2	28.286229944435508, 28.256505049175637, 26.195091871543916
172	13.0, 12.5, 11.5	12.204950727109383, 11.97136730068275, 11.433985521192536
2	19.0, 17.333333333333332, 14.5	15.6538644716915, 14.456020229321968, 15.393459016307792
257	56.5, 55.0, 59.5	60.62180862003319, 57.42460704189027, 57.547848740363065
36	32.0, 33.0, 27.25	32.53652985601055, 28.79358047885633, 32.626834322166516
81	33.0, 34.0, 32.5	32.04841179945686, 33.680554914637256, 32.47398786100562
4	72.0, 36.0, 41.333333333333336	69.999729034517784, 40.73479066048043, 43.04663174606461
72	23.5, 20.0, 23.333333333333332	20.03475479854564, 19.39706017870372, 19.546598690697248
117	19.25, 20.0, 23.0	22.70760245200741, 24.123241748199774, 20.51959074129783
157	16.333333333333332, 14.5, 14.5	14.689561682102715, 15.061585067420127, 13.028622388726834
269	36.0, 36.0, 37.0	34.559602661641144, 38.76641639414353, 37.79083452215418
276	21.0, 22.0, 23.0	19.61601864035177, 21.899274014666524, 22.500093255814722
74	33.666666666666664, 29.333333333333332, 31.333333333333332	30.851698001865085, 30.683511092277612, 29.336528728885966

186	55.5, 54.0, 60.0	56.339370626105264, 56.98655077912755, 57.82294347918718

SUGGESTER: FOR NEWLY INCOMING PATIENTS:

What if analysis:

What-if analysis, also known as sensitivity analysis or scenario analysis, is a technique used in machine learning to understand how changes in input variables or parameters affect the output or predictions of a model. It involves systematically varying the inputs to observe how the model's outputs change in response. What-if analysis helps in understanding the robustness of the model and identifying key factors that influence its predictions. It is commonly used for decision-making, risk assessment, and optimization in various domains, including finance, healthcare, and engineering.

AIM and Purpose: This module stands as a cornerstone within the broader project dedicated to constructing a comprehensive predictive model for the progression of chronic kidney disease (CKD). Operating as an intricate recommendation system, its primary aim is to empower healthcare providers with personalized insights gleaned from individual patient data, thereby optimizing patient management strategies.

The overarching goal of this module is to meticulously analyse the actual test values of CKD patients and juxtapose them against historical data of Low-Risk patients. By identifying deviations between actual and ideal values, the module aims to provide tailored interventions aimed at reducing patient risk profiles and enhancing overall CKD management efficacy.

Data Input and Historical Data Analysis:

This module seamlessly integrates a plethora of patient data, encompassing clinical, demographic, and laboratory variables such as serum creatinine levels, estimated glomerular filtration rate (eGFR), blood pressure, and urine albumin levels. Leveraging this comprehensive dataset, the module embarks on a meticulous analysis to identify Low Risk patients, who serve as benchmarks for ideal parameter values.

Drawing insights from historical data, the module identifies patients demonstrating favourable progression, characterized by an increase in GFR of at least 15 units. These Low-Risk patients serve as invaluable reference points for delineating ideal parameter values across various metrics, enabling a nuanced understanding of optimal patient management strategies.

Comparison and Discrepancy Identification:

At the core of its functionality, the module conducts a rigorous comparison between the actual test values of current patients and the ideal benchmarks extrapolated from historical data. This meticulous examination unveils discrepancies, highlighting areas where patient

management strategies can be refined to align with ideal benchmarks, thus optimizing CKD management.

The proposed methodology involves the development of an algorithm for managing low-risk patients by creating a dynamic data frame that evolves over time. The process begins with segregating previously identified low-risk patients into a distinct dynamic data frame, which serves as a foundation for ongoing adjustments. When new patient data is introduced, the algorithm initiates its execution by generating hypothetical data. This generation involves manipulating changeable features like Haemoglobin levels and water intake, while preserving unchangeable one-time data such as diabetes and hypertension.

Crucially, the algorithm operates recursively, continually generating hypothetical data until a point is reached where the glomerular filtration rate (GFR) may increase. This is accomplished by concurrently computing the Hypothetical GFR using a Random Forest Regressor. The algorithm systematically varies the input features in a hypothetical dataset, assessing the impact on GFR. If a set of hypothetical data yields a higher GFR than the actual GFR of the patient, this dataset is deemed valid and is suggested to the doctor as a potential course of action.

This approach incorporates a dynamic and iterative process, leveraging machine learning techniques to iteratively adjust hypothetical data until an improvement in GFR is achieved. By using Random Forest Regressor to estimate GFR, the algorithm provides a data-driven and personalized recommendation system. The utilization of recursion and the integration of machine learning models make this methodology robust, allowing it to adapt to changes in patient data over time and offering valuable insights to healthcare professionals for patient care decisions.

The point to be noted is that the hypothetical data while generation is taken at most care and foreseen and supervised so as to make it more legible and legit. That is, the value generated are more scrutinised to the healthy ranges that are suggested by the doctors in front hand. The data that is fed to the model for computing the hypothetical GFR lies in a healthy human range which is previously suggested. It’s done because the low-risk patient’s aggregated values of the important features may not fall under healthy ranges as these people are still patients and we cannot recommend the high risk the exact values without surveillance as it may lead harm to them.

Recommendation Generation:

Armed with insights derived from the comparative analysis, the module generates personalized recommendations tailored to bridge the gap between actual and ideal values. These recommendations encompass a spectrum of interventions, including medication adjustments, lifestyle modifications, dietary changes, and suggestions for additional diagnostic tests, aimed at optimizing patient management and improving outcomes.

Implementation: Utilizing advanced artificial intelligence and machine learning algorithms, this module operates within an intuitive interface, facilitating seamless data input, recommendation viewing, and patient progress tracking for healthcare providers. By leveraging actual data from high-risk patients and synthesized data from low-risk individuals, the module offers actionable insights into potential enhancements to patient management strategies, thereby bolstering CKD management efficacy.

RECOMMENDER: FOR EXISTING PATIENTS:

Module Overview:

Embedded within the healthcare ecosystem, this module stands as a pivotal tool tailored for existing patients within hospital records, extending the functionality of its counterpart, the Suggester module, except for the fact that suggester runs for the new patient data and recommender runs with the latest data/visit's data that is in the database, so the latest data points of the existing patient is fed to the model. Beyond standard recommendations, it meticulously analyses patient data to offer personalized insights, with a distinctive focus on medicine recommendations. Medicine recommendation makes this module stand out from the rest of all, enabling comprehensive care strategies for enhanced patient outcomes.

