Research Article

**Chronic Kidney Disease Prediction Using Machine Learning Algorithms**

**Rishita Agarwal**

**Prof. Manish Gupta**

School of Engineering and Technology, Amity University, Madhya Pradesh

Assistant Professor, School of Engineering and Technology, Amity University, Madhya Pradesh

**ABSTRACT**

The field of biosciences have progressed to a bigger degree and have produced a lot of data from Electronic Health Records. This have brought about the intense need of information age from this colossal measure of information. Information mining techniques and AI assume a significant part in this part of biosciences. Chronic Kidney Disease (CKD) is a condition wherein the kidneys are harmed and can't channel blood as they generally do. A family background of kidney infections or disappointment, hypertension, type 2 diabetes might prompt CKD. This is an enduring harm to the kidney and chances of getting worser by time is high. The extremely normal entanglements that outcomes because of a kidney disappointment are heart infections, iron deficiency, bone illnesses, high potassium and calcium. The most pessimistic scenario circumstance prompts total kidney disappointment and requires kidney relocate to live. An early identification of CKD can work on the personal satisfaction indeed. This calls for good expectation calculation to anticipate CKD at a prior stage. Writing shows a wide scope of AI calculations utilized for the expectation of CKD. This paper utilizes information preprocessing, information change and different classifiers to anticipate CKD and furthermore proposes best Prediction structure for CKD. The aftereffects of the structure show promising consequences of better expectation at a beginning phase of CKD.

1. **Introduction**

The inability of the kidneys to play out their customary blood sifting capacity and others is called Chronic Kidney Disease (CKD). The expression "CHRONIC" portrays the sluggish corruption of the kidney cells throughout an extensive stretch of time. This illness is a significant kidney disappointment where the kidney sans blood sifting cycle and there is a weighty liquid development in the body. This prompts disturbing increment of potassium and calcium salts in the body. Presence of elevated degrees of these salts bring about different diseases in the body. The great work of kidneys is to channel additional water and squanders from blood. The effective working of this cycle is vital to adjust the salts and minerals present in our body. The high equilibrium of salts is important to control pulse, enact chemicals, assemble red platelets, and so on. A high convergence of calcium prompts different bone sicknesses and cystic ovaries in ladies. CKD additionally may prompt unexpected disease or aversion to specific meds. This state is called as Acute Kidney Injury (AKI). An expanded circulatory strain might prompt heart issues and cardiovascular failures. CKD much of the time prompts super durable dialysis or kidney transfers. A past filled with kidney sickness in the family additionally prompts high likelihood of CKD. Writing shows that just about one out of three individuals determined to have diabetes have CKD. Writing additionally presents confirmations of early distinguishing proof and care of CKD can work on the nature of the patient's life.

With the availability of biomedical data, the use of machine-learning techniques in healthcare for developing disease prediction models has become common. Further, strategies, for example, profound learning and procedures like troupe learning have incredibly worked on the prescient force of AI models. By getting highlights from Electronic Health Records (EHR), exact sickness forecast models can be created. At the patient level, a doctor can survey the beginning of CKD utilizing research facility tests by taking a gander at standard boundaries, for example, the glomerular filtration rate (GFR) and the albumin creatinine proportion. Then again, from the general wellbeing viewpoint, research center information is ordinarily not accessible for an enormous scope. Be that as it may, two sorts of information can by and large be separated from the insurance agency's data sets: conclusions and drugs for every patient's visit at the medical clinic. Expectation calculations in AI can be astutely used to foresee the event of CKD and presents a technique for early prescription. The definite audit on writing shows the use of different AI calculations to anticipate CKD.

The rest of the paper is divided into four sections. Section1 consists of the introduction, Section 2 consists of the methodology used, Section 3 consists of the proposed System, Section 4 consists of the results and data analysis, and Section 5 consists of conclusion.

1. **Methodology**

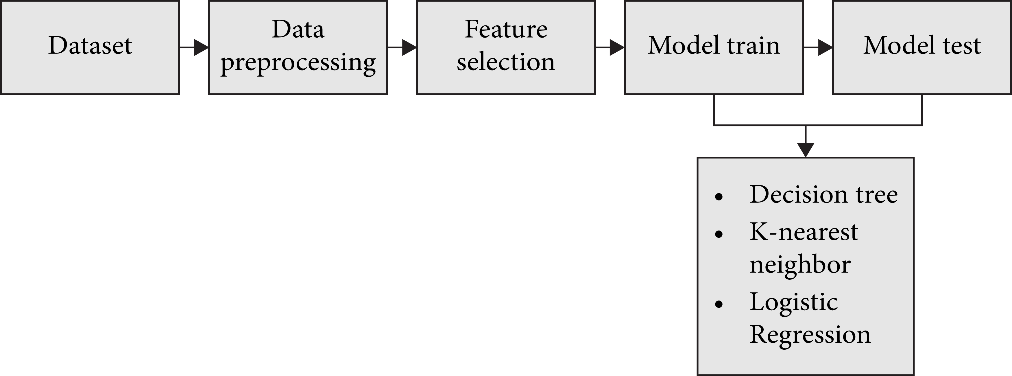
Description of the Dataset. The dataset utilized for this examination reason was the Public Health Dataset and it is dating from 2015. It contains 25 attributes. It is integer-valued 0 = no disease and 1 = disease. Now the attributes that are taken in this research purpose are described as follows:

|  |  |  |
| --- | --- | --- |
| S.NO | ATTRIBUTE | MEANING |
|  | age | age |
|  | bp | Blood Pressure |
|  | sg | Specific Gravity |
|  | al | Albumin |
|  | su | Sugar |
|  | rbc | Red Blood Cells |
|  | pc | Pus Cell |
|  | pcc | Pus Cell Clumps |
|  | ba | Bacteria |
|  | bgr | Blood Glucose Random |
|  | bu | Blood Urea |
|  | sc | Serum Creatinine |
|  | sod | Sodium |
|  | pot | Potassium |
|  | hemo | Hemoglobin |
|  | pcv | Packed Cell Volume |
|  | wc | White Blood Cell Count |
|  | rc | Red Blood Cell Count |
|  | htn | Hypertension |
|  | dm | Diabetes Mellitus |
|  | cad | Coronary Artery Disease |
|  | appet | Appetite |
|  | pe | Pedal Edema |
|  | ane | Anemia |
|  | class | Class |

1. **Proposed System**

This part depicts the dataset and contains block diagrams, flow diagrams, evaluation matrices, and the study’s procedure and methodology.

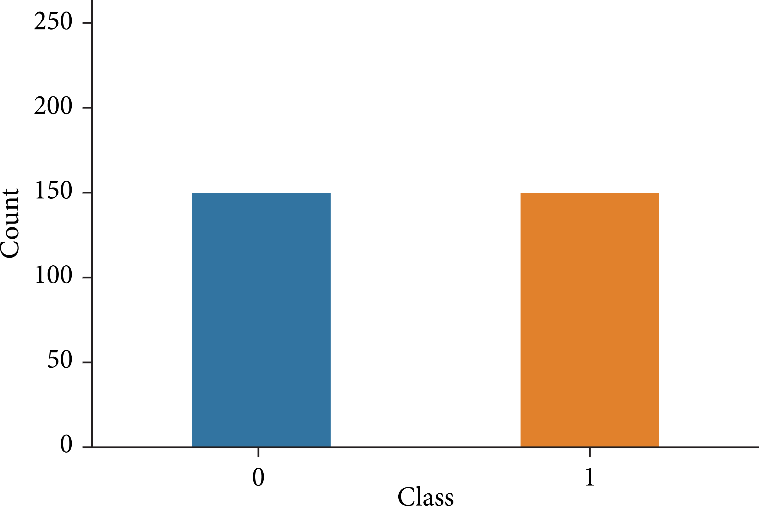
Figure [1](https://www.hindawi.com/journals/cmmm/2021/6141470/fig1/) depicts the block diagram of the proposed system. The framework utilizes the CKD prediction dataset. After pre-processing and feature selection, the DT, KNN, and logistic regression algorithms have been utilized. Every one of the parts of this outline have been talked about in the accompanying sub areas.



