**PROJECT REPORT**

**ON**

**CHRONIC KIDNEY DISEASE**

**PREDICTION USING WATSON**

**AUTO AI**

## BY

**PREETHI.K (preethikarthikeyan99@gmail.com)**

# Chronic Kidney Disease Prediction Using

# Watson Auto AI

### INTRODUCTION

#### 1.1 OVERVIEW

Chronic Kidney Disease (CKD) is a termed generally to describe disorders affecting kidney structure and function. CKD becomes a major health problem for the underdeveloped countries of southeast Asia, home to more than 2 billion people. CKD may be triggered by diabetes, high blood pressure and other disorders. Early detection and treatment can often keep CKD from getting worse to save human life.

#### 1.2 PURPOSE

This project helps in predicting whether the patient has chronic kidney disease or not using Watson Auto AI. The Watson Auto AI helps in finding the best suitable algorithm for the given data set and that algorithm will be saved as model for future prediction. The Watson Auto AI automatically trains the model and helps us in predicting the result.

### 2. LITERATURE SURVEY

#### 2.1 EXISTING PROBLEM

The overall goal of ML is to generate automatic models being able to rapidly generalize (classify) from the examples observed a priori, and it generates by designing or learning functional dependencies among the selected input (features) and output (classes) domains. Diagnosis of Chronic kidney disease (CKD), which is aimed to translate the knowledge from the extracted features (symptoms) into meaningful groups (groups of healthy individuals, CDK individual or individuals with the some other type of disorder), is thus fundamentally a ML problem. The main implication of the CKD diagnosis is the kidney damage. Several symptoms or risk factors are also associated with the CKD progress, so that, these variables could highly influence CKD recognition. By observing the progressive nature of CKD, new insights from the diagnostic computational models grounded on the ML paradigms offer the great promise to advance the diagnosis of CKD.

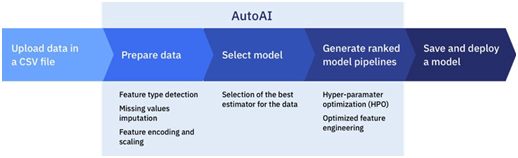
#### 2.2 PROPOSED SOLUTIONS

Several studies have suggested the models grounded on the fuzzy logic for diagnosis of CKD. The contribution of this study is to find a simple classifier which gives better classification accuracy. In our study, we consider well known machine learning methods, artificial neural network (ANN), support vector machine (SVM), k-nearest neighbor (k-NN), C4.5 decision tree and random forest (RF) to produce the highly confident model for the CKD diagnosis. To prove the efficiency and effectiveness of these machine learning tools, we experimentally validated our proposed methods on the real data (CKD data set).

### 3. THEORITICAL ANALYSIS

#### 3.1 BLOCK DIAGRAM

In this project, we have taken our data set from Kaggle Machine Learning Repository called Chronic Kidney Disease for study. In this data set there are 400 instances categorized using 25 attributes. Out of these 25 attributes there are 11 numeric and 14 Boolean attributes. In this project we used 90% instances for training purpose and 10% instances for testing purpose. Pre-processing is performed on data set to handle noisy and missing data. By this method we obtained an accuracy of 99.3%.



#### 3.2 HARDWARE/SOFTWARE DESIGNING

The AutoAI graphical tool in Watson Studio automatically analyzes your data and generates candidate model pipelines customized for your predictive modeling problem. These model pipelines are created iteratively as AutoAI analyzes your data set and discovers data transformations, algorithms, and parameter settings that work best for your problem setting. Results are displayed on a leader board, showing the automatically generated model pipelines ranked according to your problem optimization objective.

**Required service:** Watson Machine Learning service

**Data format:** Tabular: CSV files, with comma (,) delimiter

**Data size:** Less than 1 GB

Strategic investments in AI can be a game changer. To fulfill the promise of AI, organizations are tackling skill-set gaps, deployment and governance processes today. In particular, businesses are seeking an alternative where citizen data scientists can quickly get started and expert data scientists can speed experimentation time from weeks and months to minutes and hours. They need a multi modal data science and AI environment where data and analytics specialists collaborate with other experts and optimize model performance end-to-end.

