IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 12, December 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

Detection of Faulty Wafer by Using Machine Learning Model

Anshu Kumari[1], Subhanshu Babbar[2], Keshav Mishra[3] , Aakarshit Pandey[4] , Sharanya Chandran[5]

U.G Student, Department of Computer Engineering, HMRITM and Management, Bakoli, Delhi, India[1]

U.G Student, Department of Computer Engineering, HMRITM and Management, Bakoli, Delhi, India [2]

U.G Student, Department of Computer Engineering, HMRITM and Management, Bakoli, Delhi, India [3]

U.G Student, Department of Computer Engineering, HMRITM and Management, Bakoli, Delhi, India [4]

Professor, Department of Computer Engineering, HMRITM and Management, Bakoli, Delhi, India [5]

ABSTRACT: Increased, fast and tremendous growth in the semiconductor industry results in high density and IC performance in each unit area. Semiconductors are an integral part of electronicdevices that allow advances in communication, health care, computers, militarysystems, transportation, and many other systems. In most cases, however, impairmentoccurs in day-to-day production and affects the daily production of IC packages at theend of the line. Therefore, ultimately challenging the limitations of semiconductortechnology. Profile analysis and process of monitoring are critical in detecting a widerange of unconventional phenomena in the production of semiconductors, which have highly complex , coherent, and long-term production processes for improving the yield andcontrol over quality. This study is aimed at improving the error detection in semiconductors andsegregation framework for monitoring and analyzing the profile dataset performance that occurs due to a large number of variable processes associated with eliminating the reason for errors and thusminimizing irregular yield losses. We, therefore, created a model that differentiateswhether WAFER is in working condition or not. If the client has a production line andwhen various wafers are installed or deployed and when there is an error in any waferhe should manually check all the wafers and check which one is faulty and the entire production line needs to stop and replace the wrong one. So, we created a model thatdifferentiates whether a wafer works or not. The first step is to provide input data ofvarious sensors for different wafers available. The aim is then to create a machine learningmodel that can predict whether the wafers are faulty and whether it needs to be replaced or not based on inputs received fromvarious sensors. But if it is found early then that part of the production line will beclosed. So the whole production line does not disrupt. There are two categories:+1 and -1: -1 represents that the wafer is working properly in good condition and is not needed to bereplaced and +1 represents that the wafer is faulty and needs to be replaced.

KEYWORDS: Semiconductors, WAFERS, Machinelearning, Classification, Accuracy, Prediction

I. INTRODUCTION

The study of this paper will help you get knowledge of our project and alsounderstand the shortcoming in the way semiconductors are produced in manufacturing units and the importanceof "WAFERS" in the process and how it affects the process and what the problem thatwe faced and have dealt with here.

In electronics, a wafer is known as a small piece of semiconductor, such as crystalline silicon(c-Si) which is used in the manufacturing of composite circuits , in photovoltaics to producesolar cells. This wafer acts as a substrate to microelectronic devices which are built inside and above the piece. It takes on many micro-fabrication processes like doping, ionimplantation, etching, thin-film deposition and photolithographic patterning. Finally, individual microcircuits are separated from wafers and assembled asan integrated circuit. It plays an important role in production and needs to be maintained properly with carein the production line. The purpose of the semiconductor industry is to obtain highyields during the device area of all the wafers. In general, to improve the yields, similar error patterns worked as a guide in providing marketable responses for engineersto recognize errors or origin of errors. A loss of yield can be caused by a number of problems such asmechanical failure (flooding process) or human error. Error presence in thewafer can be understood as an indication of these problems occurring, the detection of these problems accurately and in time becomes an important task. Production error bydisplaying the physical and electrical properties

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

of the piece. When the back processing is complete, wafers undergo various electrical tests called"Parametric Tests"; and in these tests various electrical parameters of key devices are measured that shape the basic blocks of all integrated circuits: resistors, capacitors,diodes, transistors, inductors, etc[1].