**Fig 1: - Block Diagram for the proposed system**

* 1. **Datasets**

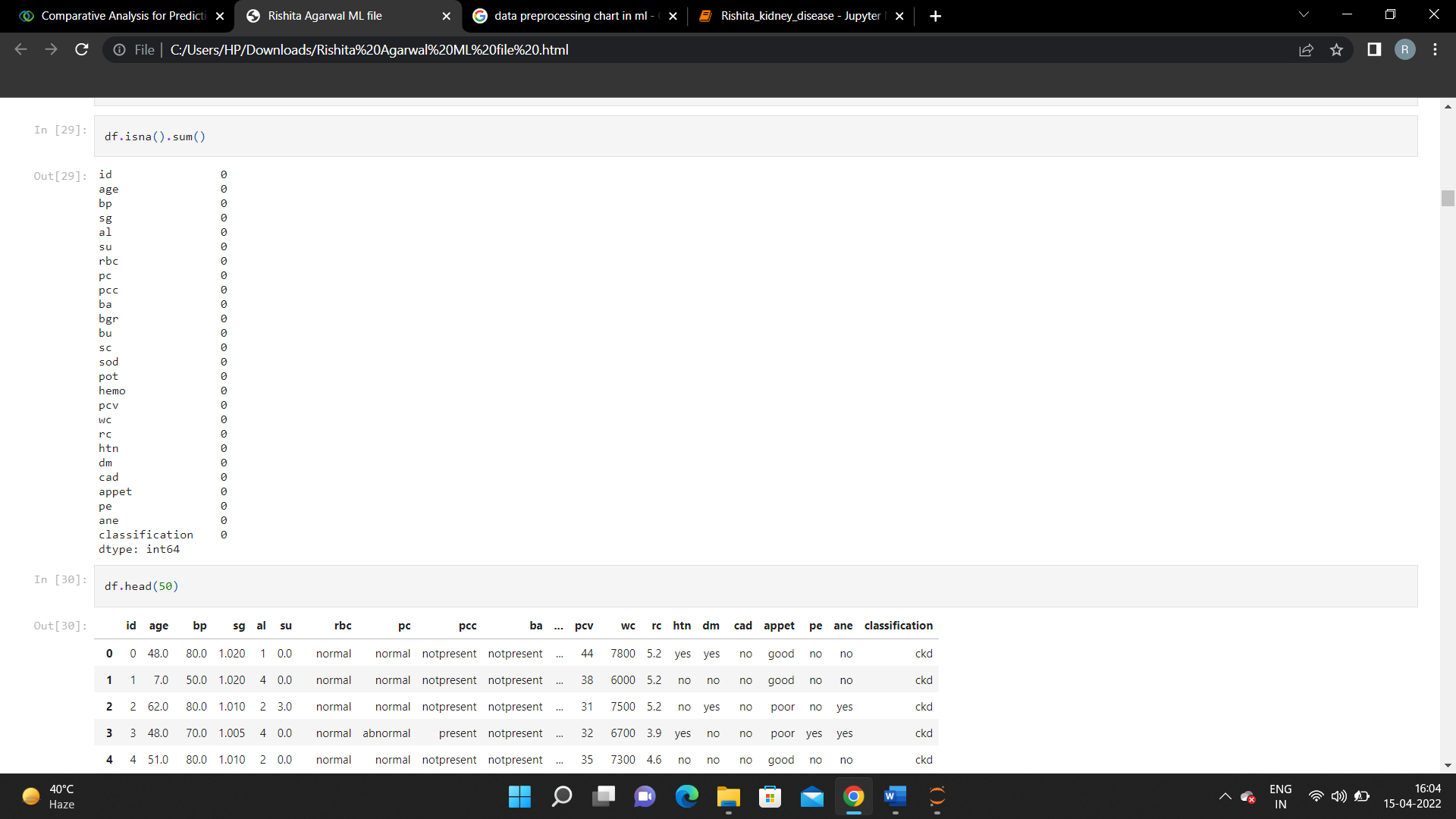
The research was led by using the CKD dataset. There are 400 rows and 26 columns in this dataset. The output column “class” has a value of either “1” or “0.” The value “0” indicates that the patient is not a CKD patient, while the value “1” shows that the patient is a CKD patient. Before preprocessing, Figure 2 displays the total number of CKD and non-CKD entries in the output column. The overall number of CKD data is 248, whereas the total number of non-CKD data is 150.



**Fig 2: Total amount of CKD and Not-CKD data.**

* 1. **Data Preprocessing**

Prior to model building, data preprocessing is required to remove unwanted noise and expectations from the dataset that might cause the model to diverge from the proper training set. This stage tackles anything that is impeding the model’s efficiency. After collecting the necessary data, it must be cleaned and prepared for model construction. The dataset is next searched for null values. However, this dataset contains no null values. Figure 3 shows there is no missing data available in this dataset.

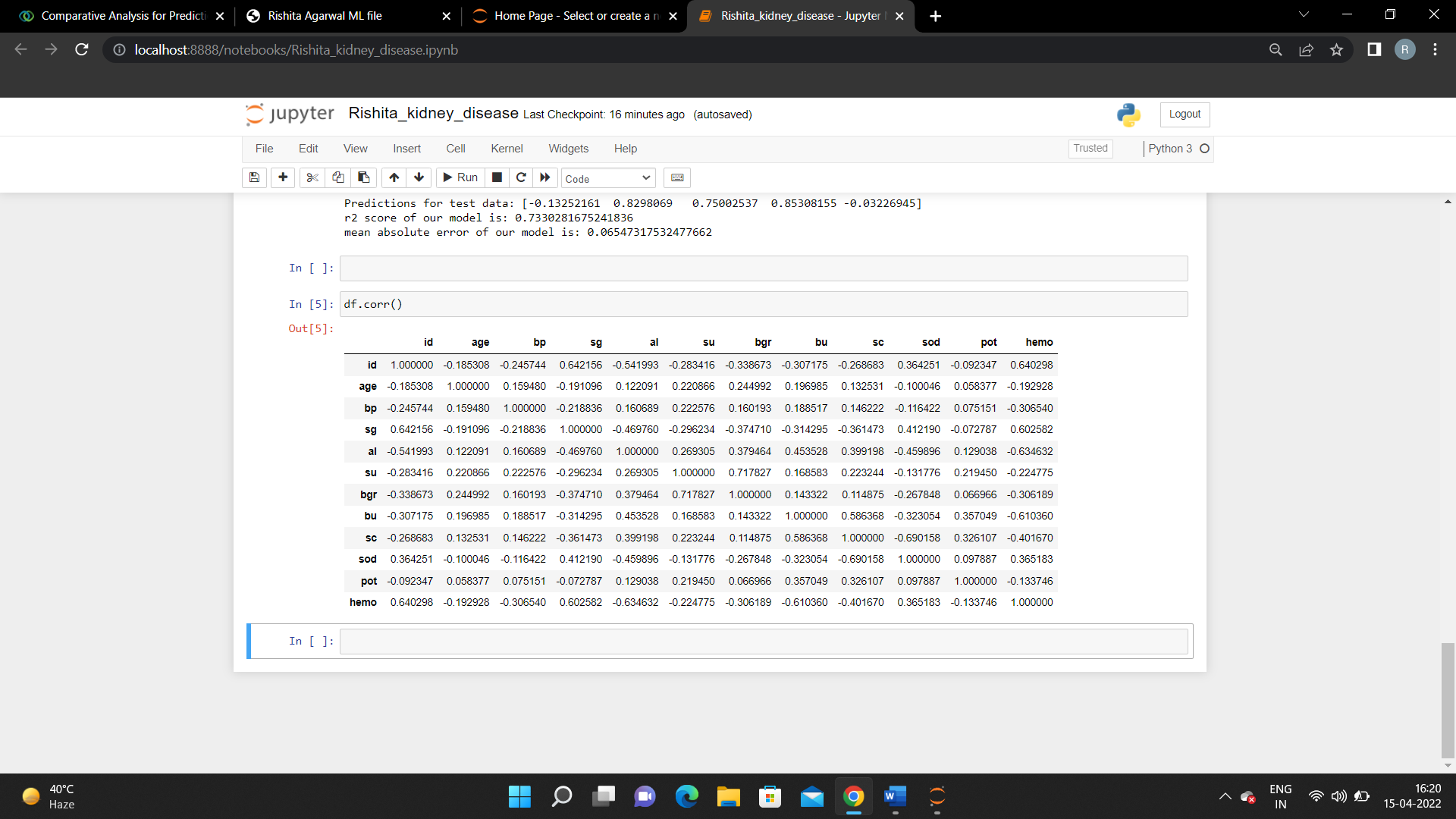


**Fig 3: No missing data**

Here, the output values “0” indicate the absence of null values. After completing data preparation and handling the unbalanced dataset, the next step is to build the model. To increment the exactness and productivity of this assignment, the information is parted into training and testing portions, with a 80/20 proportion of training to testing. Following the model’s splitting, it is trained using a number of classification techniques. The classification methods used in this research include the Linear Regression, Logistic Regression, Decision tree classification method, SVM, Random Forest and K-Neighbor Classifier.

* 1. **Feature Selection**

In this, the absolute values of the relationships between features and the class label show that blood pressure, albumin, sugar, blood urea, serum creatinine, potassium, white blood cell count, and hypertension all have positive links. Figures 4 show the feature correlation value.



**Fig 4: Feature Correlation values**

All the positively correlated features are considered for further prediction. Each albumin molecule has just five distinct sets of values. The quantity of albumin is assessed using a urine protein test. A high protein level in the urine means that the filtration units in the kidneys have been damaged by disease, fever, or intense activity. Numerous tests should be performed over many weeks to establish the diagnosis. The term serum creatinine is used interchangeably with blood creatinine and creatinine. Creatinine is the byproduct of muscle breakdown of the chemical creatine. The kidneys eliminate creatinine from the body. This test is done to find out how much creatinine is in your blood. Creatine is an element of the metabolic cycle that produces the energy needed for muscle contraction. The body produces both creatine and creatinine at the same rate. Creatinine levels in the blood can rise due to a high-protein diet, congestive heart failure, diabetic issues, and dehydration, among other factors. Creatinine levels in women should be between 0.6 and 1.1 mg/dL, while those in males should be between 0.7 and 1.3 mg/dL. Additionally, hypertension, or high blood pressure, develops whenever blood pressure against the walls of blood vessels rises. Hypertension can lead to heart attacks, strokes, and chronic kidney disease if it is not treated or managed properly. Nonetheless, CKD may result in hypertension.

* 1. **Algorithms**

The following machine learning algorithms have been used to predict chronic kidney disease: -

1. Linear Regression
2. Logistic Regression
3. Support Vector Machine (SVM)
4. Decision Tree
5. Random Forest
6. K-Neighbor Classifier

**3.1 (a) Linear Regression**

**Linear Regression** is a machine learning algorithm based on**supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Fig 5 shows the simple Linear Regression in Machine learning.