To help simplify an AI lifecycle management, AutoAI automates:

* Data preparation
* Model development
* Feature engineering
* Hyper-parameter optimization

AutoAI is available within [IBM Watson® Studio](https://www.ibm.com/in-en/cloud/watson-studio) with one-click deployment through [Watson Machine Learning](https://www.ibm.com/in-en/cloud/machine-learning).

### Data pre-processing:

Most data sets contain different data formats and missing values, but standard machine learning algorithms work with numbers and no missing values. AutoAI applies various algorithms, or estimators, to analyze, clean, and prepare your raw data for machine learning. It automatically detects and categorizes features based on data type, such as categorical or numerical. Depending on the categorization, it uses hyper-parameter optimization to determine the best combination of strategies for missing value imputation, feature encoding, and feature scaling for your data.

### Automated model selection:

The next step is automated model selection that matches your data. AutoAI uses a novel approach that enables testing and ranking candidate algorithms against small subsets of the data, gradually increasing the size of the subset for the most promising algorithms to arrive at the best match. This approach saves time without sacrificing performance. It enables ranking a large number of candidate algorithms and selecting the best match for the data.

### Automated feature engineering:

Feature engineering attempts to transform the raw data into the combination of features that best represents the problem to achieve the most accurate prediction. AutoAI uses a unique approach that explores various feature construction choices in a structured, non-exhaustive manner, while progressively maximizing model accuracy using reinforcement learning. This results in an optimized sequence of transformations for the data that best match the algorithms of the model selection step.

### Hyperparameter optimization:

Finally, a hyper-parameter optimization step refines the best performing model pipelines. AutoAI uses a novel hyper-parameter optimization algorithm optimized for costly function evaluations such as model training and scoring that are typical in machine learning. This approach enables fast convergence to a good solution despite long evaluation times of each iteration.

### 4. EXPERIMENTAL INVESTIGATIONS

So my project aim is to classify whether the patient is having Chronic Kidney Disease or not. The data set has 26 columns and 400 records. This data set is used to train the Machine Learning model. Thus the first step is   
  
 **a. Pre-processing of data set:**

The ID column from the data set is removed and further the dataset's missing values are replaced by NaN. The categorical values or nominal values are converted into binary values for training. Then the NaNs are replaced with its respective columnar means. The resultant data set is converted to csv and stored. Using the resultant data set the correlation between the attributes is found to reduce or remove the attributes to get better accuracy. The highly correlated attributes are retained by dropping the other attributes that have low correlation. Also, further in this data set, the binary values are converted into its nominal values for uploading it in the Watson Auto AI model for prediction.

**b. IBM Cloud Account:**

I have created an IBM Cloud account and created a Cloud Object Storage service to store all my data set and model. Then I created Watson Studio instance and Machine Learning Service.

**c. Model Building:**

I have created a project in Watson Studio named Chronic Kidney Disease Prediction and in that I have created a Auto AI Experiment. Once created I imported the data set final\_modified\_ckd\_for\_prediction.csv into the Auto AI experiment and provided Class as prediction column and Ran the model. After training the data set, it showed Random Forest Classifier to have accuracy of 0.992. I then saved it as model and deployed it.

**d. Application Building:**

I again created a Node Red App instance in cloud to deploy my model in an application. Once created, I downloaded the node-red-dashboard from manage palette in Node Red service. Then I created the flow using form,functions,HTTP requests,text and debug. The form contains all the variable labels, names and their types. The PreToken function will have all the global variables, api key, headers and payload. The HTTP request has the URL token through which the UI is displayed. The Pre Prediction function will get all the values from the former HTTP request and sends it to the latter HTTP request function which has the scoring end point URL where the model was trained and the output gets displayed in the Debug messages. The Output function gets the output from Debug messages and sends it to the Prediction text and gets displayed.

### 5. FLOWCHART

Insert the following nodes into the flow in Node-Red.