We have identified a problem in the production line of a semiconductor factory wherethe performance of the wafer is completely competitive in the industry and when thewafer is faulty the entire production line stops replacing or repairing the wafer so wehave tried to come up with a solution to this problem. We decided to make a machinelearning model that effectively separates the wrong wafer from the good ones so thatwe do not have to stop the whole production line and focus on the part where thefaulty wafer is and only that part of the line can be stopped and fixed without touchingthe entire production line.

The paper that gave us the idea had a theoretical proposition of the solution using an image classification in which it takes longer to train the model and even to fetch theresults. So we decided to do it with the help of some data that has been collected fromvarious sensors that are used to detect various types of faults in the “WAFERS” whichhas fastened up the process of training the module and even got the results faster.

II. METHODOLOGIES

Before starting out with our project we read and tried to understand a few research papers that mentioned the theoretical use of the methods and image classification as ameans to classify and identify the defective wafers[1] so here we tried and started tocome up with a new addition and practical implementation of these and possible optimizations in thealgorithms used to achieve the desired results. We started our work by thoroughlyunderstanding the base paper we choose and the method(image classification)[9] they used to achieve the results. So we came up with a new approach and tried searching for possible solutions and we found one where we had to use the data extracted fromvarious sensors as the input for our machine learning model. We got the raw data fromKaggle and started working on the data by arranging it, cleaning it, and performingpreprocessing techniques on the data to make it usable, understandable, and trainablefor the machine and to achieve the desired results. Here we applied K-meansclustering to cluster the data and optimized forms of Random Forest Classifier andXGboost algorithms to perform the intended tasks and get the desired results forwhich the data and process flow has been shown below in the form of a flow diagram.

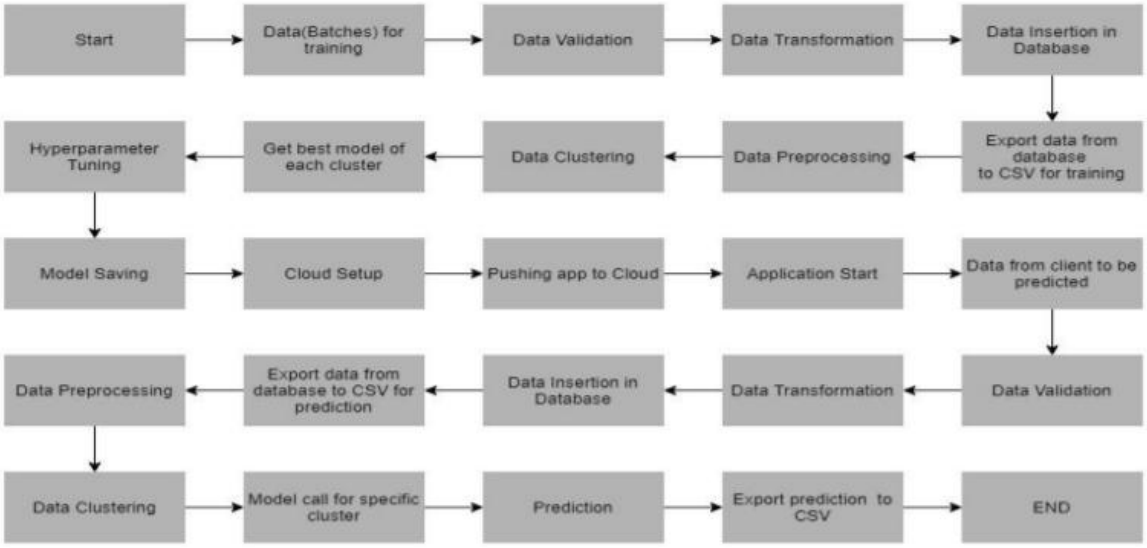

Figure 1. Classification Method

III. WORKFLOW

As initially we have a bigger problem that need to be broken into smaller problem due to many reasons like we focus on one thing at a time, as we know that while working on a big project it will be better if different people will work on different section according to their expertise so that they can easily handle the problem that may occur in a particular