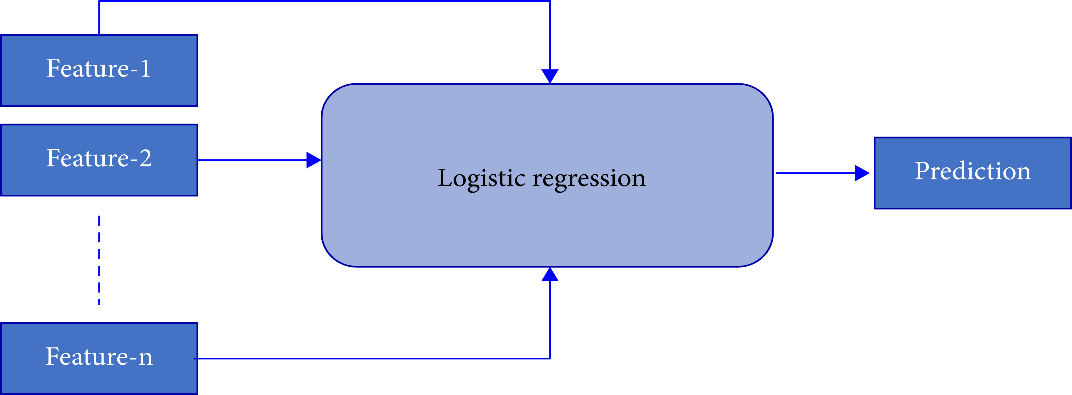


**Fig 5: - Linear Regression in Machine Learning**

**3.1 (b) Logistic Regression**

Binary outcomes are modelled using the statistical method of logistic regression, which is well known in the field. Different learning methods are used to execute logistic regression in statistical research. A variant of the neural network method was used to create the LR algorithm. This method resembles neural networks in many ways, but it is simpler to set up and use. Figure 6 shows the block diagram of Logistic Regression.

Utilizing logistic regression, the output of a categorical dependent variable is predicted. So, the output must be discrete or categorical. It may be yes or no, 0 or 1, true or false, etc., but probability values between 0 and 1 are given. Logistic regression and linear regression are used in very similar ways. Classification problems are addressed with logistic regression, and regression problems are addressed using linear regression. Instead of a regression line, we use an “S” shaped logistic function that predicts two maximum values (0 or 1). The logistic function’s curve indicates the probability of anything, such as whether cells are malignant or not, or if an animal is fat or not. Since it can classify new data using both discrete and continuous datasets, logistic regression is a common ML technique.



**Fig 6: - Block Diagram of logistic Regression**

**3.1 (c) Support Vector Machine (SVM)**

A linear model for arrangement and relapse is Support Vector Machine (SVM) that can be utilized to tackle both direct and non-direct issues. The calculation orders information utilizing a hyperplane. In this algorithm, every information thing will be plotted as a point in n-layered space (where n is the quantity of elements) with the worth of each component being the worth of a specific direction. Arrangement will be performed by observing the right hyper-plane which can separate the two classes effectively. Fig 7 shows SVM Algorithm.

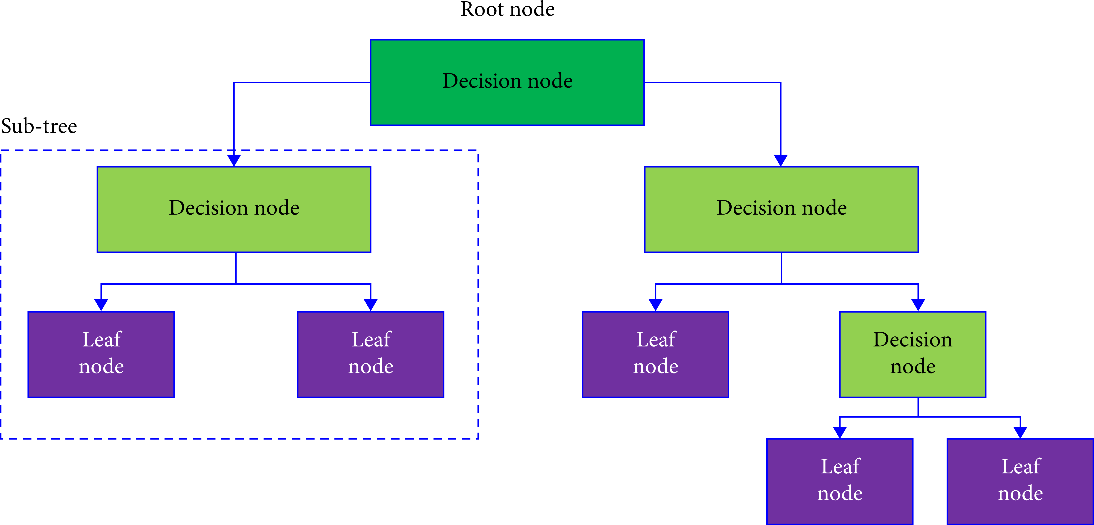


**Fig 7: - Support Vector Machine Algorithm**

**3.1 (d) Decision Tree Classifier**

The DT method is a classification and regression technique that can be used to predict both discrete and continuous characteristics. Based on the links between input columns in a dataset, the algorithm predicts discrete characteristics. It predicts the states of a column that you identify as predictable using the values of those columns, known as states. The method specifically finds the input columns that are associated with the predicted column. The DT classifier’s block diagram is shown in Figure 8.

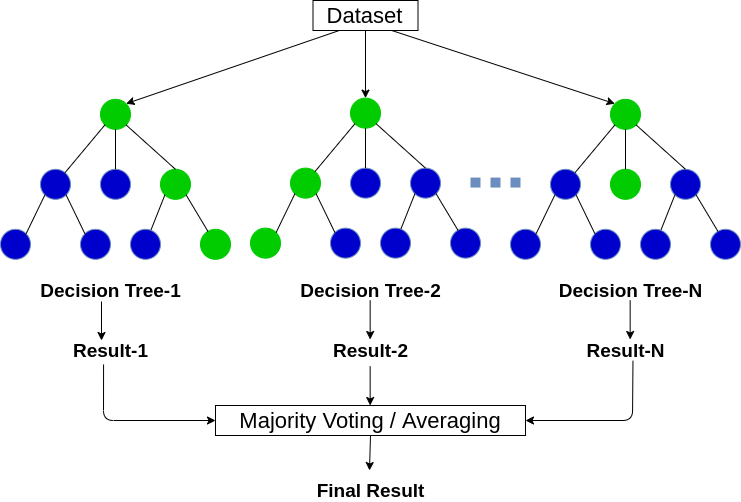
The decision tree is easy to comprehend since it replicates the phases that a person goes through while making a real-life decision. It may be quite useful in dealing with decision-making issues. It is a good idea to consider all potential solutions to an issue. Cleaning data is not as important as it is with other methods.



**Fig 8: - Block Diagram of Decision Tree Classifier**

**3.1 (e) Random Forest Regressor**

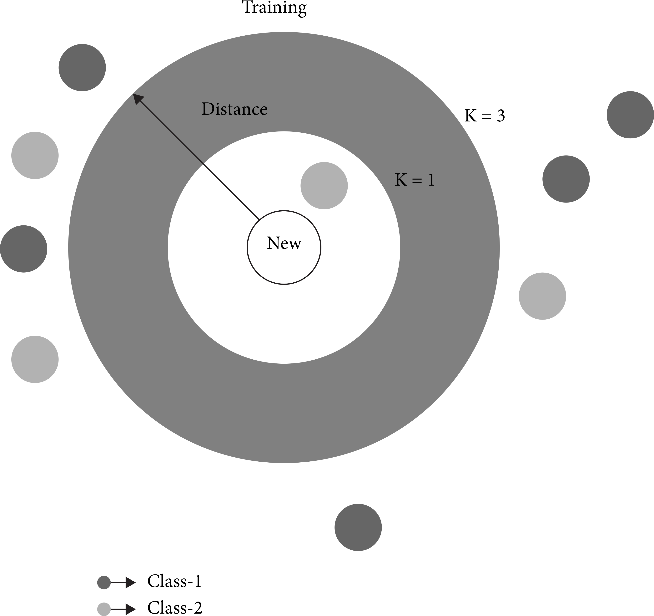
Random Forest Regressor builds numerous choice trees to go about as an outfit of grouping and relapse process. Various choice trees are built utilizing an irregular subset of the preparation informational indexes. A huge assortment of choice trees gives higher exactness of results. The runtime of the calculation is relatively quick and furthermore obliges missing information. Arbitrary woodland randomizes the calculation and not the preparation informational index. The choice class is the method of classes created by choice trees. Fig 9 shows Random Forest.



**Fig 9: - Random Forest Regressor**

* 1. **(f) K-Neighbor Classifier (KNN)**

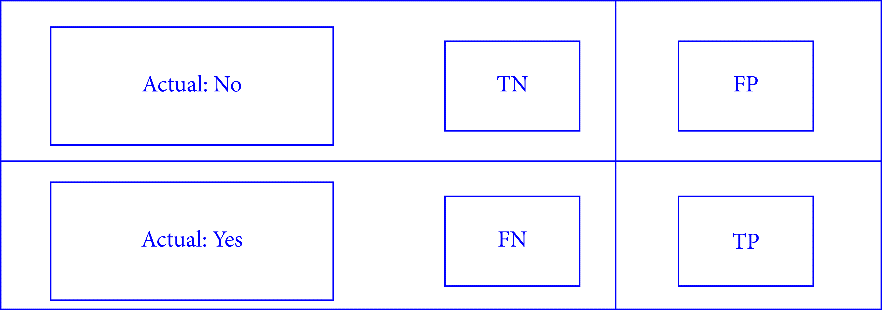
Figure [10](https://www.hindawi.com/journals/cmmm/2021/6141470/fig7/) depicts the whole KNN model’s flowchart. One of the simplest ML algorithms is KNN, which uses the supervised learning approach. A new case is assigned to a category based on how closely it resembles prior categories. This is known as the KNN technique. With the KNN method, you can store all the data you have and then classify new data based on how similar it is to the old. This suggests that the KNN technique can rapidly classify new data into well-defined categories. Though it is often utilized for classification problems, the KNN method may be used for regression as well. There are no data assumptions made by the KNN technique, which is nonparametric and also called a “lazy learner algorithm,” since it does not instantly learn from the training set but rather keeps and categorizes the data for later. If it receives new data, the KNN classifies it into a category that is quite close to the new data that was stored during training.