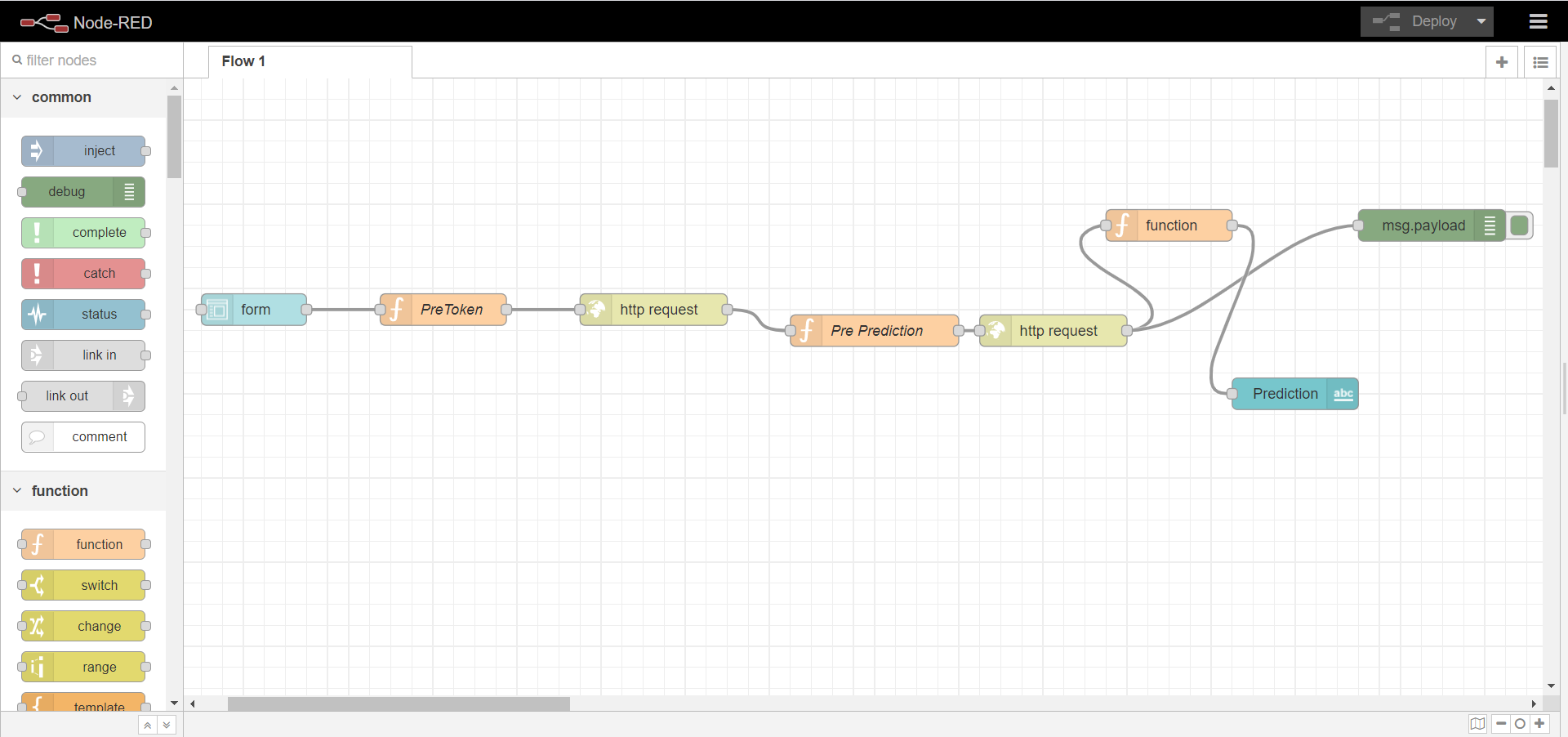
● UI\_Form

● Input

● Function

● Http request

● Text

● Debug 

### 6. RESULTS

The Auto AI model Found two best machine learning algorithms: Random Forest Classifier an Extra Tress Classifier. Accuracy (ACC) is the overall success rate of the classifier defined as

ACC = (TP + TN) / (TP + FP + TN + FN)

Where

TP - the number of true positives.

TN - the number of true negatives.

FP - the number of false positives.

FN - the number of false negatives.