Idea:

Building upon the foundation of leveraging Low risk patients as reference points, this module\pioneers the integration of medicine recommendations into patient care. By scrutinizing the medication regimens of Low-risk patients across multiple visits, it identifies pivotal medications (scrutinised by taking counting which medicines the patient takes in a lot when GFR hikes) associated with notable improvements in Glomerular Filtration Rate (GFR). This database of successful medication strategies serves as a benchmark for evaluating and refining the treatment regimens of existing patients.

Based on two different ideas of Risk predictions, medicines have been classified in distinct ways and are ranked based on the insights:

- Based on just GFR threshold:
  1. Medicines only taken by Low-risk patients (most effective medicines):

TAB.Clonazepam - 0.5mg HS
TAB.Ubicar - OD
TAB.Nicardia R 30mg TDS
TAB.Vogli MD - 0.2 mg BD
TAB.Dapaglyn 5mg OD
TAB.Carca 6.25mg BD
TAB.Arbitel 40mg OD
TAB.Zeblong 16mg OD
TAB.Carca CR - 10 mg BD
TAB.Concor AM - BD
TAB Renodapt 360mg TDS
TAB Cipcal 500mg OD

2. Medicines that weren't even taken by Low-risk patients even once:

Medicine ID	Medicine Name
1	TAB.Repace - 50mg BD
2	TAB.Ketocheck DS
3	TAB.Minipress XL - 2.5 mg HS
4	TAB.Telma AZ - OD
5	TAB.Ketosteril - 1 TDS
6	TAB.Mirbeg 50mg
7	TAB.Urimax - 0.4 mg HS
8	TAB.Zava - 50 mg OD
9	TAB.Dapavel - 5 mg OD
10	TAB.Trika 0.5mg HS
11	TAB.Amlong 5mg BD
13	TAB.Angispan TR 2.6 BD
14	TAB.Dapavel 10mg OD
15	TAB Atorvas 10mg HS
16	TAB.Ketoadd - 1
17	TAB.CTD - 12.5 mg OD
18	TAB.Ketocheck TDS
19	TAB.Thyronorm - 50 mcs OD
20	TAB Amlodac 5mg OD
21	TAB.Lobun
22	TAB.Ketosteril 2 TID
23	TAB.Shelcal - 500mg BD
24	CAP. Uprise D3 1000 units
25	TAB.Carca CR - 10 mg OD
26	TAB.Cobadex Forte OD
27	TAB Clopilet A (75/75) OD
28	TAB.Bengreat - 4 mg BD
29	TAB.Atorva 20mg HS
30	TAB.Flavedon MR 35mg BD
31	TAB.Nitrocontin 2.6mg BD
32	TAB.Sevecord 400 mg HS
33	CAP.Pregalin - 75 HS
34	TAB.Nebicard 5mg BD
35	TAB.Ketoadd 1 TDS
36	TAB.Rocaltrol - 0.25 mg OD
37	TAB.Alpha D3 OD
38	TAB Sporidex 500mg BD
39	TAB Amlong 5mg BD
40	SYP.Apitrate - 10ml BD
41	TAB.Zentel - 400mg
42	TAB.Sporlac DS
43	TAB.Megahenz 40mg OD
44	TAB.INH 300mg OD

45	TAB Rif 300 mag OD
46	TAB Pyz 750mg OD
47	TAB Tarivid 150 mg OD
48	TAB.Ecospirin - 150 mg OD
49	TAB Rapilif 8mg HS
50	TAB.Thyroxine - 50 mg OD
51	TAB.Zincovit
52	SYP.Aptivate TDS
53	TAB Telma Am (5/40) OD
54	TAB.Met xl 12.5 OD
55	TAB.Awaytox BD
56	TAB.Glynase - 5 mg BD
57	TAB.Cardace - 5mg OD
58	TAB Gluvilda 50mg OD
59	SYP Lactihep 10ml HS
60	SYP.KCL - 10ml BD
61	TAB.Ziten 20mg OD
62	CAP.Ecospirin AV 70/10
63	TAB.Stiloz - 50 mg
64	TAB.Zolsoma - 5 mg HS
65	TAB GabaPentin
66	TAB.Prazopress XL - 2.5 mg BD
67	Ketograce Sachet - OD
68	TAB.Forxiga - 5 mg OD
69	TAB Forxiga 10mg OD
70	TAB.Oxemia 100mg OD AD
71	TAB.Amtas 5mg-HS
72	TAB.Teneligliptin - 20 mg OD
73	TAB.Nexiron LP OD
74	TAB.Kerendia 10mg OD
75	TAB.Ketograce TDS
76	TAB.Neurobion Forte
77	TAB.Flomont 10 mg HS
78	TAB.Tamdura HS
79	TAB.Invokana 100mg OD
80	TAB.Amlong 5mg OD
81	TAB.Concor - 5 mg OD
82	TAB Dytor 5mg BD
83	TAB.Nicardia XL 30 TDS
84	TAB.Pregalin - 50mg HS
85	TAB.Thiamine 75 OD
86	TAB. Met XL - 25 mg OD
87	TAB Pregaba 50mg OD
88	TAB.Ferronemia OD
89	TAB.Oxemia 100mg OD
90	TAB.ELTROXIN - 50mg OD

91	TAB.Amtas AT - (5/50)mg OD
92	TAB.Atorday - 10mg HS
93	TAB Shelcal 250mg OD
94	TAB.Nebicard - 5mg OD
95	TAB.Neovit D3 - 0.25 mg AD OD
96	TAB.Met XL - 25mg BD
97	TAB.wysolone 10mg OD
98	TAB.Goutnil - 0.5mg OD
99	TAB.Pentids 400mg BD
100	TAB.Linagliptin 5mg
101	TAB.Dytor 20-10-10
102	TAB.Aminorich TDS
103	TAB.Citraphos
104	TAB.Dynalta
105	TAB.Glucobay - 50 mg BD
106	CAP.Rapilif - 4mg
107	TAB.Budenofil - 9 mg OD
108	TAB nebicard 2.5 HS
109	TAB.Wysolone 5mg
110	Wysolone 40mn OD
111	TAB.Oxemia - 50mg
112	TAB Metolar xl 12.5mg OD
113	TAB Obimet SR 500mg BD
114	TAB Amlong 10mg OD
115	Dolo 650 mg BD
116	TAB.Lipaglyn - 4 mg
117	TAB.Glucobay - 50 mg OD
118	TAB.Concor - 5mg BD
119	TAB.Obimet SR - 500mg OD
120	TAB.SA Rapid - 0.5mg OD
121	SYP.Kcit 5ml BD
122	SYP.Lactihep - 15ml BD
123	TAB.Unienzyme BD
124	TAB.Aztric - 40 mg OD
125	TAB.Nicardia R 20 mg BD
126	CAP.Clopitorva - 75/10 HS
127	TAB.Sompraz D OD
128	TAB.Flunarin 10mg BD
129	TAB.Efnocor 20 BD
130	TAB.Remylin DX OD
131	TAB.Glycinorm - 40mg OD
132	TAB.Gluvilda 25 mg OD
133	TAB.Natrillix SR 1.5 mg OD
134	TAB.Trigabantin
135	TAB.Diamicron XR - 30 mg
136	TAB.Flexijoint