**Fig 10: - KNN Classifier Working Procedure**

* 1. **(g) Confusion Matrix**

Figure 11 shows the confusion matrix. The confusion matrix rates machine learning classification models’ performance. All models were evaluated using the confusion matrix. The confusion matrix illustrates how often our models guess correctly and incorrectly. Poorly predicted values received false positives and negatives, whereas properly predicted values received genuine positives and negatives. The model’s accuracy, precision-recall trade-off, and AUC were assessed after grouping all predicted values in the matrix.

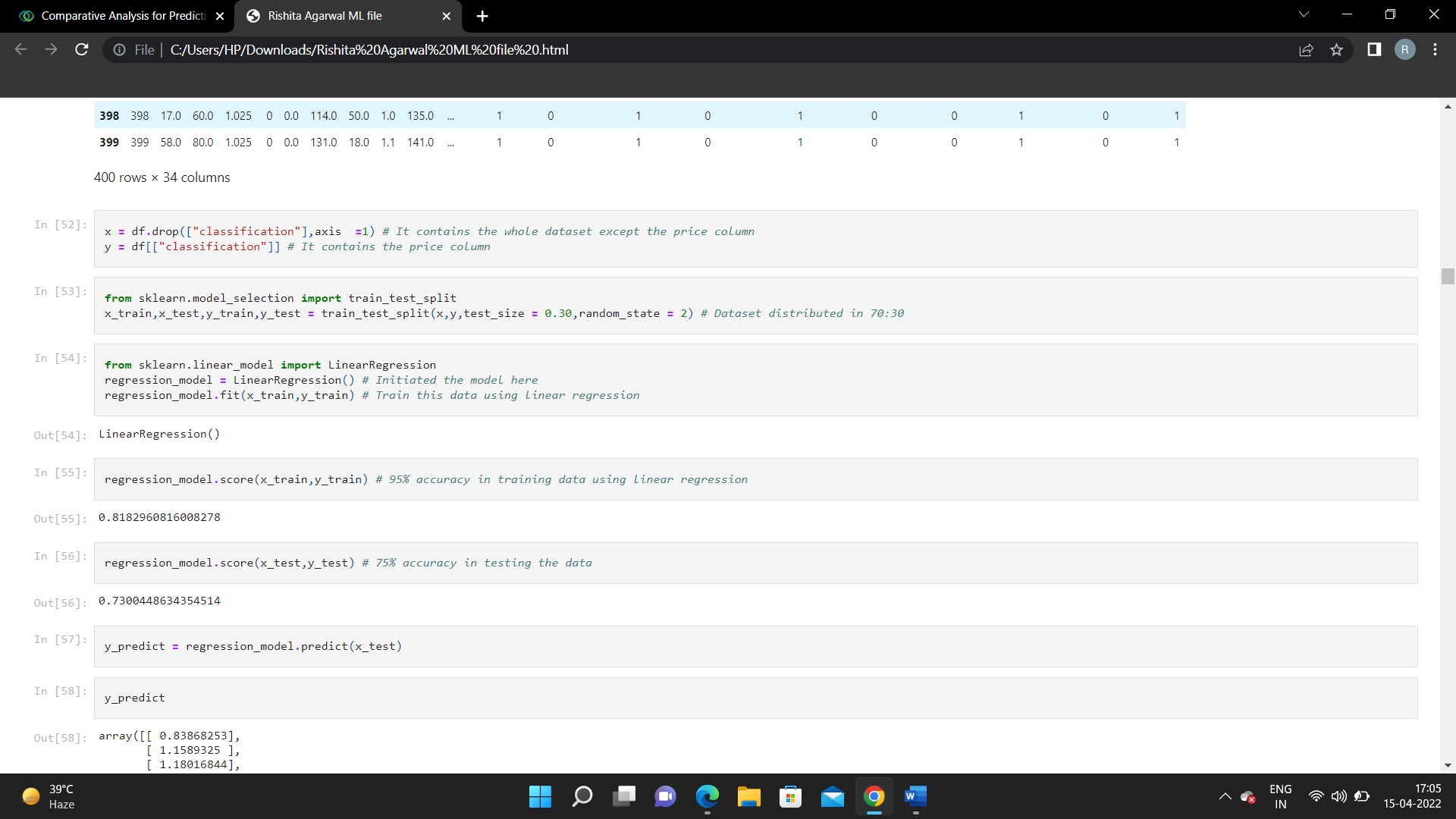


**Fig 11: - Confusion Matrix Block Diagram**

**4. Result and Data Analysis**

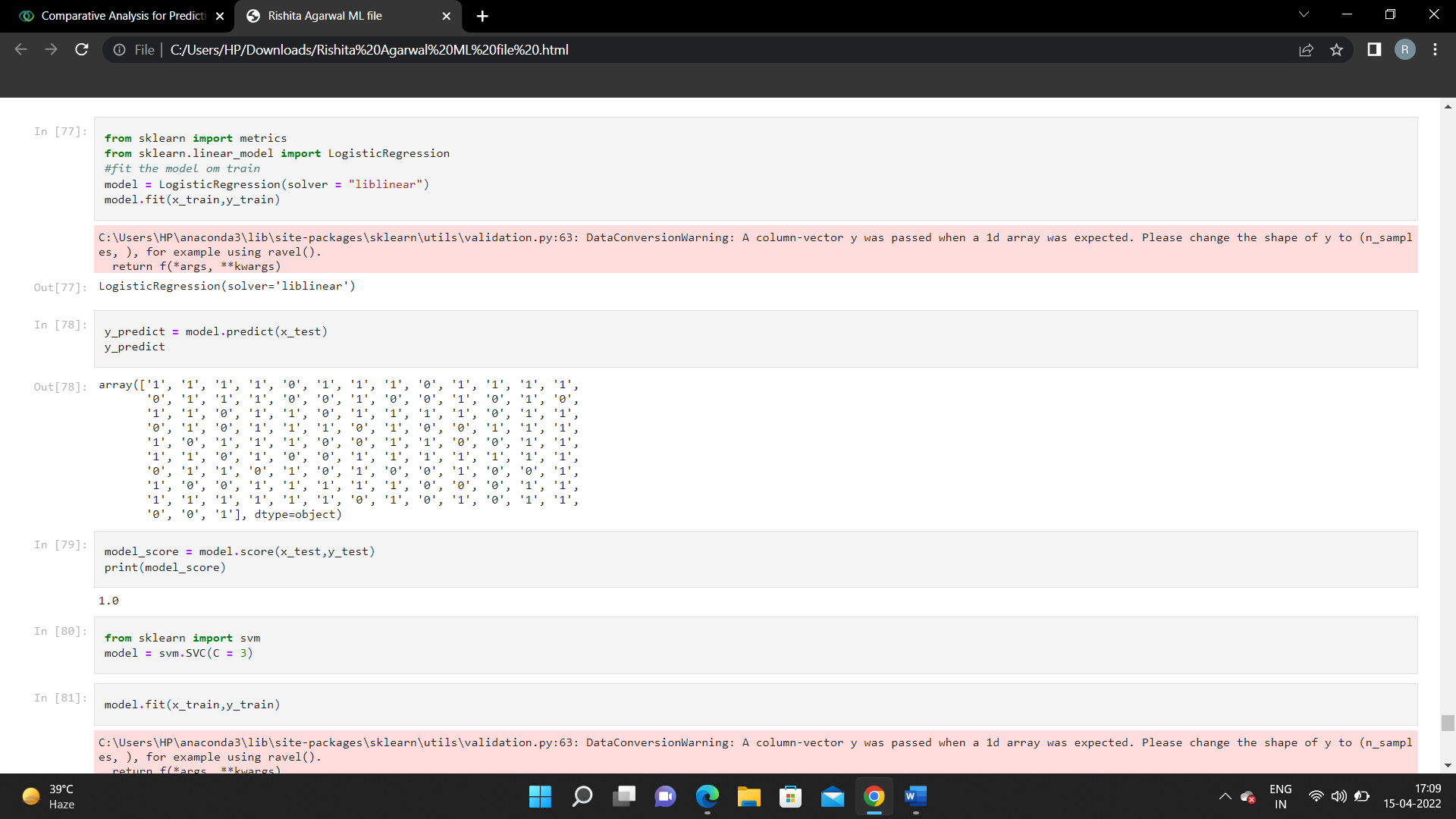
**4.1 Linear Regression**

Figure 12 shows the Linear Regression accuracy. In this case, the accuracy is 73% percent.