Table :Performance evaluation

|  |  |
| --- | --- |
| Machine Learning Algorithms sued | Accuracy(%) |
| Random Forest Classifier | 99.2 |
| Extra Trees Classifier | 98.9 |

From table by comparing the Random Forest Classifier and Extra Trees Classifier, the accuracy of Random Forest Classifier is relatively higher than the Extra Trees Classifier. The Random Forest Classifier can be adopted since it has the accuracy of 99.2% in prediction of chronic kidney disease.

Web based UI was developed by integrating all the services using NODE-RED URL for UI dashboard : <https://node-red-arpot.mybluemix.net/>ui

### 7. ADVANTAGES & DISADVANTAGES

**Pros:**

1. It can judge the importance of the feature
2. Can judge the interaction between different features
3. Not easy to overfit
4. Training speed is faster, easy to make parallel method
5. It is relatively simple to implement
6. For unbalanced data sets, it balances the error.
7. If a large part of the features are lost, accuracy can still be maintained.
8. It can come out with very high dimensional (features) data, and no need to reduce dimension, no need to make feature selection.

**Cons:**

1. For data with different values, attributes with more values will have a greater impact on random forests, so the attribute weights generated by random forests on such data are not credible.
2. Random forests have been shown to fit over certain noisy classification or regression problems.

### 8. APPLICATION

The random forest algorithm is used in a lot of different fields, like banking, the stock market, medicine and e-commerce.

In finance, for example, it is used to detect customers more likely to repay their debt on time, or use a bank's services more frequently. In this domain it is also used to detect fraudsters out to scam the bank. In trading, the algorithm can be used to determine a stock's future behavior.

In the healthcare domain it is used to identify the correct combination of components in medicine and to analyze a patient’s medical history to identify diseases.

Random forest is used in e-commerce to determine whether a customer will actually like the product or not.

### 9. CONCLUSION

The prediction of chronic kidney disease is very important and now-a-days it is the leading cause of death.The performance of Random Forest Classifier was found to be 99.2% accurate compared to Extra Trees Classifier. Through this model and Node Red application, an UI is created to view the prediction for given set of inputs.

### 10. FUTURE SCOPE

There are other possible evolutionary techniques that may be used to improve results of the proposed classifiers. In this paper, RFC and ETCare applied to detect CKD. We can also evaluate and compare the performance of the used classifiers with other existing classifiers. CKD early detection helps in timely treatment of the patients suffering from the disease and also to avoid the disease from getting worse. Early prediction of the disease and timely treatment are the need for medical sector. New classifiers can be used and their performance can be evaluated to find better solutions of the objective function in future work.

### 11. BIBILOGRAPHY

1. Auto AI with IBM Watson studio:

https://www.ibm.com/in-en/cloud/watson-studio/autoai

2. Node-RED starter application:

[https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-applicatio n/](https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/)

