iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XII Month of publication: December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76521

Prediction of Health Insurance Course Incorporating the Random Forest Algorithm

T Venkateswarlu[1], D Rammohan Reddy[2]
[1]M-Tech Scholar Department of Computer Science Engineering College: Newton's Institute of Engineering, Macharla, Andhra Pradesh, India
[2]Associate professor Department of Computer Science Engineering College: Newton's Institute of Engineering, Macharla, Andhra Pradesh, India

***Abstract:** The project focuses on claims processing, where the accuracy and timeliness of determining claim validity play a crucial role. By training the Random Forest algorithm on a comprehensive dataset consisting of patient records, medical procedures, billing codes, and claim outcomes, the system will learn complex patterns and relationships to predict the likelihood of a claim being valid or potentially fraudulent. This predictive capability will enable insurance companies to expedite the processing of legitimate claims while flagging suspicious ones for further investigation. Fraud detection is another critical aspect of health insurance operations. The project aims to utilize the ensemble learning properties of Random Forest to identify patterns indicative of fraudulent activities. By analyzing features such as billing patterns, provider behavior, and historical fraud cases, the system will build a robust model that can accurately detect and prevent fraudulent claims. This proactive approach will help insurance companies minimize financial losses and protect policyholders from the adverse effects of fraud.*

***Keywords:** Random Forest, Health Insurance, Machine learning, Fraud detection, insurance.*

I. INTRODUCTION

Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that allow computer systems to learn from data and make accurate predictions or decisions. The fundamental idea behind ML is to enable computers to automatically learn and improve from experience without being explicitly programmed. Machine learning has emerged as a powerful and versatile tool in the field of data analysis and decision-making. It has revolutionized various industries, including finance, healthcare, marketing, and technology. The ability of machine learning algorithms to learn patterns, extract insights, and make predictions from large datasets has opened up new possibilities for research and innovation.

Definition of Machine Learning Machine learning can be defined as a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models capable of learning from data and making predictions or decisions without explicit programming. Instead of relying on explicit instructions, machine learning algorithms learn from examples and experience, improving their performance over time.

A. *Key Components of Machine Learning*

1) Data: Machine learning relies on large amounts of data for training and evaluation. Datasets consist of input features (variables) and corresponding output labels or target values, allowing the algorithm to learn the underlying patterns.
2) Algorithms: Machine learning algorithms are the computational models that analyze the data and extract patterns or relationships. They can be broadly categorized into supervised learning, unsupervised learning, and reinforcement learning, depending on the type of learning task
3) Training: During the training phase, machine learning algorithms learn from the provided dataset to build a model that captures the patterns and relationships in the data. The algorithm adjusts its internal parameters iteratively to minimize the error or maximize the objective function.
4) Evaluation: Once the model is trained, it needs to be evaluated on unseen data to assess its performance and generalization ability. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure how well the model predicts the output.

II. LITERATURE SURVEY

A machine learning literature survey involves conducting a comprehensive review of existing research papers, articles, books, and other relevant publications in the field of machine learning. The purpose of a literature survey is to gain a deep understanding of the current state of the art, identify research gaps, and gather insights and knowledge that can inform our own research or project. Here are the steps involved in conducting a machine learning literature survey:

1) Define the Research Topic: Clearly define the specific area or topic within machine learning that you want to survey. This could be a broad topic like "deep learning" or a more specific area like "convolution neural networks for image classification.

2) Identify Relevant Databases and Sources: Determine the key databases, journals, conferences, and other sources where you can find relevant research papers and publications. Common databases and platforms for machine learning literature include IEEE Xplore, ACM Digital Library, arXiv, Google Scholar, and major machine learning conferences like NeurIPS, ICML, and CVPR.

3) Search and Retrieve Papers: Conduct a systematic search using appropriate keywords and filters to retrieve relevant research papers. Refine and iterate our search queries as you explore the literature. Make use of advanced search options and techniques such as Boolean operators, search filters (e.g., publication year, authors), and citation tracking.