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

section. So we also move in the same way. Firstly, we Collect the datasets from various platforms like Kaggle, MNIST, etc. When this modelis handed to any organization, batches of multiple sets of files will be uploaded by the client to the particular sharable location and share the path of the location with us so that we can perform processing, training, building the model, prediction, and deployment ofthe data. When the data is obtained, it is the raw form and our machine learningalgorithm can not work directly upon it, so it needs some preprocessing. In this, the first step is data validation and data transformation in which a number of processes are **involved;** we check whether there is any missing value present or not, if there is a missing **value,** then we fill all those missing values with NULL, and also remove the irrelevant columns that do not help in prediction, etc[13].

After performing all theseoperations, all the batch files will be aggregated and data is stored in the database.

After storing the data into the database, the necessary step is performed, which is database creation and connection that will bedone like creating a database with the given name passed. And there may be chances that the database is already created, and If the database is alreadycreated, then in that scenario we will open the connection to the database. Then one of the important steps which is a table creation in the database is done. Then finally, there will be insertion of files in the table.

Description of the data, data validation, and data insertion in the database:

Data Description:

Initially, batches of multiple sets of files will be uploaded by the client to the particular sharable location, from where we can take the data and do further processing . The data shared by them will contain Wafer names (wafer number) and 590 columns that show the reading of wafers from different sensors. The last column is the target column that will have the "Good/Bad" value for all the wafers. The last target column which is actually the "Good/Bad" column will have two unique values +1 and -1. And these two unique values represent the specific thing as:

"+1" represents Bad wafer and "-1" represents Good Wafer.

Irrespective of the training data that we have, we additionally require a "schema" file which is obtained from the client side ,which contains the necessary details regarding the training files such as: files name,Date values length in Files Name, Time values length in Files Name, Number ofColumns and Columns name and the data types of the column.

Validation of Data:

At this point, we try to carry out different validation sets on the given collection of training files with the help of schema files that we obtain from the client.

1. Name Validation- Name validation is the first priority from where we start to validate. The names of the files are checked with the provided schema pattern. The pattern that we follow here is character ,datestamp, and timestamp and also considering the length of the date and timestamp.We then check the length of date and time duration in the file name. If a value does not meet the requirements mentioned, then we transfer it to "Bad_Data_Folder" elseto "Good_Data_Folder."
2. Column Numbers - Now, we move toward checking the number of columns in the files that are present, and if they match the value given in the schema, then the folder is transferred to "Good_Data_Folder" elseto "Bad_Data_Folder."
3. Name of Columns - As all requirements are checked based on the description that are mentioned in the schema files. So, we check whether the column name should be the same as provided in schema files. If not, then the file is transferred to "Bad_Data_Folder" elseto "Good_Data_Folder."
4. Datatype is one of the most important aspects in any dataset. So, we now look into the columns data type which is indicated by the schema file. If the datatype is not correct, then the file is loaded and its contents are transferred to "bad_Data_Folder" elsewe move such files to "Good_Data_Folder."
5. In cases where all the values in a column are missing or null, we simply discard those files and move it to the Bad_Data_Folder.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

Data Insertion in Database:-

All the batch files that we have will be aggregated and stored in the database.

1. Creating and Connecting the database - Firstly we generate a new database. And there may be chances that the database is already created and If the database is alreadycreated, then in that scenario we will open the connection to the database.
2. Table creation in the database - A Good_Data_Folder is a field that is generated in the database to store the files that are inserted in the "Good_Data" folder. If the table has already been created, then the new files should not be inserted and new files are merged with the already existing table for training to be done on new as well as previous training files.
3. Passing files to the table - Every existing file in the "Good_Data_Folder" is transferred and loaded in the above-named table. We now validate the data type, if the file has valid data type in all of the column, then only the files is loaded in the table else if a file has an invalid data type in any of the columns,then the file is not crammed in the table and transferred to "Bad_Data_Folder".