### 12. APPENDIX

**A. Source code**

[{"id":"379a755f.16c60a","type":"tab","label":"Flow 1","disabled":false,"info":""},{"id":"4616f9fa.8e5a78","type":"ui\_form","z":"379a755f.16c60a","name":"","label":"","group":"6e9c8097.9a176","order":1,"width":"0","height":"0","options":[{"label":"Specific Gravity","value":"Sg","type":"number","required":true,"rows":null},{"label":"Albumin","value":"Al","type":"number","required":true,"rows":null},{"label":"Hemoglobin","value":"Hemo","type":"number","required":true,"rows":null},{"label":"Packed Cell Volume","value":"Pcv","type":"number","required":true,"rows":null},{"label":"Red Blood Cell Count","value":"Rbcc","type":"number","required":true,"rows":null},{"label":"Hypertension","value":"Htn","type":"text","required":true,"rows":null},{"label":"Diabetes Mellitus","value":"Dm","type":"text","required":true,"rows":null}],"formValue":{"Sg":"","Al":"","Hemo":"","Pcv":"","Rbcc":"","Htn":"","Dm":""},"payload":"","submit":"Predict","cancel":"cancel","topic":"","x":170,"y":260,"wires":[["987deb07.35b108"]]},{"id":"987deb07.35b108","type":"function","z":"379a755f.16c60a","name":"PreToken","func":"global.set(\"Sg\",msg.payload.Sg)\nglobal.set(\"Al\",msg.payload.Al)\nglobal.set(\"Hemo\",msg.payload.Hemo)\nglobal.set(\"Pcv\",msg.payload.Pcv)\nglobal.set(\"Rbcc\",msg.payload.Rbcc)\nglobal.set(\"Htn\",msg.payload.Htn)\nglobal.set(\"Dm\",msg.payload.Dm)\nvar apikey=\"nIRjhISx910dr1kYf7uCwWpp4MxI4ab0Zy5q3xY5Kvmp\";\nmsg.headers={\"Content-Type\":\"application/x-www-form-urlencoded\"}\nmsg.payload={\"grant\_type\":\"urn:ibm:params:oauth:grant-type:apikey\",\"apikey\":apikey}\nreturn msg;","outputs":1,"noerr":0,"initialize":"","finalize":"","x":350,"y":260,"wires":[["3875565c.f7b1da"]]},{"id":"3875565c.f7b1da","type":"http request","z":"379a755f.16c60a","name":"","method":"POST","ret":"obj","paytoqs":"ignore","url":"https://iam.bluemix.net/identity/token","tls":"","persist":false,"proxy":"","authType":"","x":550,"y":260,"wires":[["8ba7d0a6.647eb"]]},{"id":"8ba7d0a6.647eb","type":"function","z":"379a755f.16c60a","name":"Pre Prediction","func":"var Sg=global.get('Sg')\nvar Al=global.get('Al')\nvar Hemo=global.get('Hemo')\nvar Pcv=global.get('Pcv')\nvar Rbcc=global.get('Rbcc')\nvar Htn=global.get('Htn')\nvar Dm=global.get('Dm')\nvar token=msg.payload.access\_token\nvar instance\_id=\"f36085ae-a511-408c-8fa5-6a5900b8f5aa\"\nmsg.headers={\"Content-Type\":\"application/json\",\"Authorization\":\"Bearer \"+token ,\"ML-Instance-ID\":instance\_id}\nmsg.payload={\"input\_data\": [{\"fields\": [\"Specific Gravity\", \"Albumin\", \"Hemoglobin\", \"Packed Cell Volume\", \"Red Blood Cell Count\", \"Hypertension\", \"Diabetes Mellitus\"],\"values\": [[Sg,Al,Hemo,Pcv,Rbcc,Htn,Dm]]}]}\nreturn msg;","outputs":1,"noerr":0,"initialize":"","finalize":"","x":760,"y":280,"wires":[["eef491dd.9a155"]]},{"id":"eef491dd.9a155","type":"http request","z":"379a755f.16c60a","name":"","method":"POST","ret":"obj","paytoqs":"ignore","url":"https://us-south.ml.cloud.ibm.com/v4/deployments/9da1dab4-5204-444b-a90b-d3575f80d0c1/predictions","tls":"","persist":false,"proxy":"","authType":"","x":930,"y":280,"wires":[["fabeae19.bb97c","cb35e3cf.d49d2"]]},{"id":"fabeae19.bb97c","type":"function","z":"379a755f.16c60a","name":"","func":"msg.payload=msg.payload.predictions[0].values[0][0]\nreturn msg;","outputs":1,"noerr":0,"initialize":"","finalize":"","x":1040,"y":180,"wires":[["d10139c9.086cd8"]]},{"id":"d10139c9.086cd8","type":"ui\_text","z":"379a755f.16c60a","group":"6e9c8097.9a176","order":2,"width":0,"height":0,"name":"","label":"Prediction","format":"{{msg.payload}}","layout":"row-spread","x":1160,"y":340,"wires":[]},{"id":"cb35e3cf.d49d2","type":"debug","z":"379a755f.16c60a","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","statusVal":"","statusType":"auto","x":1290,"y":180,"wires":[]},{"id":"6e9c8097.9a176","type":"ui\_group","z":"","name":"Chronic Kidney Disease Prediction","tab":"8b2f4ce7.7a627","order":1,"disp":true,"width":"8","collapse":false},{"id":"8b2f4ce7.7a627","type":"ui\_tab","z":"","name":"Home","icon":"dashboard","disabled":false,"hidden":false}]

**B. UI output Screenshot.**