137	TAB.Homo 16 LC
138	TAB.Clopilet Av (75/20)mg OD
139	TAB.Rejoint -1/2 BD
140	TAB.Remo - 100 mg BD
141	TAB.Atora - 40mg
142	TAB.Rocaltrol - 0.25 mg
143	TAB.Glycomet
144	TAB.JARDIANCE - 10mg
145	TAB.Feburic - 80 mg
146	TAB.Evion LC
147	TAB.Myfortic 360 mg BD
148	CAP.Magmax - 400mg BD
149	TAB.Indomethacin - 25mg BD
150	TAB.Betacap TR - 20mg OD
151	TAB.Indomethacin - 25mg OD
152	TAB.Febutaz - 20 mg
153	TAB.Revololxl -25 ml OD
154	TAB.Amlong -2.5mg
155	TAB.Restyl - 0.25mg HS
156	TAB.Acitrom
157	TAB.Tonact - 10mg HS
158	TAB Cardivas 3.125mg BD
159	TAB.Pruvict -2mg HS
160	TAB.Aten - 25 mg OD
161	TAB Predmet 8mg TDS
162	TAB Revlamer 400mg TDS
163	TAB Deplata 75mg OD
164	TAB Rosuvas 5mg HS
165	TAB.Nebiflo - 2.5 mg OD
166	TAB.Amaryl M - 1mg OD
167	TAB.Feburic - 20 mg OD AD
168	TAB.Astat - 10 mg HS
169	TAB Cilacar (10/40) OD
170	TAB.Metaprolol xl - 25mg OD
171	TAB.Tonact - 20 mg HS
172	TAB.Prax - 5 mg OD
173	TAB.Amlogard - 2.5 mg OD
174	TAB.Deplatt A 150 OD
175	TAB.Nitrosun 5mg HS
176	TAB.Aldactone 50mg BD
177	TAB.Calcigard R - 10 mg BD
178	TAB.Lobet 100mg BD
179	TAB.Bronac - 600mg OD
180	TAB.Isordil - 5 mg TDS
181	TAB.Eptoin 75mg BD
182	TAB.Roliten - 2mg HS

183	TAB.Calcitriol 0.25mg OD
184	TAB.Nicardia XL - 30 mg BD
185	TAB.Nodosis - 500 mg (2-2-1)
186	TAB.Fludrocort - 0.1 mg OD
187	TAB Alprax 0.25mg HS
188	TAB.Nicardia XL 60 mg BD
189	TAB.Cilacar - 20 mg BD
190	TAB.Prazopress XL - 5 mg BD
191	TAB.Ketograce BD
192	TAB.Oxemia 50 mg OD AD
193	TAB Orofer XT OD
194	K-Bind powder 15mg BD
195	TAB.Amaryl - 1mg BD
196	TAB.Aldactone 25mg OD
197	TAB.Benadon - 100 mg OD
198	TAB Deplat 75mg OD
199	TAB.Nexiron LP
200	TAB.Minipress XL - 5 mg BD
201	TAB Eltroxin 100mg OD
202	TAB.Amtas 10mg HS
203	TAB.Sandacol OD
204	TAB.Alphadol - 0.5 mg OD
205	TAB. Neurobion forte 1 OD
206	TAB.Palmova
207	TAB.Nitrotong 2.6 mg BD
208	TAB.SA Rapid - 1mg OD
209	TAB.Cordarone 100 mg BD
210	TAB.Urispas - 100mg BD
211	TAB.Fludrocort - 0.1 mg AD OD
212	TAB.Rejuneuron Forte - OD
213	TAB.Storvas - 40 mg HS
214	TAB Alprax 0.5mg
215	TAB.Isolazine 20/37.5 TDS
216	TAB.Plavix - 75 mg OD
217	TAB Linotril 0.5 mg
218	SYP.URILYSER - 15ml BD
219	TAB Met XL 50mg OD
220	TAB.Vymada 50 mg
221	TAB. Betaloc 25mg OD
222	TAB.Zavamet - (50/500) OD
223	TAB.Omnacortil 5mg OD
224	TAB.wysolone 20mg OD
225	TAB.Livogen
226	TAB.Cardivas - 6.25 mg OD
227	Febustat 40mg
228	TAB Feburic 40mg OD

229	TAB.Metpure XL
230	TAb.folygel - 5 mg OD
231	TAB Cilaheart T BD
232	TAB.MOXON 0.2 mg - OD
233	TAB.Zytanix - 2.5 mg AD
234	TAB.Zyloric - 100mg BD
235	TAB Epitril 0.5 mg OD
236	TAB.Novamox 250 TDS
237	SYP.KCL - 15ml TDS
238	TAB.Isordil 10 mg TLD
239	TAB.Bactrim ds HS
240	TAB.Cobadex czs 100
241	CAP.Lobun Forte OD
242	TAB.Enteclavir - 0.5 mg

- Based on blending both GFR threshold and Stage deviation:
  1. Medicines only taken by Low-risk patients (most effective ones):

Medicine Name
TAB.Mega 3 - 1 TDS
TAB.Carnitor - 500mg BD
TAB.Trika 05mg HS
TAB.Sodamint - 500mg TDS
TAB.Supriclox - 500mg TDS
TAB.Amlong -2.5mg
TAB.Restyl - 0.25mg HS
TAB.Dapaglyn 5mg OD
TAB.Atora - 40mg
TAB.Revololxl 25mg HS
TAB.MOXON 0.2 mg - OD
TAB.Remo - 100 mg BD
TAB.Flexijoint
TAB.Bifilac BD
TAB.Synprotik OD
TAB.Onglyza 2.5mg OD
TAB.Raciper-D - BD
TAB Semiglynase 2.5 mg OD
TAB Cranmed OD
TAB Levoflox 500mg OD
TAB Rosolip 5mg HS
TAB Sorbitrate 10mg TDS
TAB Amlogard 5mg OD
SYP Corex
TAB Dynapress 4mg HS
TAB.One Up TDS
TAB.Cresar AM OD
CAP.Gemcal OD