4) Evaluate and Select Papers: Assess the relevance and quality of the retrieved papers based on their titles, abstracts, and keywords. Skim through the papers to determine if they are closely related to our research topic. Pay attention to influential papers, recent publications, and papers from reputable authors and institutions.

5) Read and Summarize Papers: Carefully read the selected papers and extract key information. Take notes on the main objectives, methodologies, datasets used, key findings, and limitations of each paper. Summarize the papers in our own words and organize them based on themes, methodologies, or any other relevant categories.

6) Analyze and Synthesize Information: Analyze the findings, methodologies, and results presented in the papers.

7) Identify common trends, patterns, and insights across the literature. Compare and contrast different approaches, methodologies, or models used in the surveyed papers. Look for research gaps, areas that need further exploration, or potential opportunities for our own research.

8) Document and Organize the Survey: Create a systematic and well-structured document to record our literature survey findings. Include detailed summaries of each paper, key insights, references, and any relevant figures or tables. Organize the document in a logical manner to facilitate easy navigation and retrieval of information.

9) Critical Evaluation and Discussion: Critically evaluate the strengths and weaknesses of the surveyed papers. Discuss any limitations, biases, or potential sources of error in the existing literature. Provide a balanced view of the state of the art and highlight areas where further research is needed. Identify gaps, research questions, or opportunities that can guide our own research or project.

10) Citations and References: Ensure proper citation and referencing of the surveyed papers in our literature survey document. Follow the appropriate citation style (e.g., APA, MLA) and maintain consistency throughout the document.

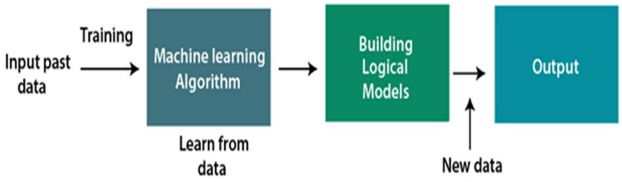
11) Continuous Updating: Keep track of new publications and research in the field of machine learning even after completing the initial literature survey. Machine learning is a rapidly evolving field, and staying updated with the latest research is essential. A literature survey in machine learning involves conducting a comprehensive review and analysis of existing research papers, articles, and publications related to a specific topic or problem in the field of machine learning.

III. MACHINE LEARNING ALGORITHMS

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:

.



At a broad level, machine learning can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

A. *Supervised Learning*

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not. The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering.

Supervised learning can be grouped further in two categories of algorithms:

- Classification
- Regression

B. *Unsupervised Learning*

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

- Clustering
- Association

C. *Reinforcement Learning*

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

IV. PROPOSED METHODOLOGY

A. *Step-by-Step Methodology for Health Insurance Prediction using Random Forest in Machine Learning*

1) Step 1: Data Collection: Gather a comprehensive dataset from a reliable source, including information such as patient demographics, medical procedures, diagnosis codes, billing codes, claim outcomes, and any relevant features for health insurance prediction.

2) Step 2: Data Preprocessing: Perform data cleaning and preprocessing steps to handle missing values, outliers, and inconsistencies in the dataset. This may involve techniques such as imputation, normalization, and handling categorical variables.
3) Step 3: Feature Selection: Select the most relevant features from the dataset that are likely to have an impact on health insurance prediction. Consider factors such as feature importance, correlation analysis, and domain knowledge.
4) Step 4: Data Split: Split the dataset into training and testing subsets. Typically, use a significant portion (e.g., 70-80%) for training the Random Forest model and the remaining portion for evaluating its performance.
5) Step 5: Model Training: Train a Random Forest classifier on the training dataset. Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. Adjust the hyperparameters, such as the number of trees, maximum depth, and minimum samples per leaf, based on experimentation and cross-validation techniques.
6) Step 6: Model Evaluation: Evaluate the trained Random Forest model on the testing dataset to assess its performance. Use appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), to measure the model's effectiveness in health insurance prediction.
7) Step 7: Hyperparameter Tuning: Fine-tune the hyperparameters of the Random Forest model to optimize its performance further. This can be done using techniques like grid search or randomized search, varying the values of the hyperparameters and evaluating the model's performance on a validation dataset.
8) Step 8: Feature Importance Analysis: Analyze the feature importance provided by the Random Forest model to identify the key factors influencing health insurance prediction. This analysis can help in gaining insights into the most significant variables and understanding their impact on the predictions.
9) Step 9: Model Deployment: Once satisfied with the performance of the Random Forest model, deploy it for real-world health insurance prediction tasks. Ensure that the model is integrated with appropriate data pipelines and systems for seamless prediction and decision-makin
10) Step 10: Continuous Monitoring and Maintenance: Regularly monitor the performance of the deployed model and update it as new data becomes available. Over time, retrain the model with updated datasets to ensure its accuracy and relevance in the dynamic health insurance domain. Throughout the entire process, it is essential to maintain proper documentation, record experimental results, and follow best practices in machine learning to ensure reproducibility and transparency in the research.