After then Model Training, In Model Training:-

1. Data Export from Db - Export data from Db in a CSV file to be used with model training.
2. Data Preprocessing

a) If the columns contain null values, then impute them using the KNN imputer with 3 parameters then depending on the parameter we decide which value we need to replace with the missing value[12].

b) If a column has zero standard deviation, then it is removed. It can only be zero if the mean and actual values are equal.

3. Clustering - KMeans is a programming algorithm that is used to create clusters in the pre-processed data. The algorithm is commonly used to find the number of clusters in a particular dataset. The complete perfect or optimum number of clusters are selected with the help of an elbow plot , and there is also another method for selecting the number of clusters and that one is mainly for thedynamic selection of the number of clusters, in that case we use the "KneeLocator"function. The main idea is to implement the different algorithms for different clusters so that we are able to get the best algorithm for each cluster. Since we are using the K Means model for training the preprocessed data andis saved for further use in prediction.

4. Model Selection - After creating a cluster, we need to find the best model for it. This step can be done by looking into the AUC score for the two algorithms that we are using "Random Forest" and "XGBoost". After applying these algorithms we check which particular algorithm will work best in each cluster that we have, and this can be done by looking into AUC(Area Under Curve) Score. Therefore, we calculate the AUC scores of both the models and select the best modelwith the best score for every cluster. Every model is then saved for all the clusters present to be used in predictions.

IV. PREDICTION DATA DESCRIPTION:

Initially the client will send data in batches of multiple files and they placed their file at a particular sharable location, from where we can take the data and do further processing . The data shared by them will contain Wafer names (wafer number) and 590 columns of different sensor reading for each wafer.Irrespective of the training files, we also need a "schema" file from the client side ,that will contains all the necessary information regarding the training files such as: File Names, Date values length in FileName, Time values length in FileName, Number of Columns, Name of the Columns and their datatype. Then the prediction takes place in the deployed platform to retrieve the results from the data provided by the client using the cloud(HEROKU)[11] environment. And in return we will provide the output with the list of all the sensors and tell whether they are faulty or not.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

Image of Prediction along with deployment:

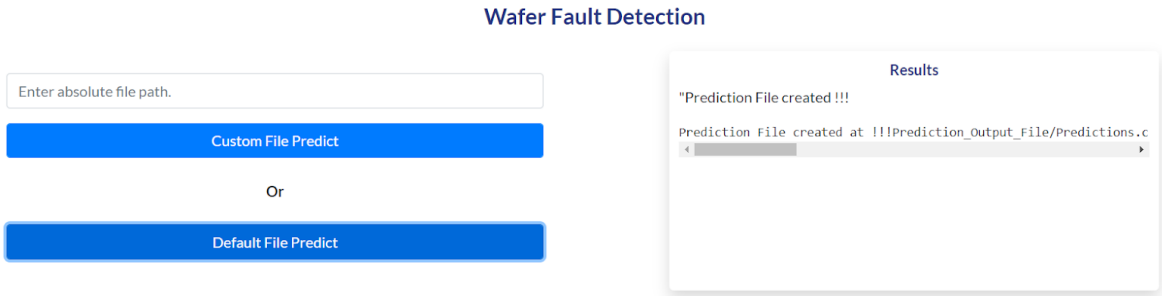

Figure 2: Final Prediction

Some Images of Prediction.csv file which is generated after prediction is done:

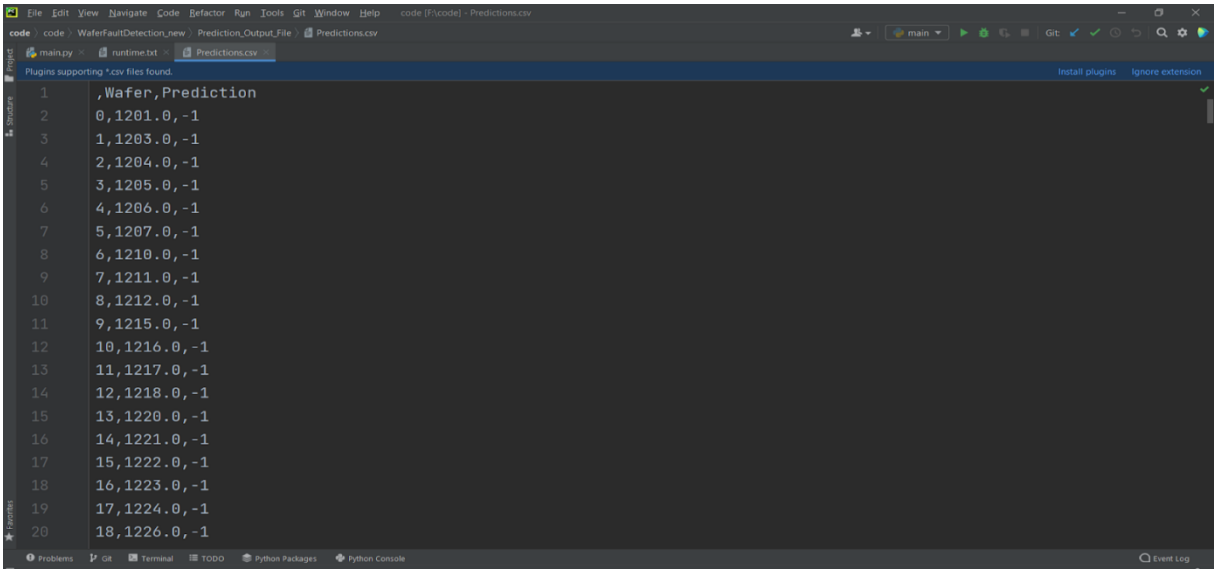

Figure 3: Prediction.csv

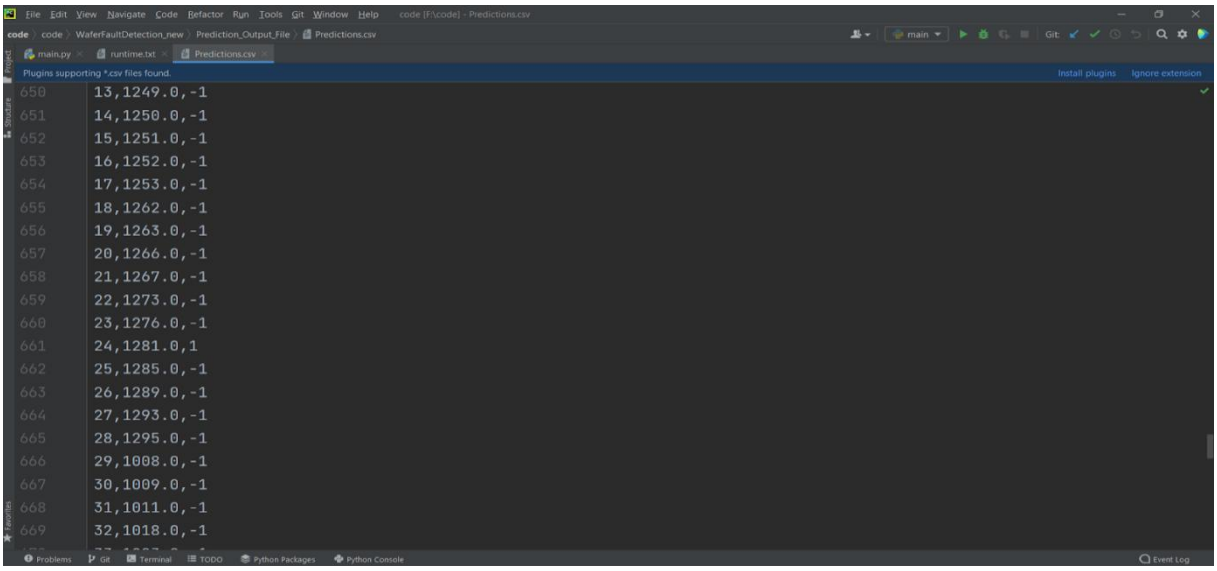

Figure 4: Prediction.csv

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

V. DESCRIPTION OF USED ALGORITHMS

1. K-Means Clustering Algorithm

K-Means Clustering falls under Unsupervised Learning Algorithms. Here, clustering means the process of dividing the datasets into groups, consisting of similar data-points.Clustering is generally of three types exclusive, overlapping and hierarchical clustering and K-Means is an example of exclusive clustering. The main aim of this algorithm is to group similar elements or data points into a cluster. Here K represents the total number of clusters that need to be created for the process, i.e. if K=4, there are four clusters, for K=5, there are five clusters and so on[3]. By looking into the working of K-Means Algorithm, it is understood that it follows the sequential step in which firstly our main aim is to select the number K so that we decide the number of cluster presents after that we select random K points and assign each data point to their closest centroid, which forms K clusters then we calculate the variance and place a new centroid of each cluster. Then every datapoint is assigned to the new closest centroid of each cluster. If any reassignment occurs, then again we need to calculate variance and place a new centroid of each cluster and move ahead otherwise else we need to end. And finally the model is getting ready[4].

2. Random Forest Algorithm

Random Forest is one of the algorithms that comes under supervised machine learning technique. This algorithm is used for both Classification and Regression problems in machine learning. This algorithm mainly depends on the concept of **ensemble learning,** that is a procedure for combining multiple classifiers to solve a complex problem and for improving the performance of the model[5]. Random Forest is also known as a bagging technique. In this, there are various base learner models and these models are decision trees in Random forest. And in this we select some rows and columns (row and feature sampling) from the dataset and provide to each base learner. After providing the data to each decision tree prediction is done by each base learner and based on the majority votes of prediction, the final output is predicted of which is getting selected[6]. Therefore a higher number of trees in the forest ultimately leads to higher accuracy and prevents the problem of overfitting[5].

3. XGBoost Algorithm

The algorithm XGBoost comes under ensemble learning method. An ensemble learning method provides a systematic solution that combines the predictive power of multiple learners. At the end we get a single model as a result that provides the aggregated output combined from several models.
Those models that are involved in ensembles, also known as base learners, could be either from the same learning algorithm or different learning algorithms[7][8]. The two widely used ensemble learners are bagging and boosting. XGBoost is a boosting technique.

VI. RESULT

During the research we collected the data, performed operations and with the help ofour algorithms we got faster and better performing results and output just asclose to what was expected. After obtaining the output, an output file is generated which contains the status of working wafer and faulty wafer. And finally “we will handover that output file to the client that helps them to get faulty wafers at an early stage and does need any manual intervention.”

VII. CONCLUSIONS

The results presented in this paper show that the use of theMachine Learning model in classifying wafers can be a feasible alternate method to replace manualchecking. In a future research paper, we can also perform and use different machinelearning models on each different cluster and after that on the basis of AUC scores ofall models, we can select the best model having the highest AUC score. Also, we have a probability ofincreasing the number of training samples used for future investigations in order to skew the results towards the best.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012050|

ACKNOWLEDGEMENT

We would like to express our thanks and gratitude to our guide “Ms. SharanyaChandran” for her able guidance and support in completing our project and also for providing us with all the facilities that were required.

REFERENCES

1. Dhanshree M. Gharde[1], Prof. Jayant Adhikari[2], Prof. Priyanka Bhende[3], Dr. Narendra Chaudhari[4] https://www.irjet.net/archives/V8/i5/IRJET-V8I5191.pdf
2. https://www.javatpoint.com/python-tutorial
3. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning
4. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
5. https://www.javatpoint.com/machine-learning-random-forest-algorithm
6. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
7. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
8. https://en.wikipedia.org/wiki/XGBoost
9. YefanZhou http://www.joig.net/index.php?m=content&c=index&a=show&catid=50&id=203
10. https://www.fullstackpython.com/flask.html
11. https://www.analyticsvidhya.com/blog/2021/10/a-complete-guide-on-machine-learning-model-deployment-using-heroku/
12. https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/
13. https://www.zuar.com/blog/data-transformation-types-process-benefits-and-definition/

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details