TAB.Renerve OD
TAB Sporidex 250mg
TAB.Tegretol - 100mg HS
TAB.Aten - 25 mg BD
CAP.Neovit D3 -OD
TAB.Renodapt s 360mg
TAB.Omnacortil 5mg OD
TAB.Concor AM - BD
TAB.Isordil 10 mg TLD
TAB Alprax 0.5mg
TAB.Aldactone 50mg BD
SYP.KCL - 15ml TDS
TAB.wysolone 20mg OD
SYP.Kcit 5ml BD
TAB Trental 400mg OD
CAP.Uprise D2 60000
SYP.Zincovit - 10 ml OD
TAB Cilaheart T BD
TAB.Tide - 10 mg BD
TAb.folygel - 5 mg OD
TAB.Tamdura HS
TAB Domstal BD
TAB.Nicardia R 30mg TDS
TAB.Predmet - 32 mg OD
TAB.Glynase - 2.5 mg BD
TAB.Zyloric - 100 mg OD
CAP.Vibact - OD
SYP.Sucrafil - 1tsp BD
TAB.Amlogard 5mg BD
TAB.Isordil - 5 mg TDS
TAB.Clobazam - 10 mg OD
TAB.Labetalol - 100 mg OD
TAB.Amtas AT - (5/50)mg OD
TAB.Volix - 0.2 mg BD
TAB.CTD - 6.25mg OD
TAB cilacar 5mg HS
TAB Aspirin 75mg OD
TAB.Flunarin 10mg HS
CAP.Megaza
TAB.Carca 6.25mg BD
TAB.CIpcal - 500 BD
TAB Nitrovet 10mg HS
TAB.Cardivas - 6.25 mg BD
SYP Mucaine 2 tsp BD
TAB.Raciper - 40 mg BD
TAB.Acitrom

CAP.Silofast - 4 mg HS
TAB.Sodamint - 500 mg BD
TAB.Gabapentin 100mg HS
TAB.Thyroxine - 50 mg OD
TAB. Betaloc 25mg OD
TAB.Brilinta - 90 mg BD
TAB.Carca CR - 10 mg BD
TAB.Nexiron LP AD
TAB Cardivas 3.125mg BD
TAB Tryptomer 10mg HS
TAB Pyz 750mg OD
TAB.Xtor 10mg HS
TAB.SA Rapid - 1mg OD
TAB Crocin 500mg
TAB.Zeblong 16mg OD

2. Medicines that were not at all taken by Low Risk patients:

Medicine ID	Medicine Name
1	TAB.Telma AZ - OD
2	TAB.Mirbeg 50mg
3	TAB.Angispan TR 2.6 BD
4	TAB.CTD - 12.5 mg OD
5	CAP.T3 D3 - 1 OD
6	TAB Amlodac 5mg OD
7	TAB Amlong 5mg BD
8	TAB.Zentel - 400mg
9	TAB.Met xl 12.5 OD
10	CAP.Ecospirin AV 70/10
11	TAB GabaPentin
12	TAB.Ketograce TDS
13	TAB Pregaba 50mg OD
14	TAB Shelcal 250mg OD
15	TAB.Pentids 400mg BD
16	TAB.Linagliptin 5mg
17	TAB.Dytor 20-10-10
18	TAB.Flomont 10 mg HS
19	TAB.Dynalta
20	TAB Renochlor (1) BD
21	TAB.Wysolone 5mg
22	TAB Metolar xl 12.5mg OD
23	TAB Obimet SR 500mg BD
24	TAB Amlong 10mg OD
25	TAB.Glucobay - 50 mg OD
26	TAB.Flunarin 10mg BD
27	TAB.Efnocor 20 BD

28	TAB.Remylin DX OD
29	TAB.Evion LC
30	TAB.Indomethacin - 25mg BD
31	TAB.Indomethacin - 25mg OD
32	TAB.Tonact - 10mg HS
33	TAB Deplata 75mg OD
34	TAB Rosuvas 5mg HS
35	TAB.Nebiflo - 2.5 mg OD
36	TAB.Amaryl M - 1mg OD
37	TAB.Glycinorm M - (40/500) mg BD
38	TAB.Tonact - 20 mg HS
39	TAB.Prax - 5 mg OD
40	TAB.Amlogard - 2.5 mg OD
41	TAB.Nitrosun 5mg HS
42	TAB.Eptoin 75mg BD
43	TAB.Calcitriol 0.25mg OD
44	TAB.Fludrocort - 0.1 mg OD
45	TAB.Prazopress XL - 5 mg BD
46	TAB.Ketograce BD
47	TAB.Oxemia 50 mg OD AD
48	TAB Orofer XT OD
49	K-Bind powder 15mg BD
50	TAB.Nexiron LP
51	TAB Eltroxin 100mg OD
52	TAB.Amtas 10mg HS
53	TAB.Sandacol OD
54	TAB. Neurobion forte 1 OD
55	TAB.Palmova
56	TAB.Cordarone 100 mg BD
57	TAB.Warf 5mg OD
58	TAB.Fludrocort - 0.1 mg AD OD
59	TAB.Rejuneuron Forte - OD
60	TAB.Storvas - 40 mg HS
61	TAB Feburic 40mg OD
62	TAB.Metpure XL
63	TAB.Zytanix - 2.5 mg AD

Note: The 4 above tables are static here but dynamic in the application, i.e., it is not fixed and there will definitely be changes and alterations occurring dynamically over time.

Methodology:

The methodology is rooted in a meticulous analysis of medication profiles and their correlation with GFR outcomes among Low-risk patients. Leveraging Cosine similarity computation, the module juxtaposes the medication profile of a high-risk patient against this dataset. This mathematical technique discerns similarities between medication profiles, facilitating the identification of potentially beneficial medications.

Cosine Similarity Computation: Treating each list (medicine list) as a vector in a Multi-Dimensional Space and calculating its cosine angle. Closer the angle to 1, more similar the lists are. First the string form of medicines are converted into integer or numerical lists with the help of a function called CountVectorizer. The vectors are compared using cosine similarity, and the function identifies the index of the most similar medication list. The list at this index is considered the most comparable to the medications the patient has taken. The function then further refines this list by excluding medications that the patient has already taken, providing a final recommendation of suggested medications that the patient might find beneficial based on their medical history. The cosine similarity measure, in this context, facilitates the identification of medication lists that align closely with the patient's historical usage, aiding in personalized medication recommendations.

By recommending the addition of specific medications to the patient’s regimen, the module aims to address treatment gaps and optimize care pathways, ultimately fostering GFR improvement and enhancing patient well-being.

Implementation:

Operational within an advanced artificial intelligence framework, this module seamlessly integrates with existing healthcare infrastructure. Through its intuitive interface, healthcare providers gain access to actionable insights derived from patient data and medication profiles. The incorporation of medication recommendations elevates the module’s utility, empowering providers to make informed decisions and tailor treatment plans to individual patient needs. By facilitating the integration of personalized medication recommendations into patient care, the module plays a pivotal role in optimizing CKD management strategies and driving improved patient outcomes.

Validation for Recommender and Suggester:

The validation process for the above two modules were done manually to see whether the suggestions and recommendations are actually logical and does it fall in the healthy ranges.

And whether the medicines are suggested right:

21	160	181
Current - Haemoglobin Level 13.9 Approximate - Haemoglobin Level 15.175421865334815 Current - Tota Water Intake noise removed 2.0 Approximate - Tota Water Intake noise removed 2.483295246528243 Current - Random Blood Sugar 82 Approximate - Random Blood Sugar 108 Current - Systolic 140 Approximate - Systolic 140 Current - Diastolic 80 Approximate - Diastolic 100 Current - Urine Albumin 3 Approximate - Urine Albumin 1 Living Area should be changed from the existing. Current - BMI 22.7 Approximate - BMI 21.795548909011686 Worker should be changed from the existing. 23 is the Actual GFR	Current - Haemoglobin Level 14.1 Approximate - Haemoglobin Level 13.56148469579957 Current - Tota Water Intake noise removed 2.5 Approximate - Tota Water Intake noise removed 1.8182318434962945 Current - Random Blood Sugar 170 Approximate - Random Blood Sugar 112 Current - Systolic 130 Approximate - Systolic 140 Current - Diastolic 80 Approximate - Diastolic 100 Current - Urine Albumin 3 Approximate - Urine Albumin 2 Living Area should be changed from the existing. Current - BMI 24.24 Approximate - BMI 19.51593698442218 Worker should be changed from the existing. 34 is the Actual GFR	Current - Haemoglobin Level 11.0 Approximate - Haemoglobin Level 16.458248587187434 Current - Tota Water Intake noise removed 2.5 Approximate - Tota Water Intake noise removed 2.4734312531608778 Current - Random Blood Sugar 279 Approximate - Random Blood Sugar 140 Current - Systolic 120 Approximate - Systolic 140 Current - Diastolic 80 Approximate - Diastolic 60 Current - Urine Albumin 3 Approximate - Urine Albumin 3 Living Area should remain the same. Current - BMI 25.88 Approximate - BMI 21.370657401296466 Worker should be changed from the existing. 10 is the Actual GFR 25.0 is the hypothetical GFR

38.0 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Urimax - 0.4 mg HS * TAB.Nodosis - 500mg BD * TAB.Isordil - 5 mg TDS * TAB Trika 0.25mg HS * TAB.Becosules OD * TAB.Ecospirin - 150 mg OD * TAB.Daxid - 25 mg HS * TAB.Betacap TR - 20mg OD * CAP Becosules OD * TAB.Nodosis - 500mg TDS	37.7 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Nephrosave OD * TAB.Telma - 40 mg * TAB.Storvas - 10mg HS * TAB.JARDIANCE - 25mg * TAB.Glucobay - 50mg TDS * TAB.Feburic - 40mg OD	Medicines that could help the GFR increase for this patient: * TAB.Urimax - 0.4 mg HS * TAB.Nodosis - 500mg BD * TAB.Atorvas - 10mg * TAB.Isordil - 5 mg TDS * TAB Trika 0.25mg HS * TAB.Becosules OD * TAB.Ecospirin - 150 mg OD * TAB.Daxid - 25 mg HS * TAB.Betacap TR - 20mg OD * CAP Becosules OD
20	**248**	**55**
Current - Haemoglobin Level 10.7 Approximate - Haemoglobin Level 12.360420237908482 Current - Tota Water Intake noise removed 2.5 Approximate - Tota Water Intake noise removed 2.659024608992267 Current - Random Blood Sugar 96 Approximate - Random Blood Sugar 108 Current - Systolic 110 Approximate - Systolic 130 Current - Diastolic 70 Approximate - Diastolic 90 Current - Urine Albumin 3 Approximate - Urine Albumin 2	Current - Haemoglobin Level 13.5 Approximate - Haemoglobin Level 14.149179008533343 Current - Tota Water Intake noise removed 3.0 Approximate - Tota Water Intake noise removed 2.7731802383131905 Current - Random Blood Sugar 106 Approximate - Random Blood Sugar 111 Current - Systolic 130 Approximate - Systolic 90 Current - Diastolic 70 Approximate - Diastolic 60 Current - Urine Albumin 3 Approximate - Urine Albumin 0	Current - Haemoglobin Level 14.0 Approximate - Haemoglobin Level 16.384925262394663 Current - Tota Water Intake noise removed 3.0 Approximate - Tota Water Intake noise removed 2.3675211540962877 Current - Random Blood Sugar 341 Approximate - Random Blood Sugar 130 Current - Systolic 110 Approximate - Systolic 120 Current - Diastolic 70 Approximate - Diastolic 60 Current - Urine Albumin 3 Approximate - Urine Albumin 2

Living Area should be changed from the existing. Current - BMI 24.5 Approximate - BMI 24.65182775497688 Worker should be changed from the existing. 23 is the Actual GFR 38.0 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Repace - 25mg * TAB.Nephrosave OD * TAB.Nodosis - 500mg BD * TAB.Dapaglyn 5mg OD \ufeff * TAB.Repace - 50mg OD	Living Area should be changed from the existing. Current - BMI 27.2 Approximate - BMI 24.41795255216006 Worker should be changed from the existing. 34 is the Actual GFR 43.2 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Cilacar - 10mg OD * TAB.bisferxt OD * TAB.Oxemia - 100mg * TAB.Glimy * TAB.Nodosis - 500mg BD * TAB.Storvas - 20mg HS * TAB.Rantac - 150mg BD * TAB.Dytor - 10mg OD * TAB.Alphadol - 0.25 mg * TAB.Pruvict - 1mg HS * TAB Montair * TAB.Pan 40 mg OD * TAB.Feburic - 40mg OD	Living Area should be changed from the existing. Current - BMI 19.14 Approximate - BMI 22.457935988258438 Worker should remain the same. 29 is the Actual GFR 41.2 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Feburic - 40mg OD * CAP.T3 D3 - 1 OD * TAB.Nodosis - 250mg BD * TAB.Shelcal - 500mg OD * TAB.Nephrosave OD * TAB.Rejoint -1 BD * TAB.Eltroxin - 75mg OD * TAB Aten 50mg OD
83	**264**	**141**
Current - Haemoglobin Level 9.1 Approximate - Haemoglobin Level 17.099050017374577 Current - Tota Water Intake noise removed 2.5 Approximate - Tota Water Intake noise removed 1.535073667096932	Current - Haemoglobin Level 9.4 Approximate - Haemoglobin Level 14.623140643070027 Current - Tota Water Intake noise removed 2.0 Approximate - Tota Water Intake noise removed 2.844726664897638	Current - Haemoglobin Level 12.4 Approximate - Haemoglobin Level 16.51329243810394 Current - Tota Water Intake noise removed 2.0 Approximate - Tota Water Intake noise removed 2.562452607301471