**Fig 12: - Accuracy of Linear Regression**

**4.2 Logistic Regression**

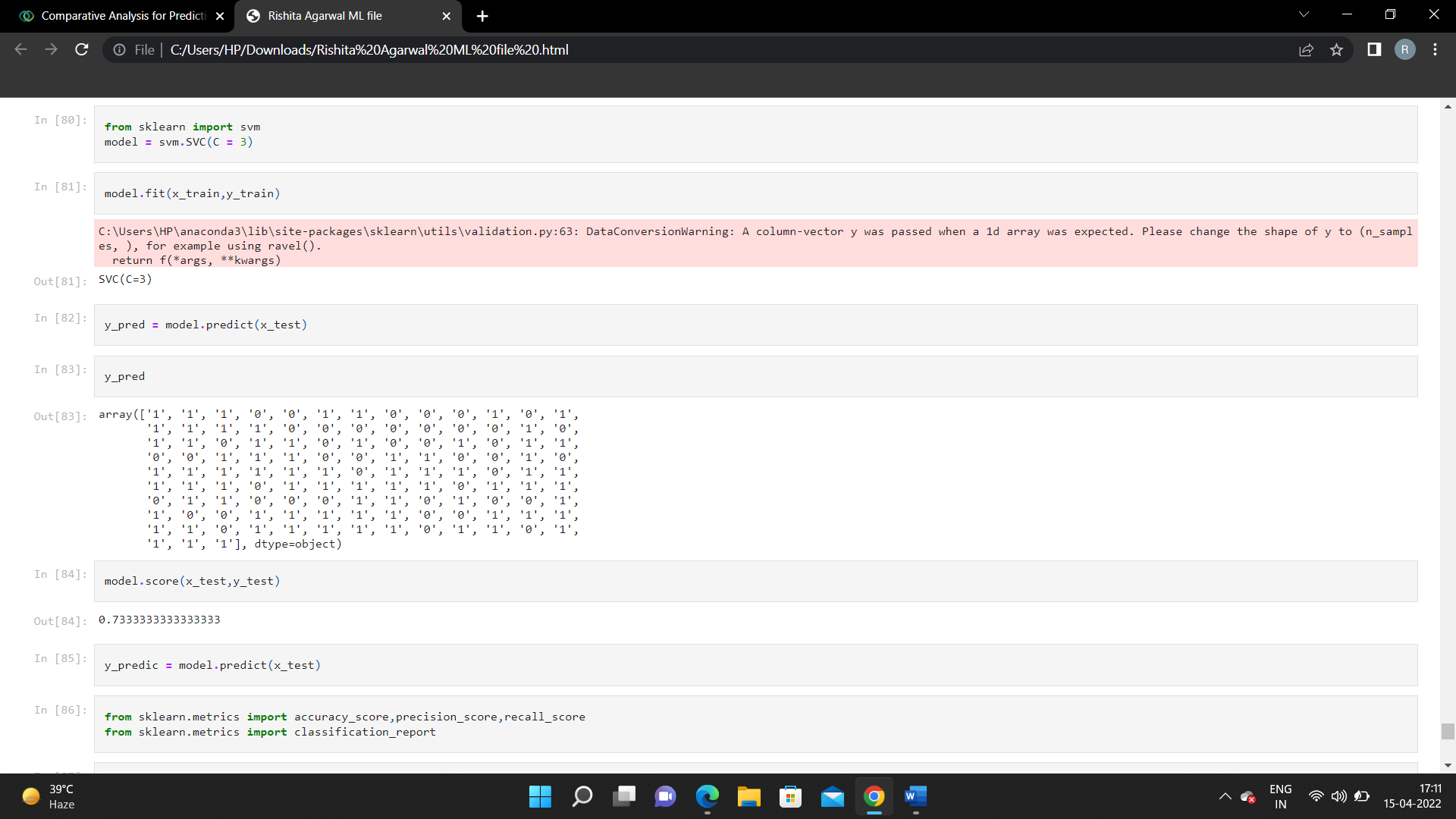
Figure 13 shows the Logistic Regression model score. In this case, the accuracy is 73% percent.



**Fig 13: - Model Score of LR**

**4.3 Support Vector Machine (SVM)**

Fig 14 shows the accuracy of the model SVM. In this case, the accuracy is 73% percent.