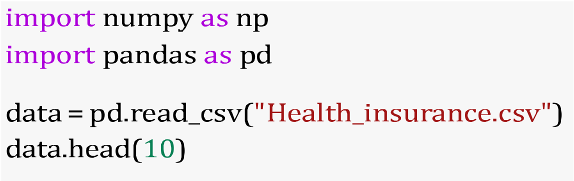
V. EXPERIMENTAL RESULTS

A. Installing Anaconda and Python

To learn machine learning, we will use the Python programming language in this project. So, in order to use Python for machine learning, we need to install it in our computer system with compatible IDEs (Integrated Development Environment).

B. Step-by-Step procedure of Random Forest algorithm

1) Step 1: import the libraries



```
import numpy as np
import pandas as pd

data = pd.read_csv("Health_insurance.csv")
data.head(10)
```

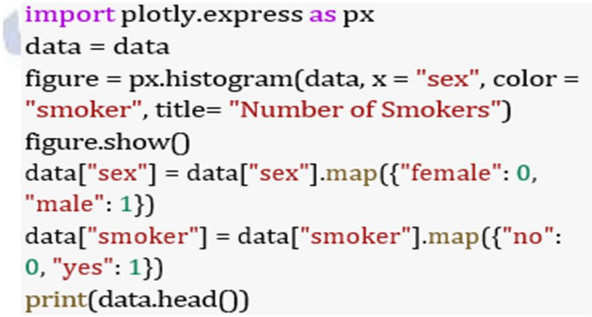
2) Step 2: Visualize the data

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692
10	25	male	26.220	0	no	northeast	2721.32080

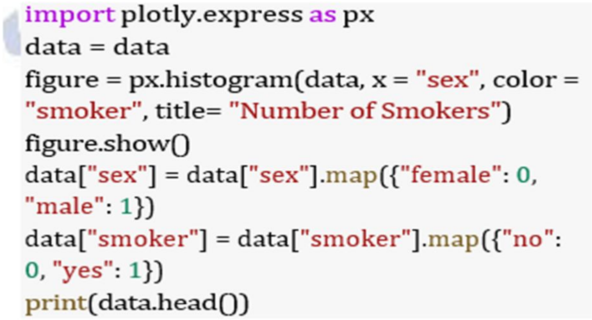
3) Step 3: pre-processing method to avoid the null data

```
data.isnull().sum()
age  0 sex  0 bmi  0 children 0 smoker 0
region 0 charges 0 dtype: int64
```

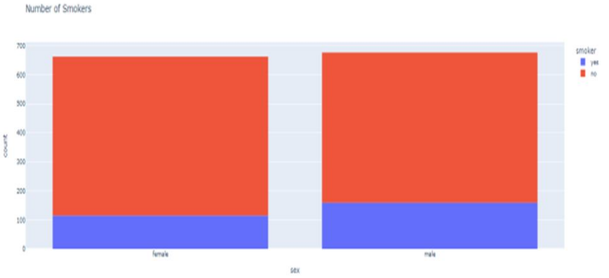
4) Step 4: Data visualization for input data



```
import plotly.express as px
data = data
figure = px.histogram(data, x = "sex", color =
"smoker", title= "Number of Smokers")
figure.show()
data["sex"] = data["sex"].map({"female": 0,
"male": 1})
data["smoker"] = data["smoker"].map({"no":
0, "yes": 1})
print(data.head())
```