Current - Random Blood Sugar 126 Approximate - Random Blood Sugar 138 Current - Systolic 110 Approximate - Systolic 140 Current - Diastolic 70 Approximate - Diastolic 100 Current - Urine Albumin 3 Approximate - Urine Albumin 0 Living Area should remain the same. Current - BMI 22.74 Approximate - BMI 19.424019407902072 Worker should remain the same. 40 is the Actual GFR 44.5 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * CAP.T3 D3 - 1 OD * TAB.Nodosis - 250mg BD * TAB.Rejoint -1 BD * TAB.Eltroxin - 75mg OD * TAB Aten 50mg OD	Current - Random Blood Sugar 178 Approximate - Random Blood Sugar 190 Current - Systolic 140 Approximate - Systolic 140 Current - Diastolic 100 Approximate - Diastolic 70 Current - Urine Albumin 3 Approximate - Urine Albumin 3 Living Area should be changed from the existing. Current - BMI 24.61 Approximate - BMI 23.58842257524553 Worker should be changed from the existing. 17 is the Actual GFR 32.0 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB Montair * TAB Minipress XL 5mg OD * TAB.Bronac - 600MG BD * TAB.Atorvas - 10mg * TAB.Clopilet 75 OD * TAB.Aldactone 50mg BD	Current - Random Blood Sugar 208 Approximate - Random Blood Sugar 150 Current - Systolic 130 Approximate - Systolic 140 Current - Diastolic 70 Approximate - Diastolic 90 Current - Urine Albumin 3 Approximate - Urine Albumin 1 Living Area should remain the same. Current - BMI 25.71 Approximate - BMI 21.62899844607593 Worker should remain the same. 34 is the Actual GFR 43.4 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Telma - 40 mg OD * TAB.Nephrosave Forte - 1 OD * CAP.Renerve Plus OD * TAB.Storvas - 10mg HS * TAB Folvite 5mg OD * TAB.Clopilet 75 OD * TAB.Cardivas CR - 10mg OD * TAB cilacar 5mg HS * TAB. Telma - 20 mg * TAB.Cilacar - 10mg OD * TAB.Nodosis - 500mg TDS * TAB.Dapaglyn 5mg OD \ufeff * TAB.Ranadyl

		* TAB.Obimet SR - 500mg BD * TAB.Storvas - 20mg HS * TAB.Clavix AS - OD
100	**175**	**195**
Current - Haemoglobin Level 11.8 Approximate - Haemoglobin Level 13.634783970992565 Current - Tota Water Intake noise removed 2.5 Approximate - Tota Water Intake noise removed 1.505064997262992 Current - Random Blood Sugar 100 Approximate - Random Blood Sugar 95 Current - Systolic 110 Approximate - Systolic 110 Current - Diastolic 70 Approximate - Diastolic 90 Current - Urine Albumin 3 Approximate - Urine Albumin 2 Living Area should remain the same. Current - BMI 49.79 Approximate - BMI 24.77112119418634 Worker should be changed from the existing. 15 is the Actual GFR 30.0 is the hypothetical GFR	Current - Haemoglobin Level 11.3 Approximate - Haemoglobin Level 16.772037696208205 Current - Tota Water Intake noise removed 1.0 Approximate - Tota Water Intake noise removed 2.1224722434827026 Current - Random Blood Sugar 112 Approximate - Random Blood Sugar 88 Current - Systolic 110 Approximate - Systolic 140 Current - Diastolic 70 Approximate - Diastolic 100 Current - Urine Albumin 3 Approximate - Urine Albumin 2 Living Area should remain the same. Current - BMI 33.15 Approximate - BMI 19.419387004501736 Worker should remain the same. 7 is the Actual GFR 22.0 is the hypothetical GFR	Current - Haemoglobin Level 10.3 Approximate - Haemoglobin Level 12.888843006553772 Current - Tota Water Intake noise removed 2.5 Approximate - Tota Water Intake noise removed 1.6763724203971613 Current - Random Blood Sugar 90 Approximate - Random Blood Sugar 102 Current - Systolic 120 Approximate - Systolic 140 Current - Diastolic 70 Approximate - Diastolic 70 Current - Urine Albumin 3 Approximate - Urine Albumin 3 Living Area should be changed from the existing. Current - BMI 25.74 Approximate - BMI 21.749230344556057 Worker should be changed from the existing. 13 is the Actual GFR 22.3 is the hypothetical GFR

Medicines that could help the GFR increase for this patient: * TAB.Telma - 40 mg OD * TAB.Nephrosave Forte - 1 OD * CAP.Renerve Plus OD * TAB.Storvas - 10mg HS * TAB Folvite 5mg OD * TAB.Cardivas CR - 10mg OD * TAB cilacar 5mg HS * TAB. Telma - 20 mg * TAB.Cilacar - 10mg OD * TAB.Feburic - 40mg OD * TAB.Nephrosave OD * TAB.Nodosis - 500mg TDS * TAB.Dapaglyn 5mg OD \ufeff * TAB.Ranadyl * TAB.Obimet SR - 500mg BD * TAB.Clavix AS - OD	Medicines that could help the GFR increase for this patient: * TAB.Cilacar - 10mg OD * TAB.bisferxt OD * TAB.Oxemia - 100mg * TAB.Glimy * TAB.Storvas - 20mg HS * TAB.Dytor - 10mg OD * TAB.Alphadol - 0.25 mg * TAB.Pruvict - 1mg HS * TAB Montair * TAB.Pan 40 mg OD * TAB.Feburic - 40mg OD	* TAB.Nicardia R 30mg TDS * TAB.Revololxl 25mg BD * TAB Minipress XL 5mg OD * TAB.Dytor - 10mg OD * TAB.Pruvict - 1mg HS * TAB.Feburic - 40mg OD
60	**215**	**71**
Current - Haemoglobin Level 9.7 Approximate - Haemoglobin Level 16.34221693261607 Current - Tota Water Intake noise removed 2.0 Approximate - Tota Water Intake noise removed 2.339156351184292 Current - Random Blood Sugar 83 Approximate - Random Blood Sugar 147 Current - Systolic 140 Approximate - Systolic 140 Current - Diastolic 80	Current - Haemoglobin Level 9.5 Approximate - Haemoglobin Level 14.283711840449024 Current - Tota Water Intake noise removed 2.0 Approximate - Tota Water Intake noise removed 2.096735867897697 Current - Random Blood Sugar 89 Approximate - Random Blood Sugar 101 Current - Systolic 120 Approximate - Systolic 140 Current - Diastolic 80	Current - Haemoglobin Level 10.8 Approximate - Haemoglobin Level 13.87994229657241 Current - Tota Water Intake noise removed 2.0 Approximate - Tota Water Intake noise removed 1.9243269813898318 Current - Random Blood Sugar 204 Approximate - Random Blood Sugar 216 Current - Systolic 110 Approximate - Systolic 90 Current - Diastolic 70

Approximate - Diastolic 100 Current - Urine Albumin 1 Approximate - Urine Albumin 1 Living Area should remain the same. Current - BMI 23.97 Approximate - BMI 24.72755008790746 Worker should remain the same. 15 is the Actual GFR 30.0 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Feburic - 40mg OD * CAP.Uprise D2 60000 * TAB.Repace - 50mg BD * TAB.Dapaglyn 5mg OD \ufeff * TAB.Zincovit * TAB.Kerendia 10mg OD * TAB.Rantac 150 mg OD * TAB.Met XL - 25mg BD * TAB.Concor - 2.5mg OD	Approximate - Diastolic 80 Current - Urine Albumin 3 Approximate - Urine Albumin 4 Living Area should remain the same. Current - BMI 23.12 Approximate - BMI 24.389437090852343 Worker should remain the same. 7 is the Actual GFR 22.0 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Cilacar - 10mg OD * TAB.bisferxt OD * TAB.Oxemia - 100mg * TAB.Glimy * TAB.Storvas - 20mg HS * TAB.Rantac - 150mg BD * TAB.Alphadol - 0.25 mg * TAB.Pruvict - 1mg HS * TAB Montair * TAB.Pan 40 mg OD	Approximate - Diastolic 60 Current - Urine Albumin 3 Approximate - Urine Albumin 2 Living Area should be changed from the existing. Current - BMI 25.54 Approximate - BMI 22.119374565746785 Worker should remain the same. 14 is the Actual GFR 29.0 is the hypothetical GFR Medicines that could help the GFR increase for this patient: * TAB.Telma - 40 mg OD * TAB.Nephrosave Forte - 1 OD * CAP.Renerve Plus OD * TAB.Storvas - 10mg HS * TAB Folvite 5mg OD * TAB.Clopilet 75 OD * TAB.Cardivas CR - 10mg OD * TAB cilacar 5mg HS * TAB. Telma - 20 mg * TAB.Cilacar - 10mg OD * TAB.Feburic - 40mg OD * TAB.Nephrosave OD * TAB.Dapaglyn 5mg OD \ufeff * TAB.Ranadyl * TAB.Obimet SR - 500mg BD * TAB.Storvas - 20mg HS * TAB.Clavix AS - OD

DEPLOYMENT:

Streamlit is an open-source Python library designed to simplify the process of building interactive web applications for data science and machine learning tasks. It allows developers to create powerful, customized, and visually appealing applications directly from Python scripts, with minimal effort and without the need for web development expertise. Streamlit provides a high-level API that abstracts away the complexities of web development, enabling users to focus on data analysis, visualization, and model deployment.

Key Features:

1. Simple and Intuitive API: Streamlit offers a straightforward and easy-to-understand API that allows users to create interactive web applications using familiar Python syntax. Developers can use functions and decorators provided by Streamlit to add interactive widgets, such as sliders, buttons, and text inputs, to their applications effortlessly.

2. Built-in Components: Streamlit provides a rich set of built-in components and widgets for data visualization, including charts, tables, maps, and plots. These components can be easily customized and integrated into applications to visualize data and present insights effectively.

3. Real-time Updates: One of the standout features of Streamlit is its ability to automatically update the web application in real-time as the underlying Python script is modified. This allows developers to see changes instantly and iterate rapidly during the development process, enhancing productivity and efficiency.

4. Seamless Integration with Python Libraries: Streamlit seamlessly integrates with popular Python libraries used in data science and machine learning, such as pandas, matplotlib, scikit-learn, and TensorFlow. This enables users to leverage the full power of these libraries within their Streamlit applications, facilitating data analysis, model training, and inference.

5. Deployment Flexibility: Streamlit applications can be deployed easily to various platforms, including local servers, cloud platforms, and containerized environments. Streamlit provides a command-line interface (`streamlit run`) for running applications locally and offers options for deploying applications to cloud providers like Heroku, AWS, and Google Cloud Platform.

Usage:

1. Data Exploration and Analysis: Streamlit is commonly used for building interactive data exploration and analysis tools. Developers can create applications that allow users to upload datasets, perform data manipulation and transformation, visualize data using charts and plots, and derive insights interactively.

2. Model Prototyping and Evaluation: Streamlit is well-suited for prototyping machine learning models and evaluating their performance. Developers can build applications that showcase model predictions, evaluate model metrics, and visualize model outputs in real-time, enabling rapid experimentation and iteration.

3. Dashboarding and Reporting: Streamlit enables the creation of dynamic dashboards and reports that summarize and communicate key findings from data analysis projects. Developers can design applications that provide interactive visualizations, summary statistics, and actionable insights, facilitating decision-making and communication within organizations.

4. Education and Demonstration: Streamlit is used extensively for educational purposes, allowing instructors and educators to create interactive tutorials, demos, and workshops for teaching data science and machine learning concepts. Students can interact with these applications to gain hands-on experience and deepen their understanding of complex topics.

5. Custom Applications: Streamlit is highly customizable, allowing developers to create tailored web applications for specific use cases and domains. Whether it's building a recommendation system, a sentiment analysis tool, or a financial forecasting application, Streamlit provides the flexibility and functionality needed to bring ideas to life.

In summary, Streamlit is a powerful and versatile tool for building interactive web applications for data science and machine learning. Its simplicity, flexibility, and integration with Python libraries make it a popular choice among developers for prototyping, deploying, and sharing data-driven applications with ease.

MySQL is a widely used open-source relational database management system (RDBMS) that is renowned for its reliability, scalability, and performance. Developed by MySQL AB, it's now owned by Oracle Corporation. MySQL is utilized across various industries and applications, from small websites to large-scale enterprise systems. Here's a detailed description along with its usage:

1. Relational Database Management System (RDBMS): MySQL is an RDBMS, which means it organizes data into tables with rows and columns, allowing users to establish relationships between different sets of data. This relational structure provides flexibility in storing and querying data, making MySQL suitable for a wide range of applications.

2. Features and Capabilities:

- Data Storage: MySQL can handle structured data, including text, numbers, dates, and binary data, efficiently.

- Data Manipulation: Users can perform various operations on the data stored in MySQL, such as inserting, updating, deleting, and querying records using SQL (Structured Query Language).

- Transactions: MySQL supports transactions, ensuring data integrity by allowing multiple operations to be grouped together and executed as a single unit, either successfully or not at all.

- Security: MySQL provides robust security features, including user authentication, access control, encryption, and auditing, to protect sensitive data from unauthorized access and malicious activities.

- Scalability: MySQL offers scalability options to handle growing amounts of data and increasing workload demands. It supports replication, clustering, partitioning, and sharding for distributing data and workload across multiple servers.