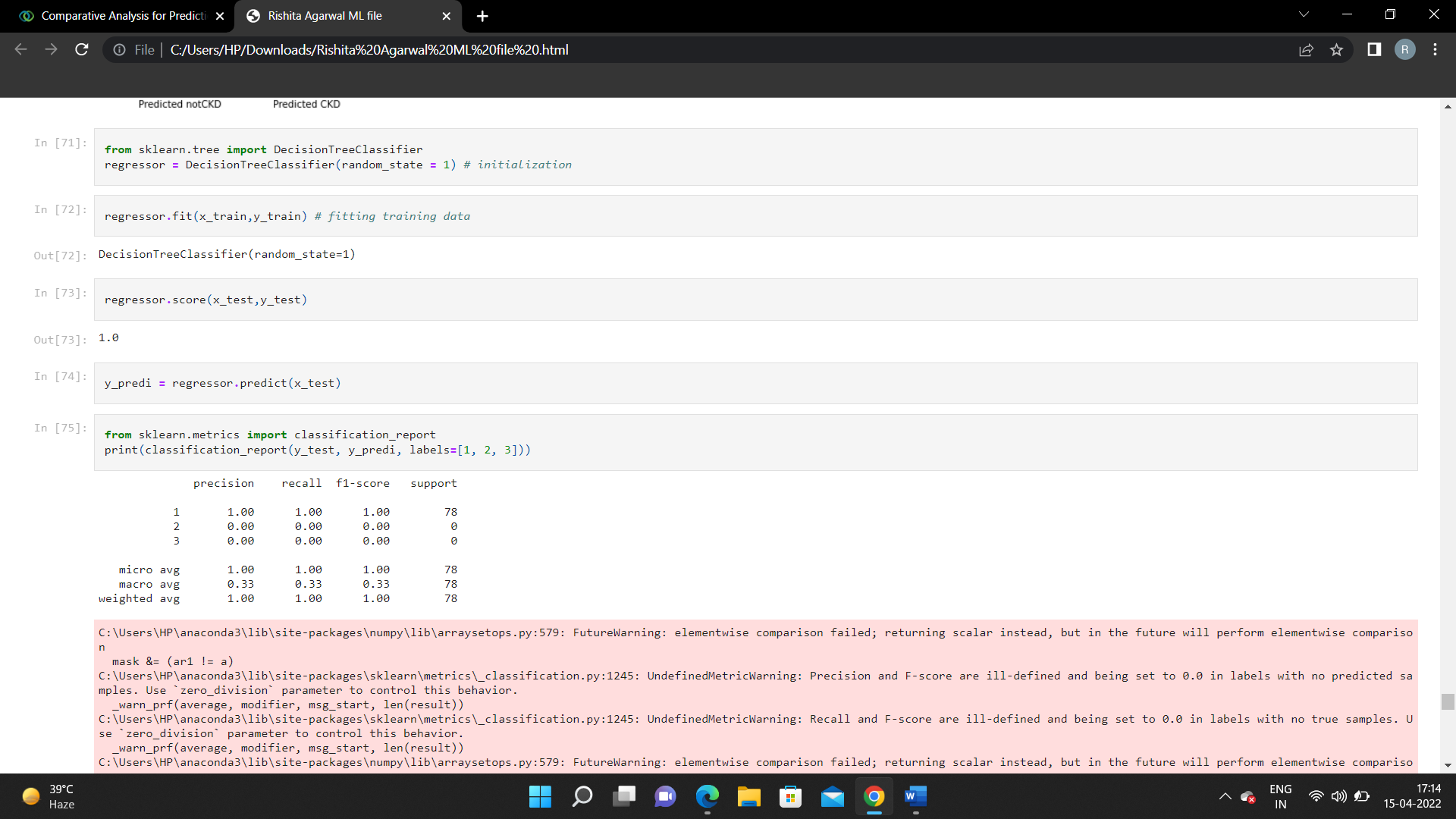


**Fig 14: - Accuracy of SVM**

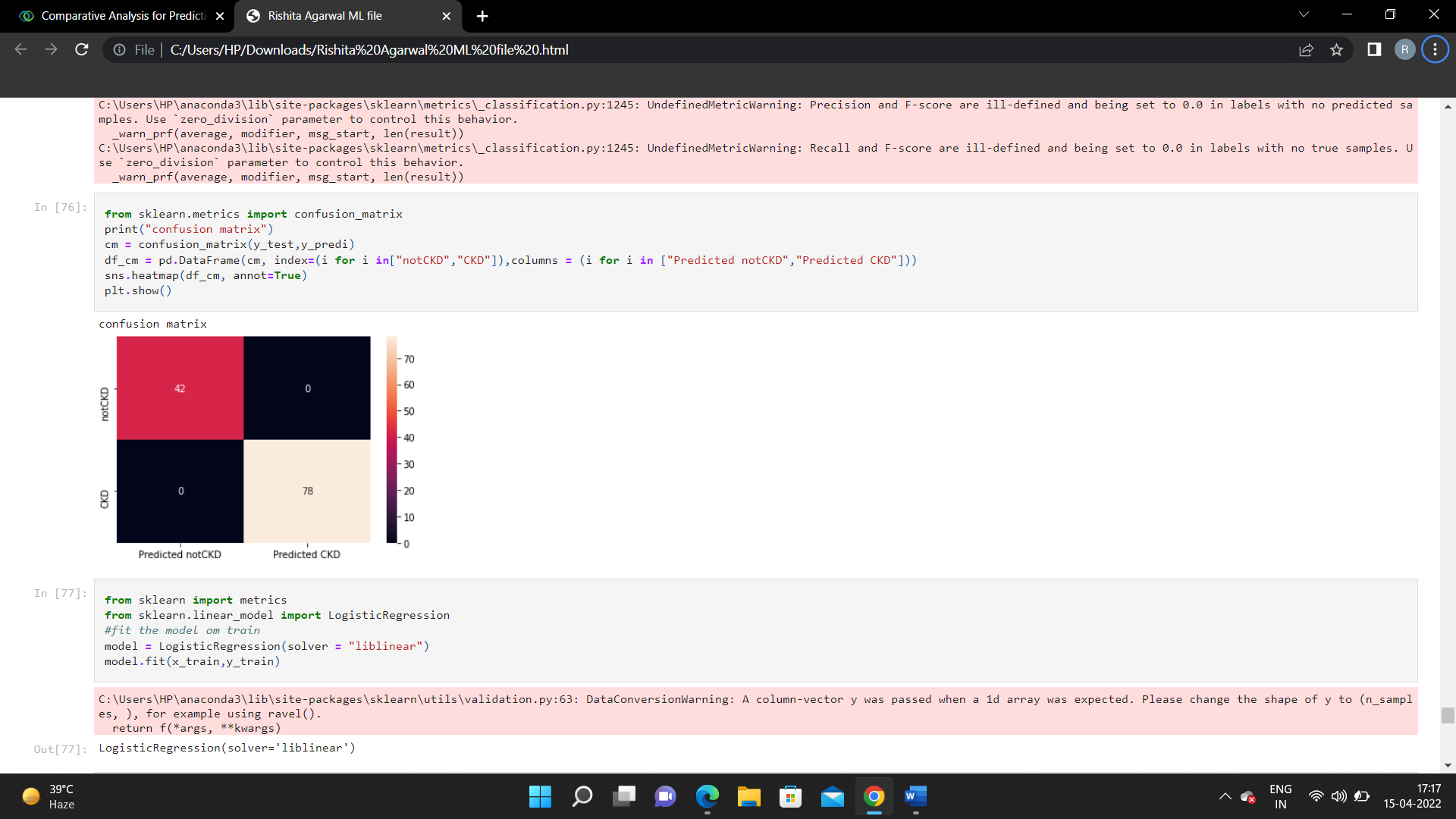
**4.4 Decision Tree Classifier**

Fig 15 shows the classification report of the model using Decision Tree Classifier Algorithm. In this case, the accuracy is 100 percent.

Fig 16 shows the Confusion Matrix of DT.



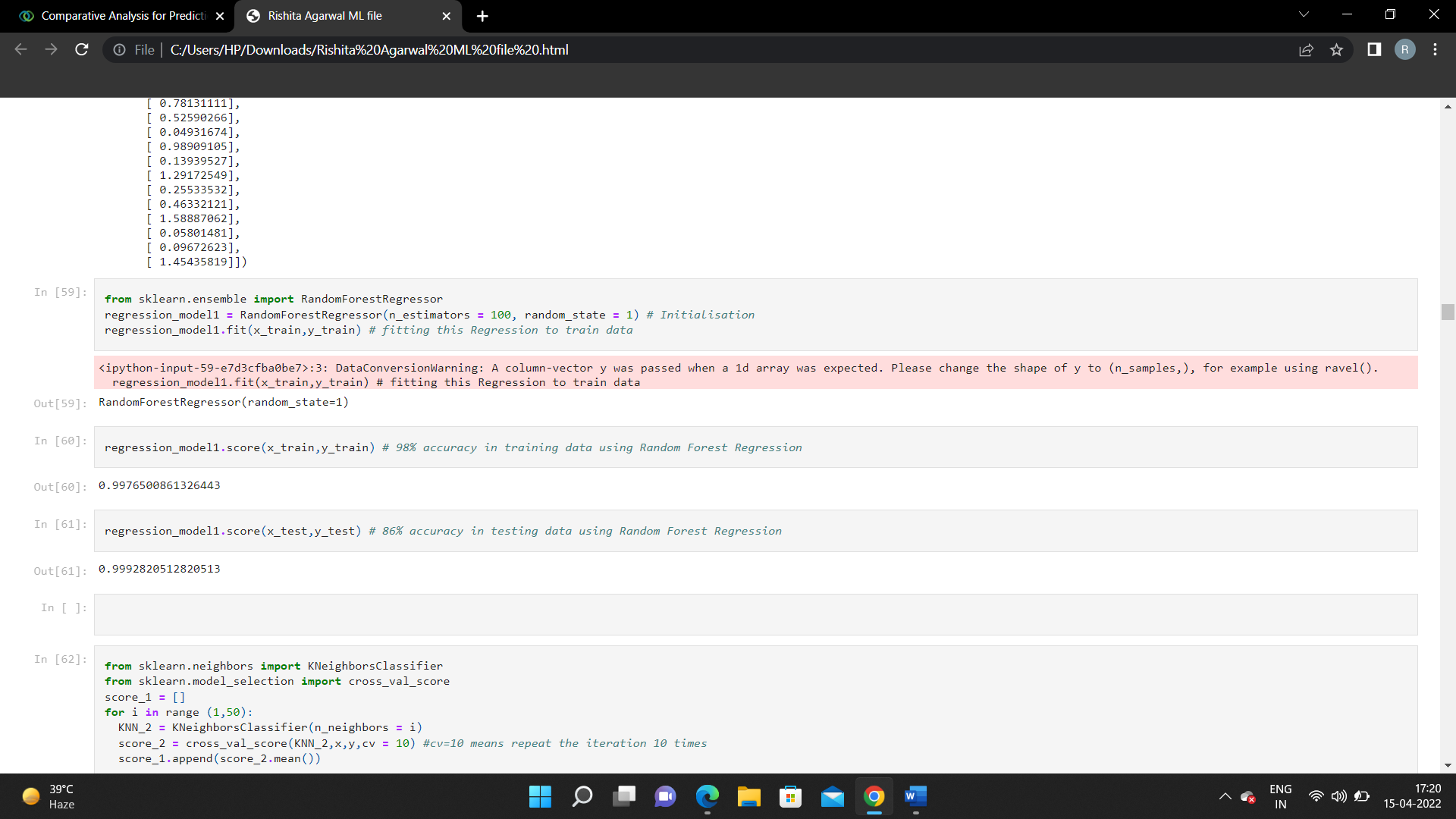
**Fig 15: - Classification Report of DT**



**Fig 16: - Confusion Matrix of DT**

**4.5 Random Forest Regressor**

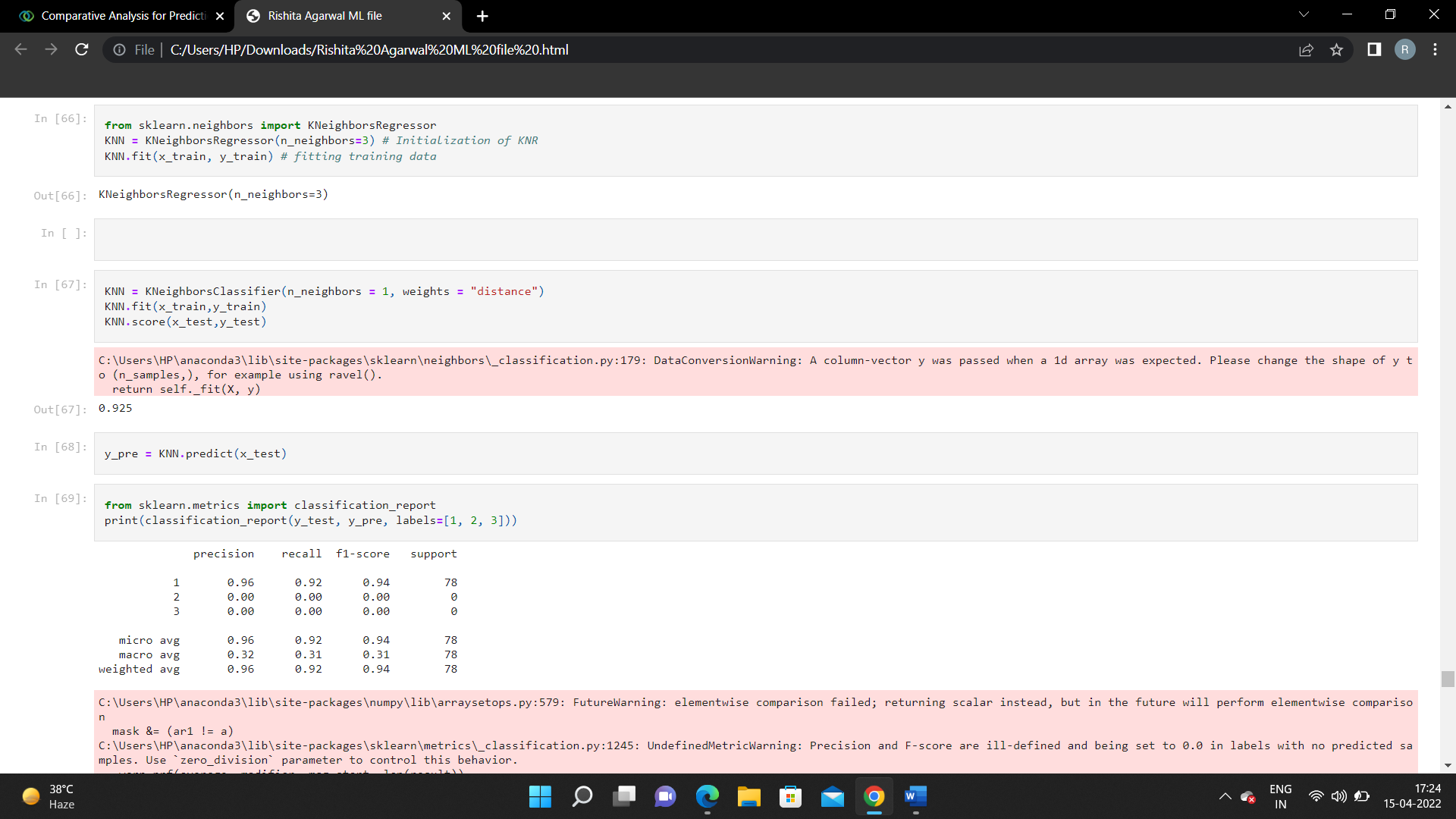
Accuracy of the model using Random Forest Regressor is 99 percent. This is depicted in fig 17.