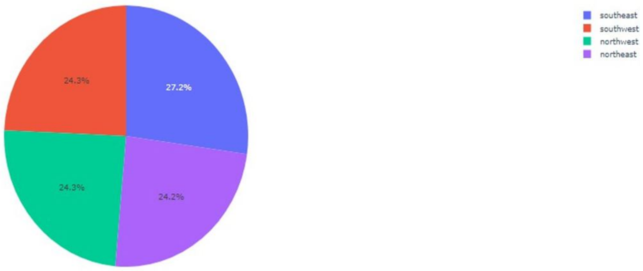
```
import plotly.express as px
data = data
figure = px.histogram(data, x = "sex", color =
"smoker", title= "Number of Smokers")
figure.show()
data["sex"] = data["sex"].map({"female": 0,
"male": 1})
data["smoker"] = data["smoker"].map({"no":
0, "yes": 1})
print(data.head())
```



```
pie = data["region"].value_counts()
regions = pie.index

population = pie.values

fig = px.pie(data, values=population,
names=regions)
```



5) *Step 5: Correlated data*

	age	sex	bmi	children	smoker	charges
age	1	-0.020856	0.109272	0.042469	-0.025019	0.299008
sex	-0.020856	1	0.046371	0.017163	0.076185	0.057292
bmi	0.109272	0.046371	1	0.012759	0.00375	0.198341
children	0.042469	0.017163	0.012759	1	0.007673	0.067998
smoker	-0.025019	0.076185	0.00375	0.007673	1	0.787251
charges	0.299008	0.057292	0.198341	0.067998	0.787251	1

6) *Step 6: Data splitting*

```
x = np.array(data[["age", "sex", "bmi",
"smoker"]])

y = np.array(data["charges"])

from sklearn.model_selection import
train_test_split

xtrain, xtest, ytrain, ytest =
train_test_split(x, y, test_size=0.2
```

7) *Step 7: Random forest Model generator*

```
from sklearn.ensemble import
RandomForestRegressor

forest = RandomForestRegressor()
forest.fit(xtrain, ytrain)
```

8) *Step 8: Health insurance Prediction ypred = forest.predict(xtest)*

```
data = pd.DataFrame(data={"Predicted Premium Amount": ypred})
print (data.head(20))
Predicted                          Premium  Amount
0                                     10306.57708
1                                      5661.90978
2                                     28403.36684
3                                     9815.212534
4                                     34630.60522
5                                        7912.693
6                                      2519.18398
```

7		14882.21941
8	5992.482954	
9	8908.600083	
10	19331.50003	
11	7211.838239	
12	7805.723976	
13	45995.04517	
14	48605.31655	
15	45099.97632	
16	10435.06536	
17	43234.00425	
18	9699.98735	
19	23516.97236	

VI. CONCLUSION

In conclusion, the Random Forest algorithm offers a powerful and versatile tool for health insurance prediction, delivering accurate and interpretable results. Its robustness, feature importance analysis, and ability to handle complex datasets make it a valuable asset in the field of health insurance analytics and decision-making. Future research could focus on further improving the model's performance, exploring ensemble techniques, or integrating additional data sources to enhance prediction accuracy and address specific challenges in the health insurance domain.