- High Availability: MySQL provides features like replication, failover, and backup and recovery mechanisms to ensure high availability and fault tolerance, minimizing downtime and data loss.

- Performance Optimization: MySQL offers tools and techniques for optimizing database performance, such as indexing, query optimization, caching, and profiling, to achieve better responsiveness and throughput.

- Compatibility: MySQL is compatible with various operating systems, programming languages, and development frameworks, making it versatile and easy to integrate into existing software ecosystems.

3. Usage:

- Web Development: MySQL is extensively used in web development for powering dynamic websites and web applications. It serves as the backend database for content management systems (e.g., WordPress, Drupal), e-commerce platforms (e.g., Magento, WooCommerce), social networks, forums, and other online services.

- Data Warehousing and Analytics: MySQL is utilized for storing and analyzing large volumes of data in data warehousing and analytics applications. It serves as the storage engine for business intelligence (BI) tools, reporting platforms, and data analytics frameworks.

- Enterprise Applications: MySQL is deployed in enterprise environments for managing business-critical data and applications, such as customer relationship management (CRM), enterprise resource planning (ERP), supply chain management (SCM), and financial systems.

- Mobile and IoT Applications: MySQL is used in mobile and Internet of Things (IoT) applications for storing and syncing data between devices, managing user profiles and preferences, and facilitating real-time data processing and analysis.

- Software as a Service (SaaS): MySQL is employed by SaaS providers to power multi-tenant cloud applications and services, where multiple users or organizations share a common database infrastructure while maintaining data isolation and security.

- Education and Research: MySQL is widely adopted in educational institutions and research organizations for teaching database concepts, conducting experiments, and prototyping applications due to its ease of use, availability, and community support.

In summary, MySQL is a versatile and robust relational database management system that caters to a diverse range of use cases and industries. Its rich feature set, performance, scalability, and reliability make it a preferred choice for developers, businesses, and organizations worldwide. Whether it's powering websites, managing enterprise data, or

driving innovation in emerging technologies, MySQL continues to play a pivotal role in the world of data management and application development.

Certainly! Let's delve into the usage of each of these packages:

1. scikit-learn:

- Description: scikit-learn is a popular machine learning library in Python, providing simple and efficient tools for data mining and data analysis. It is built on top of NumPy, SciPy, and matplotlib.

- Usage: scikit-learn offers various machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more. Users can preprocess data, train models, and evaluate their performance using this library. It also includes utilities for data splitting, cross-validation, hyperparameter tuning, and model selection. The library is well-documented and widely used in both academia and industry for building machine learning pipelines.

2. statsmodels:

- Description: statsmodels is a Python module that provides classes and functions for estimating many different statistical models and performing statistical tests. It is heavily influenced by R and is designed for statistical modeling and hypothesis testing.

- Usage: statsmodels offers a wide range of statistical models, including linear regression, generalized linear models, time series analysis, and more. Users can perform hypothesis tests, explore data, and estimate parameters using this library. It also provides tools for visualization and diagnostics of statistical models. statsmodels is often used in academic research, econometrics, and social sciences for statistical analysis and modeling.

3. pandas:

- Description: pandas is a powerful data manipulation and analysis library for Python. It provides data structures like Series and DataFrame, which allow users to efficiently work with structured data.

- Usage: pandas is widely used for data cleaning, transformation, and analysis tasks. It offers functionality for reading and writing data from various formats like CSV, Excel, SQL databases, and more. Users can perform operations like indexing, slicing, filtering, grouping, merging, and reshaping data with pandas. It also includes tools for handling missing data, time series data, and categorical data. pandas is a fundamental tool in the data science toolkit and is often used in conjunction with other libraries like NumPy and scikit-learn.

4. sqlite3:

- Description: sqlite3 is a lightweight, embedded SQL database engine written in C. In Python, the sqlite3 module provides a simple way to work with SQLite databases directly from Python code.

- Usage: sqlite3 allows users to create, connect to, and manipulate SQLite databases using Python. Users can execute SQL queries, manage transactions, and interact with databases programmatically. SQLite databases are self-contained, serverless, and file-based, making them suitable for small to medium-sized applications or prototyping. sqlite3 is often used for local data storage, caching, and lightweight applications where a full-fledged database server is not necessary.

5. matplotlib and seaborn:

- Description: matplotlib is a comprehensive library for creating static, interactive, and animated visualizations in Python. seaborn is a statistical data visualization library built on top of matplotlib, providing a higher-level interface for creating attractive and informative statistical graphics.

- Usage: matplotlib offers a wide range of plotting functions and styles for creating various types of plots, including line plots, bar plots, scatter plots, histograms, heatmaps, and more. seaborn extends matplotlib by providing additional statistical plots and themes that make it easier to create visually appealing plots for exploratory data analysis and presentation. Both libraries are extensively used in data visualization, scientific computing, and data-driven storytelling. They offer flexibility and customization options to create publication-quality visualizations for conveying insights from data.

6. Numpy

:NumPy is a fundamental Python library for numerical computing, providing support for large, multi-dimensional arrays and matrices. With its extensive collection of mathematical functions, NumPy facilitates efficient data manipulation and computation. It serves as a cornerstone for many scientific computing tasks, including linear algebra operations, statistical analysis, and signal processing. NumPy's array-oriented computing paradigm allows for concise and expressive code, enabling faster execution compared to traditional Python lists. Additionally, NumPy seamlessly integrates with other libraries in the scientific Python ecosystem, such as SciPy, pandas, and Matplotlib. Its versatility and performance make it indispensable for data scientists, researchers, and engineers tackling complex numerical problems.

NEW ENTRY:

When entering a new entry into the database of the applications there may be many anomalies as people tend to make mistakes and those mistakes can lead to an Exception or any error in the pipelines in the Streamlit application. And there are millions of way a person could make mistakes in data entry.

To avoid those silly tiny yet impactful mistakes, we validate the data entered and if the entered one does not pass on the validation the data is asked for a verification from the data entry operator or the doctors

.

The validations made are as follows:

Attribute	Lowest value possible	Highest value possible	Range
Systolic pressure	0	210	0,210
Age	10	98	10,98
Height (CMS)	100	195	100,195
Weight (kgs)	20	210	20,210
GFR	0	120	0,120
Serum Creatinine	0	30	0,30
Haemoglobin	0	18	0,18
Total water intake	1	6	1,6
Total Count	1000	20000	1000,20000
Random Blood sugar	0	500	0,500
Diastolic Pressure	0	130	0,130
Urea	0	300	0,300

CURRENT RESULT:

An intuitive AI model which analyses patient data, recommends biological and lifestyle suggestions, medicine recommendation and changes, predicts the Glomerular Filtration rate for the next three years for both existing and newly incoming patients.