**Fig 17: - Accuracy of Random Forest Regressor**

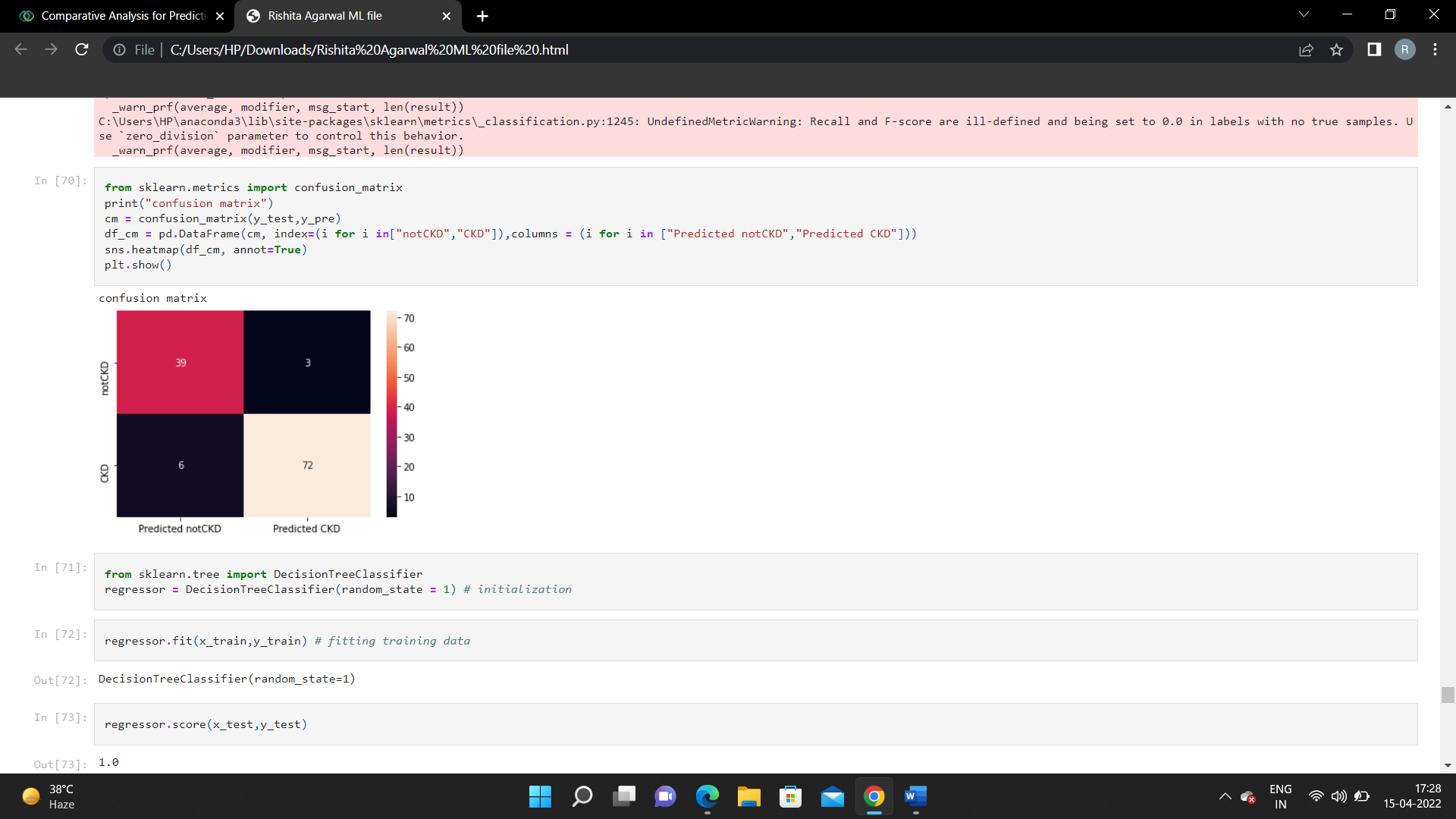
**4.6 K-Neighbor Classifier (KNN)**

The accuracy of the model using KNN is 92 percent. The classification report of KNN is shown in fig 18.



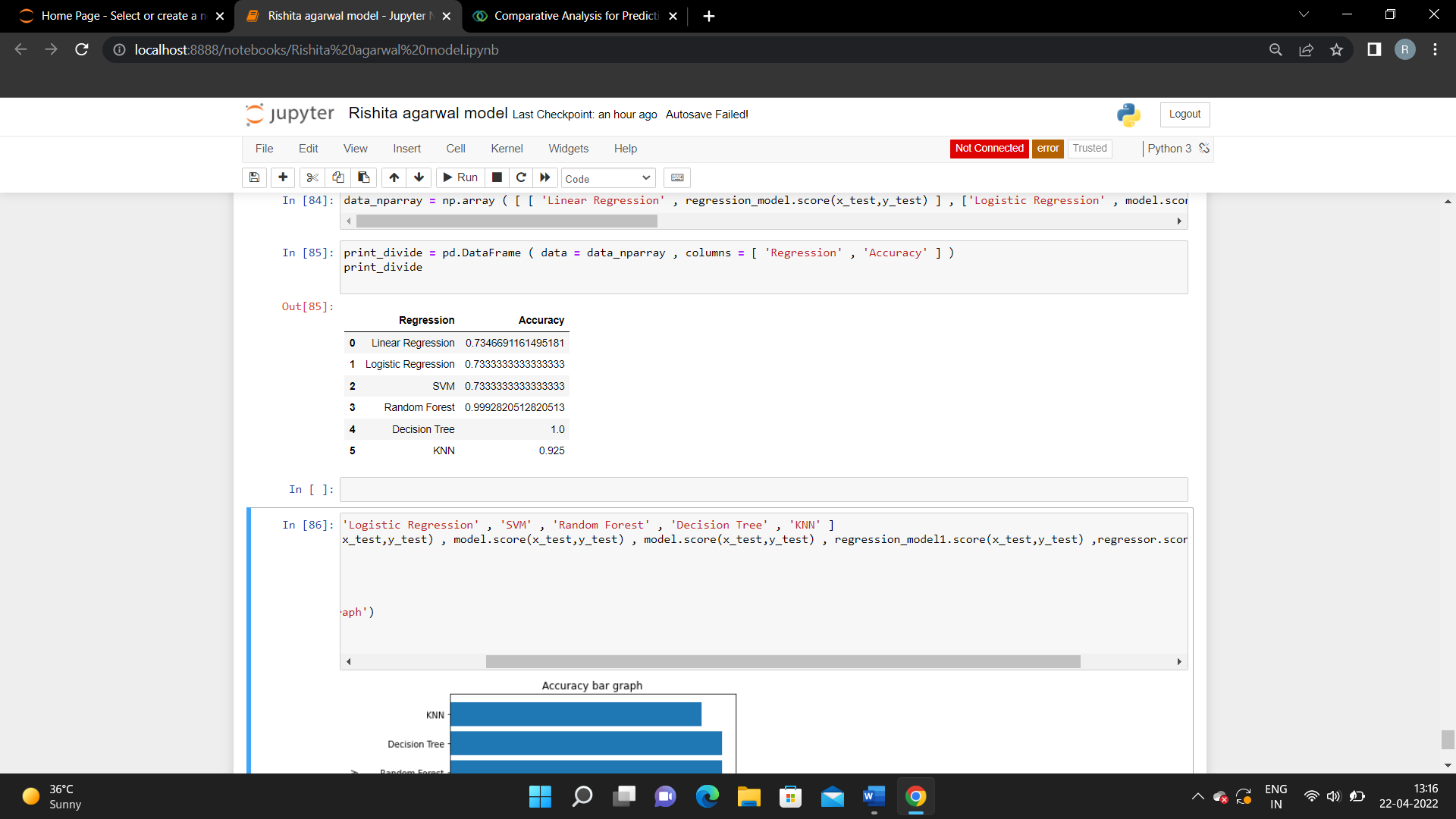
**Fig 18: - Accuracy and Classification report of KNN**

Confusion Matrix of KNN is depicted in figure 19.



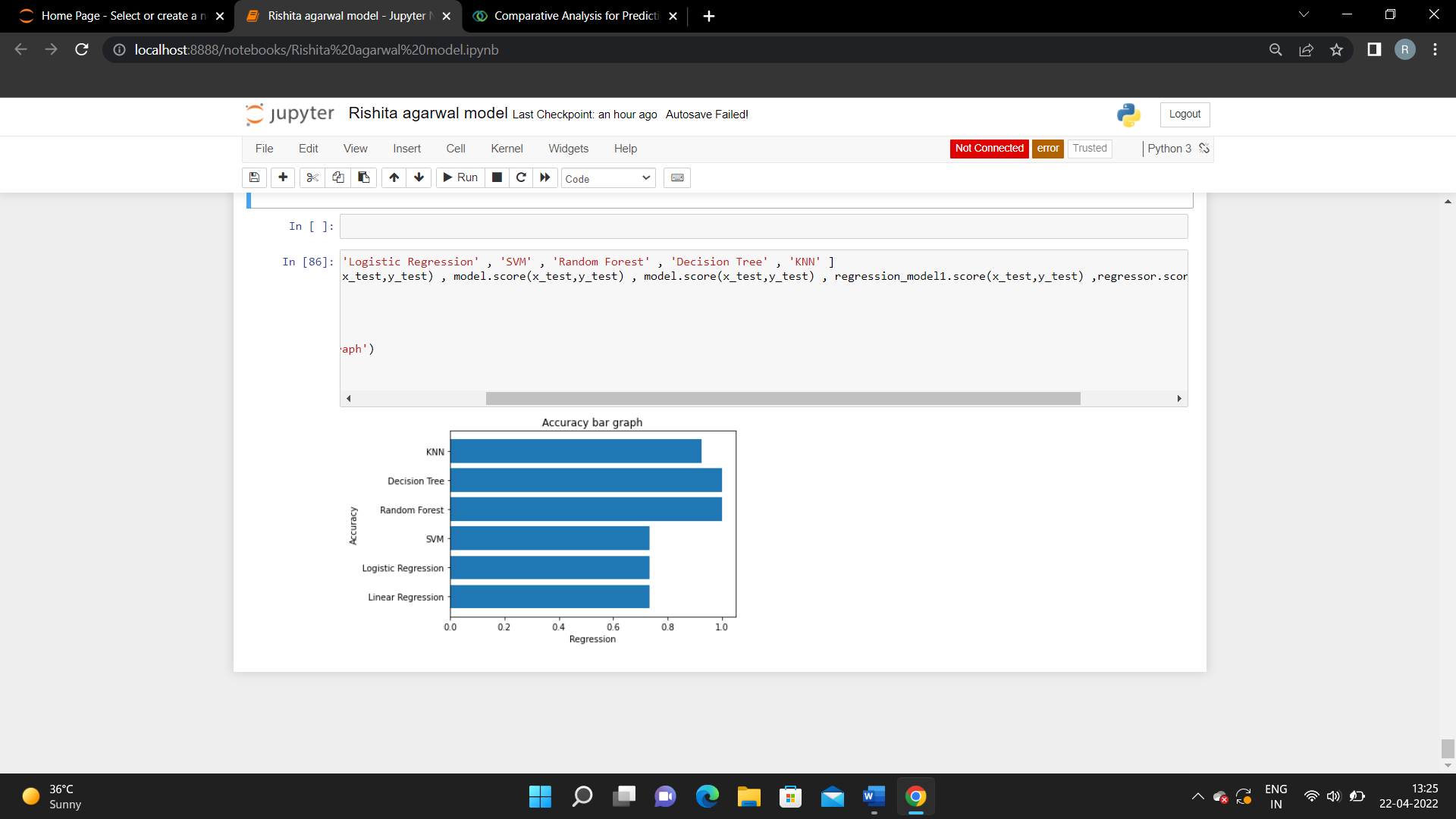
**Fig 19: - Confusion Matrix of KNN**

**4.7 Model Conclusion**



**Fig 20: - Accuracy of different models**

The chart clearly indicates that Decision Tree is the best among the many models in the framework. Using Decision Tree, this paper achieved 100 percent accuracy. The Random Forest Regressor also achieved good accuracy i.e., 99 percent. But the same paper is achieving the poor quality by using SVM i.e., 73 percent. Fig.20 and Fig.21 shows the accuracy of different models respectively.



**Fig 21: - Bar-graph showing the accuracy of different models**

**5. Conclusion**

According to the findings of the study, the Random Forest Regressor approach and Decision Tree can be used to predict chronic kidney disease more accurately. According to the study, their precision was 99 percent, and their accuracy was 100 percent. Compared to prior research, the accuracy percent of the models used in this investigation is considerably higher, indicating that the models used in this study are more reliable than those used in previous studies. When cross validation measurements are used in the prediction of chronic kidney disease, the Random Forest Regressor method outperforms the other processes. Future research may build on this work by developing a web application that incorporates these algorithms and using a bigger dataset than the one utilized in this study. This will aid in the achievement of improved outcomes as well as the accuracy and efficiency with which healthcare practitioners can anticipate kidney issues. This will enhance the dependability of the framework as well as the framework’s presentation. The hope is that it would encourage people to seek early treatment for chronic renal disease and to make improvements in their lives.

**References**

1. A. S. Levey, R. Atkins, J. Coresh et al., “Chronic kidney disease as a global public health problem: approaches and initiatives–a position statement from kidney disease improving global outcomes,” *Kidney International*, vol. 72, no. 3, pp. 247–259, 2007.View at: [Publisher Site](https://doi.org/10.1038/sj.ki.5002343) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Chronic%20kidney%20disease%20as%20a%20global%20public%20health%20problem:%20approaches%20and%20initiatives%E2%80%93a%20position%20statement%20from%20kidney%20disease%20improving%20global%20outcomes&author=A.%20S.%20Levey&author=R.%20Atkins&author=J.%20Coresh%20et%20al.&publication_year=2007)
2. V. Jha, G. Garcia-Garcia, K. Iseki et al., “Chronic kidney disease: global dimension and perspectives,” *The Lancet*, vol. 382, no. 9888, pp. 260–272, 2013.View at: [Publisher Site](https://doi.org/10.1016/S0140-6736(13)60687-X) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Chronic%20kidney%20disease:%20global%20dimension%20and%20perspectives&author=V.%20Jha&author=G.%20Garcia-Garcia&author=K.%20Iseki%20et%20al.&publication_year=2013)
3. N. R. Hill, S. T. Fatoba, J. L. Oke et al., “Global prevalence of chronic kidney disease – a systematic review and meta-analysis,” *PLoS One*, vol. 11, no. 7, article e0158765, 2016.View at: [Publisher Site](https://doi.org/10.1371/journal.pone.0158765) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Global%20prevalence%20of%20chronic%20kidney%20disease%20%E2%80%93%20a%20systematic%20review%20and%20meta-analysis&author=N.%20R.%20Hill&author=S.%20T.%20Fatoba&author=J.%20L.%20Oke%20et%20al.&publication_year=2016)
4. H. Nasri, “World kidney day 2014; chronic kidney disease and aging: a global health alert,” *Iranian Journal of Public Health*, vol. 43, no. 1, pp. 126-127, 2014.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=World%20kidney%20day%202014;%20chronic%20kidney%20disease%20and%20aging:%20a%20global%20health%20alert&author=H.%20Nasri&publication_year=2014)
5. G. Abraham, S. Varughese, T. Thandavan et al., “Chronic kidney disease hotspots in developing countries in South Asia,” *Clinical Kidney Journal*, vol. 9, no. 1, pp. 135–141, 2016.View at: [Publisher Site](https://doi.org/10.1093/ckj/sfv109) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Chronic%20kidney%20disease%20hotspots%20in%20developing%20countries%20in%20South%20Asia&author=G.%20Abraham&author=S.%20Varughese&author=T.%20Thandavan%20et%20al.&publication_year=2016)
6. K. T. Mills, T. Xu, W. Zhang et al., “A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010,” *Kidney International*, vol. 88, no. 5, pp. 950–957, 2015.View at: [Publisher Site](https://doi.org/10.1038/ki.2015.230) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20systematic%20analysis%20of%20worldwide%20population-based%20data%20on%20the%20global%20burden%20of%20chronic%20kidney%20disease%20in%202010&author=K.%20T.%20Mills&author=T.%20Xu&author=W.%20Zhang%20et%20al.&publication_year=2015)
7. B. Ene-Iordache, N. Perico, B. Bikbov et al., “Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study,” *The Lancet Global Health*, vol. 4, no. 5, pp. e307–e319, 2016.View at: [Publisher Site](https://doi.org/10.1016/S2214-109X(16)00071-1) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Chronic%20kidney%20disease%20and%20cardiovascular%20risk%20in%20six%20regions%20of%20the%20world%20(ISN-KDDC):%20a%20cross-sectional%20study&author=B.%20Ene-Iordache&author=N.%20Perico&author=B.%20Bikbov%20et%20al.&publication_year=2016)
8. S. Anand, M. A. Khanam, J. Saquib et al., “High prevalence of chronic kidney disease in a community survey of urban Bangladeshis: a cross-sectional study,” *Glob Health*, vol. 10, no. 1, p. 9, 2014.View at: [Publisher Site](https://doi.org/10.1186/1744-8603-10-9) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=High%20prevalence%20of%20chronic%20kidney%20disease%20in%20a%20community%20survey%20of%20urban%20Bangladeshis:%20a%20cross-sectional%20study&author=S.%20Anand&author=M.%20A.%20Khanam&author=J.%20Saquib%20et%20al.&publication_year=2014)
9. L. Ali, K. Fatema, Z. Abedin et al., “Screening for chronic kidney diseases among an adult population,” *Saudi Journal of Kidney Diseases and Transplantation*, vol. 24, no. 3, p. 534, 2013.View at: [Publisher Site](https://doi.org/10.4103/1319-2442.111049) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Screening%20for%20chronic%20kidney%20diseases%20among%20an%20adult%20population&author=L.%20Ali&author=K.%20Fatema&author=Z.%20Abedin%20et%20al.&publication_year=2013).