REFERENCES

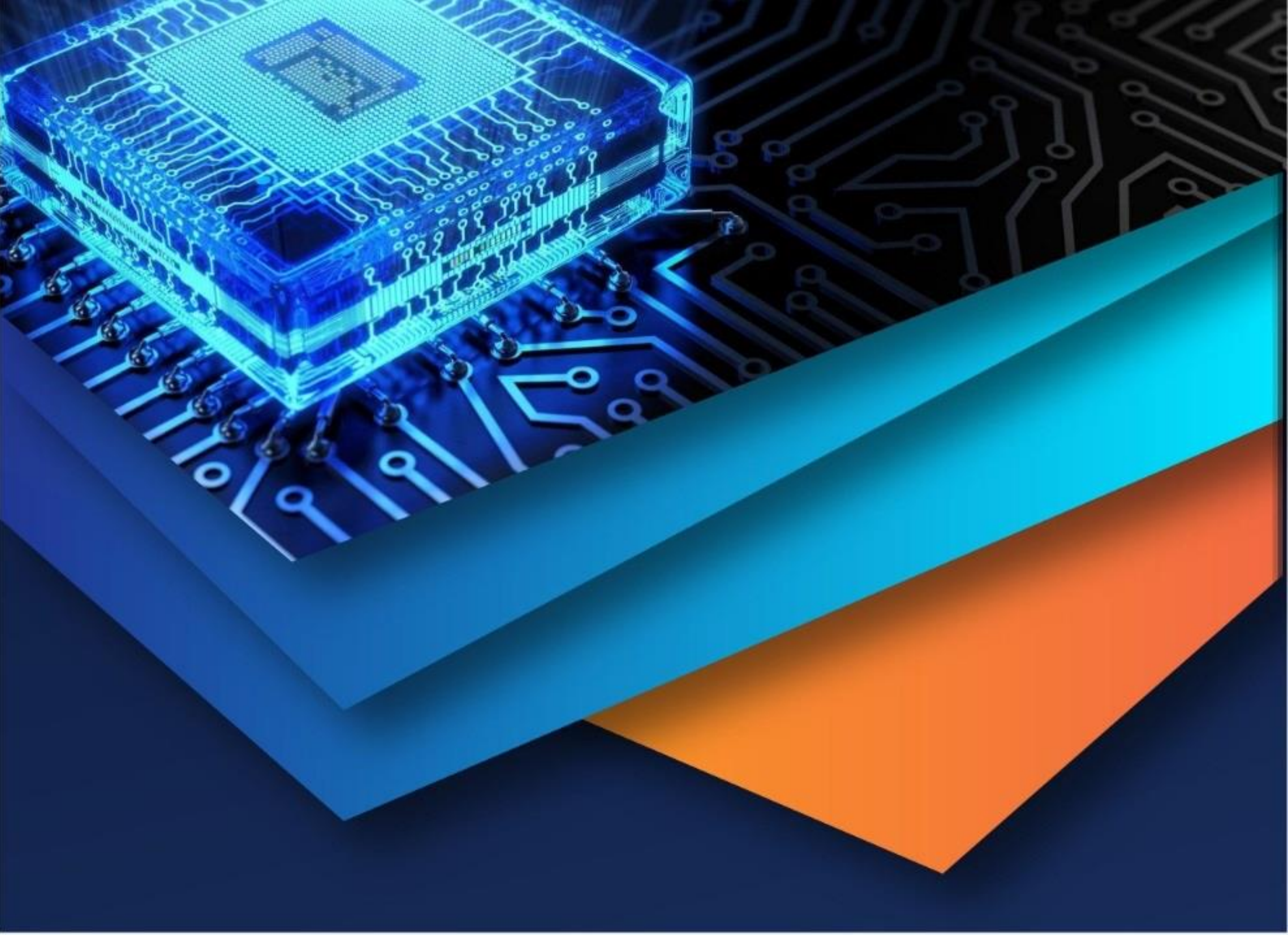
[1] Raghavan P., El Gayar N. “Fraud detection using machine learning and deep learning”, 2019 international conference on computational intelligence and knowledge economy (ICCIKE), IEEE (2019), pp. 334-339.

[2] Awoyemi J.O., Adetunmbi A.O., Oluwadare S.A., “Credit card fraud detection using machine learning techniques: A comparative analysis”, 2017 international conference on computing networking and informatics (ICCNI), IEEE (2017), pp. 1-9.

[3] Breiman L.,”Random forests Machine Learning”, 45 (1) (2001), pp. 5-32.

[4] Eshghi A., Kargari M. “Introducing a new method for the fusion of fraud evidence in banking transactions with regards to uncertainty Expert Systems with Applications”, 121 (2019), pp. 382-392.

[5] Eweoya I., Adebiyi A., Azeta A., Azeta A.E. “Fraud prediction in bank loan administration using decision tree”, Journal of Physics: Conference Series, 1299 (1) (2019).



10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI