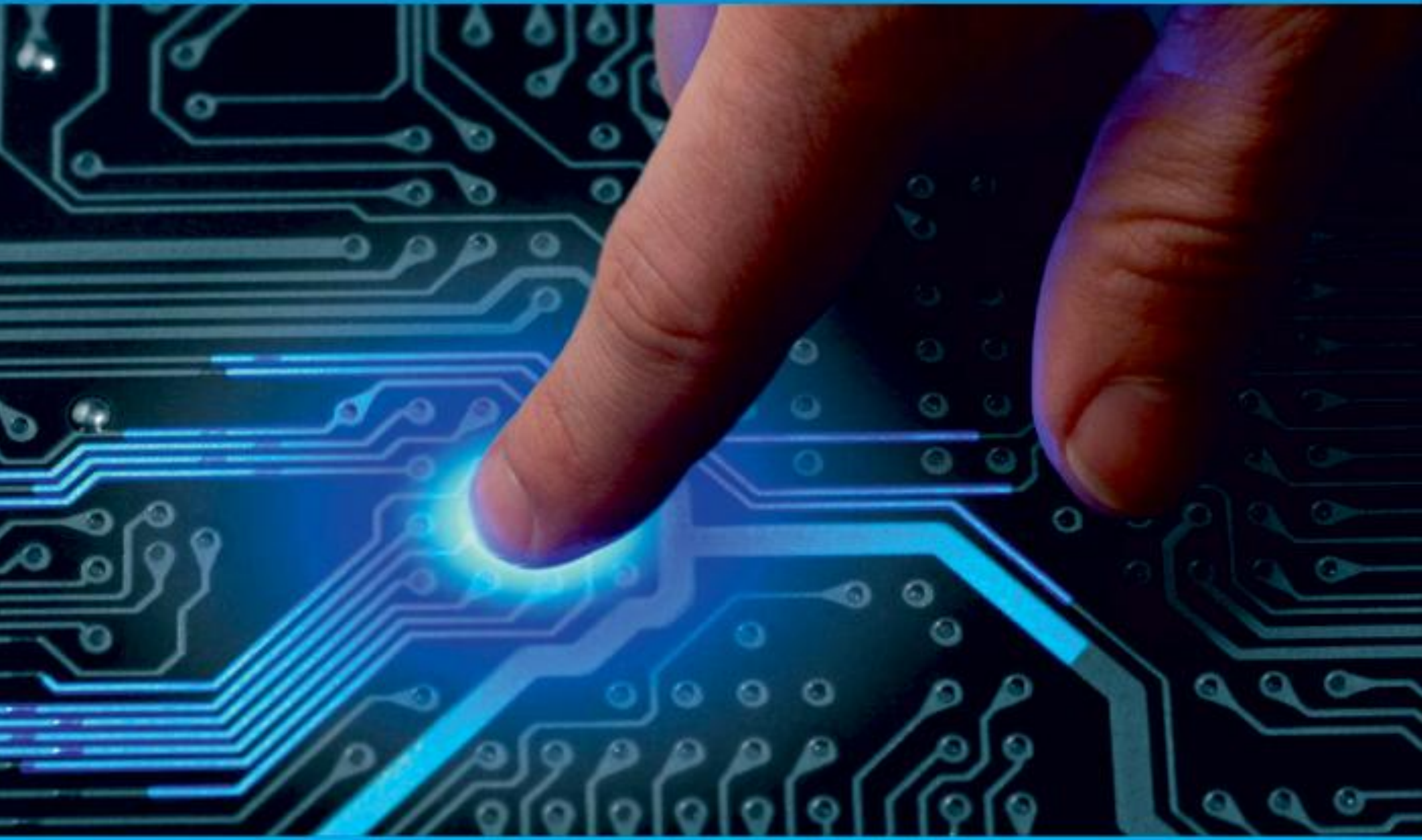
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205079 |

Future Navigator: Machine Learning Approaches to Career Planning

Monishaa K V , Dr. Srikanth V

Student, School of CS and IT, Department of MCA, JAIN (Deemed to be) University, Bengaluru, India

Associate Professor, School of CS and IT, Department of MCA, JAIN (Deemed to be) University, Bengaluru, India

ABSTRACT: This study presents a comprehensive approach to guiding individuals in selecting suitable career paths following higher secondary education. By leveraging a diverse dataset and employing a combination of handcrafted rules, clustering algorithms, and different modelling, we offer personalized career development recommendations. Additionally, we introduce a computerized career counselling system that utilizes objective assessments to predict optimal career paths based on factors such as skills assessment, interests, and academic performance. Our approach aims to minimize the likelihood of individuals choosing unsuitable career paths, thus enhancing the decision-making process for transitioning from education to the workforce. Through the integration of machine learning algorithms and objective assessments, this research contributes to advancing career guidance methodologies and empowering individuals in making informed career decisions.

KEYWORDS: Machine Learning, supervised classification algorithms, Career Guidance System, Random Forest model, Decision tree.

I. Introduction

When college students graduate, they're faced with a lot of career options. It can feel overwhelming trying to pick a job that fits their long-term goals. The same goes for people who are already working—they might wonder if switching jobs or going back to school will help them reach their ambitions. At times like these, people often turn to others with similar backgrounds to see what decisions they made and where they ended up. Instead of just asking a few friends, we're offering a way for people to learn from thousands of others like them. We want to help them find the best career steps to reach their goals.[1][2]

II. Related work

We're looking at people's career paths by mapping out their education and work experiences over time. Each job or educational milestone is like a point on a map. Education includes where someone went to school, what they studied, and what degree they earned. Work experience includes their job title and where they worked. Each person's career path is like a series of these points connected in a sequence. Each point has different features that describe it. Given where someone is in their career now and where they want to go, we're trying to find the best path forward for them—the one with the highest chance of reaching their goal.

This project has some challenges. For example, we need to figure out how to turn someone's background information into useful features. And with so many different job titles out there, we need to find a way to group similar ones together to avoid having too many options to consider. We also need to pick the right method for estimating parameters and giving recommendations about career paths.

The rest of this paper is structured like this: In Section 2, we'll talk implementation, about how we collected and prepared the data. Then, in Section 3, we'll explain the model and approach we used for this problem. Section 4 will present some of the main results we found and analyse them. Finally, in Section 5, we'll wrap up the project and talk about possible future directions.[1]

III. Implementation Strategy

1. Data Collection:

Collecting information is a vital step in any machine learning venture, as the exactness and adequacy of calculations depend on the quality of input information. For anticipating understudy career ways, we accumulate different parameters such as scholarly scores, specializations, programming abilities, interface, extracurricular exercises, and individual subtle elements. Information is sourced from representatives at diverse schools, haphazardly produced, and recovered from school databases.[3][4]

2. Information Preprocessing:

Once information is collected, it needs to be organized and cleaned to be valuable. This includes dealing with invalid values, invalid sections, and undesirable

information. Cleaning guarantees that the information is in a usable arrange by supplanting inaccurate values and evacuating insignificant data. Information preprocessing guarantees that the information is organized and prepared for investigation.

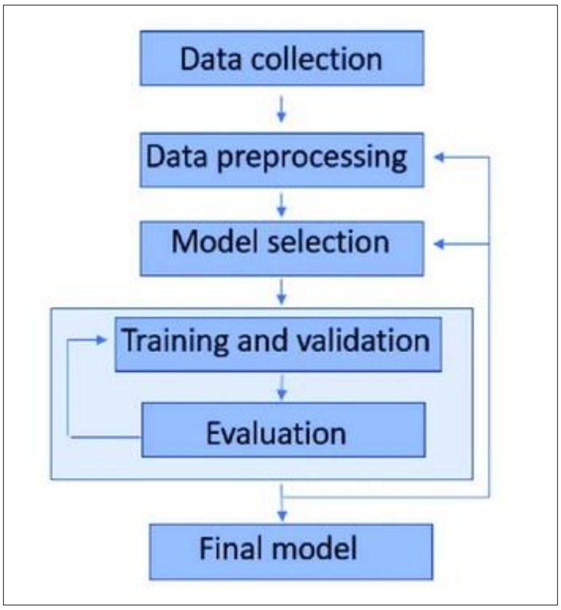


Figure 1: Process flow diagram

There are different strategies of information pre-processing:

- **Information Cleaning**
  - Lost Information
  - Boisterous Data
- **Information Change**
  - Normalization
  - Property selection
- **Information Diminishment**
  - Aggregation
  - Dimensionality Reduction

3. Preparing and Testing After preprocessing:

The information is part into preparing and testing sets. Around 80% of the information is utilized for preparing the machine learning demonstrate, whereas the remaining 20% is saved for testing its execution. Preparing includes educating the demonstrate to make forecasts based on the input information, whereas testing evaluates the model's exactness by comparing its forecasts to known results. If the demonstrate accomplishes tall precision on the test information, it is considered solid for advance use.

4. Classification and Prediction/Evaluation:

Classification is performed utilizing methods such as calculated relapse, credulous Bayes, and back vector machine (SVM) based on understudy interface, pastimes, and scholastic execution. Expectation is made utilizing information collected from different understudies, permitting us to expect their future career ways based on their characteristics and scholastic achievements.

5. Fitting the model:
The nitty gritty steps of the machine learning classification strategies that has been utilized in the proposed work are talked about here one by one.

IV. ALGORITHMS

MACHINE LEARNING ALGORITHM

1.Logistic Regression:
Logistic regression is one of the most important Machine Learning classification algorithms. It is used to predict the probability of categorical dependent variables. It is especially useful when the dependent variable is binary, i.e., 1 (meaning success, yes) or 0 (i.e., failure, no).When the dependent variable is X, the logistic regression model predicts the likelihood (P) that the dependent variable will be 1. It does this by using a mathematical function on the input variables.

Logistic regression is a statistical analysis technique that has become an essential part of machine learning. It allows you to perform predictive analytics on historical data. It is also useful for data preprocessing, as it allows you to categorize the data into pre-defined groups during ETL (Extract, Transform, Load) to prepare the data for further analyisis.

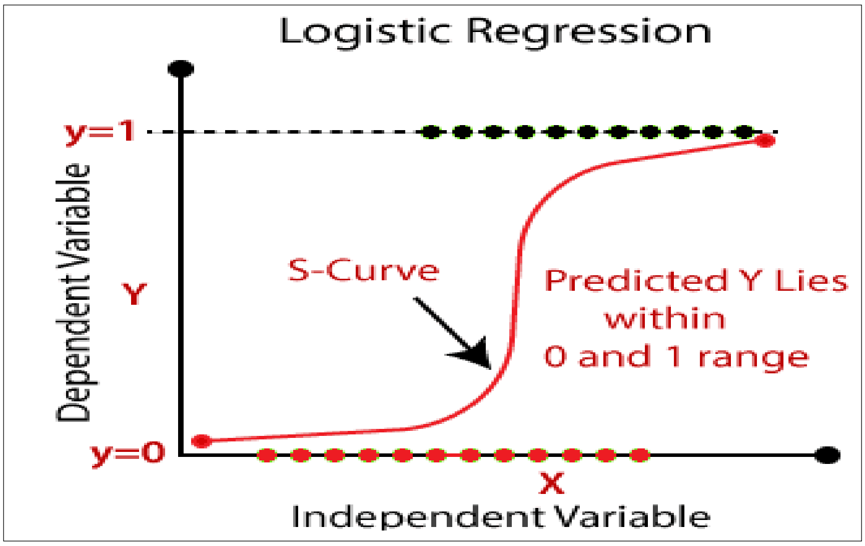


Figure 2: Logistic regression

Figure 2 illustrates the conceptual framework of logistic regression, depicting its role in predicting binary outcomes based
on input variables.

Logistic regression finds application across various domains:

- In healthcare, it aids in identifying risk factors for diseases and formulating preventive measures.
- Weather forecasting apps utilize logistic regression to predict phenomena like snowfall and weather conditions.
- Voting apps leverage logistic regression to gauge the likelihood of voters supporting a particular candidate.
- In the insurance sector, this algorithm is deployed to assess the probability of policyholders passing away before the policy term ends, based on factors like gender, age, and health status.
- Banking institutions utilize logistic regression to predict the likelihood of loan default by applicants, considering variables such as income, credit history, and debt obligations.
- This versatile algorithm plays a pivotal role in predictive modeling across diverse industries, offering valuable insights into binary outcome probabilities and aiding in informed decision-making processes.[13]

2. Naive Bayes:
Naive Bayes is a classification technique grounded in the Naïve Bayes Theorem, which operates under the assumption of independence among predictors. In essence, this classifier posits that the presence of a specific feature within a

classis unrelated to the presence of any other feature. Despite its apparent simplicity, Naive Bayes emerges as a remarkably potent algorithm for predictive modeling.
The model comprises two fundamental probabilities derived directly from the training data:
1.The probability of each class, and
2.The conditional probability for each class given each value of x.

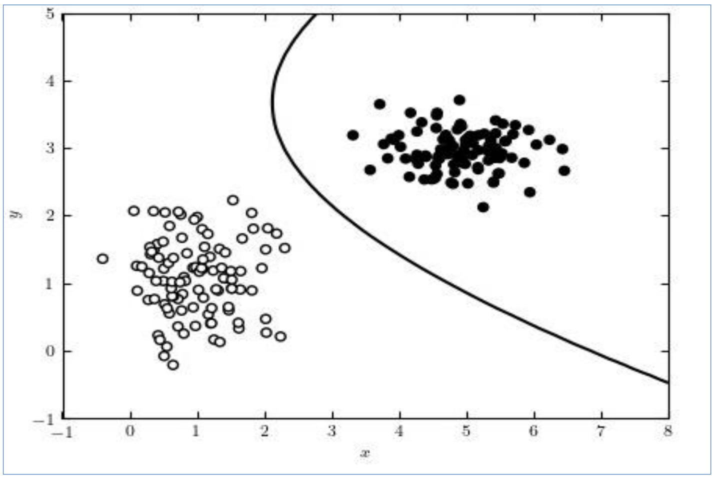


Figure 3: Naiver BayesBayes

Once these probabilities are computed, the model becomes equippedto generate predictions for new data using Bayes Theorem.

$$\boxed{P(A|B) = P(B|A) * P(A) / P(B)}$$

Theorem is expressed as follows:

Here,
P(A|B) denotes the probability of A occurring given evidence B has already occurred,
P(B|A) represents the probability of B occurring given evidence A has already occurred,
P(A) signifies the probability of A occurring, and
P(B) indicates the probability of B occurring.
Naive Bayes amalgamates these diverse strands of information to deliver robust predictions regarding the class for a new set of features. By synthesizing the conditional probabilities of each feature given a specific class, along with the prior probabilities of each class, Naive Bayes offers a principled approach to classification, enabling accurate predictions in a variety of real-world scenarios.

3. Support Vector Machines (SVM):

Support Vector Machines (SVM) are a class of powerful, supervised machine learning models capable of performing both classification and regression tasks. At its core, SVM aims to find the optimal separating hyperplane that distinguishes between classes of data points in an n-dimensional space, where n represents the number of features.
The decision boundary, or the hypothesis plane, is mathematically represented as w^Tx + b = 0, where w is the weight vector, x is the feature vector, and b is the bias. Surrounding this decision boundary are two additional planes: the positive plane (w^Tx + b = +1) and the negative plane (w^Tx + b = -1), which respectively represent the margins closest to the positive and negative classes. The distances from these planes to the nearest data points (support vectors) are denoted as d+ and d-, aiming to maximize the margin between these planes to enhance the model's generalization ability.
In a theoretical visualization, data points of one class might be marked as stars (positive points) and those of another class as circles (negative points), with the decision boundary clearly separating them. This separation showcases the

SVM's capability to classify linearly separable data effectively. However, SVM is also equipped to handle non-linearly separable data through the use of kernel tricks, enabling the model to find a separating hyperplane in the transformed feature space.

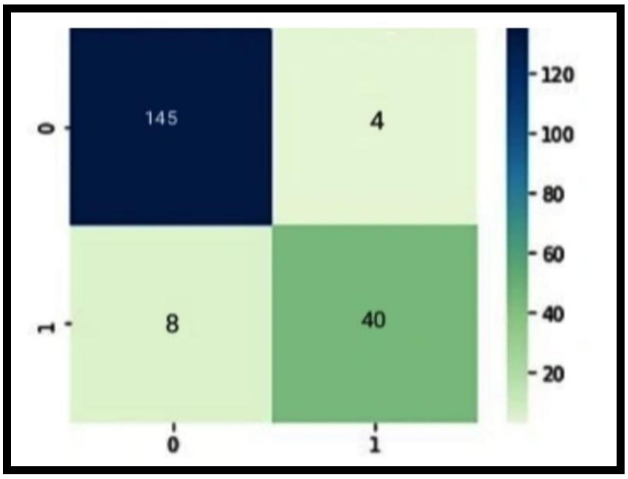
SVM's methodology of using support vectors to define decision boundaries and its flexibility in handling both linear and non-linear datasets make it an effective tool for various applications. It is particularly noted for its robustness in cases where the dimensionality of the feature space is high relative to the number of samples, reducing the risk of overfitting and ensuring computational efficiency.[5][6][7]

V. Simulation Results

1.(Doctor)Logistic Regression analysis:

Accuracy = 145+40/145+4+8+40
= 93.90

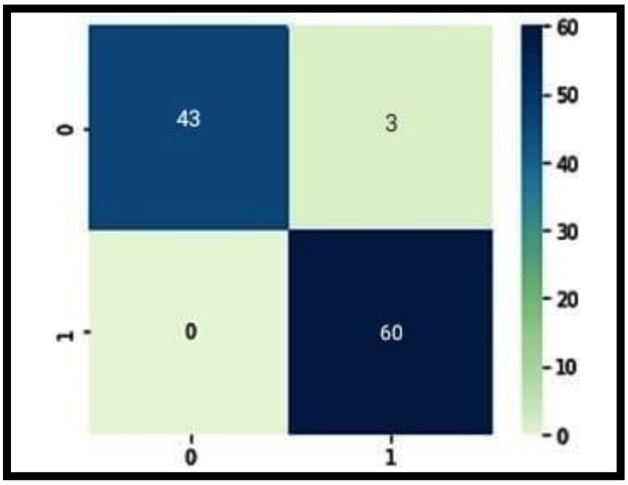
Hence the accuracy obtained is 93.90%.



2.(CA)Naive Bayes:

Accuracy= 43+60/43+3+0+60
= 97.16

Hence the accuracy obtained is 97.16%.

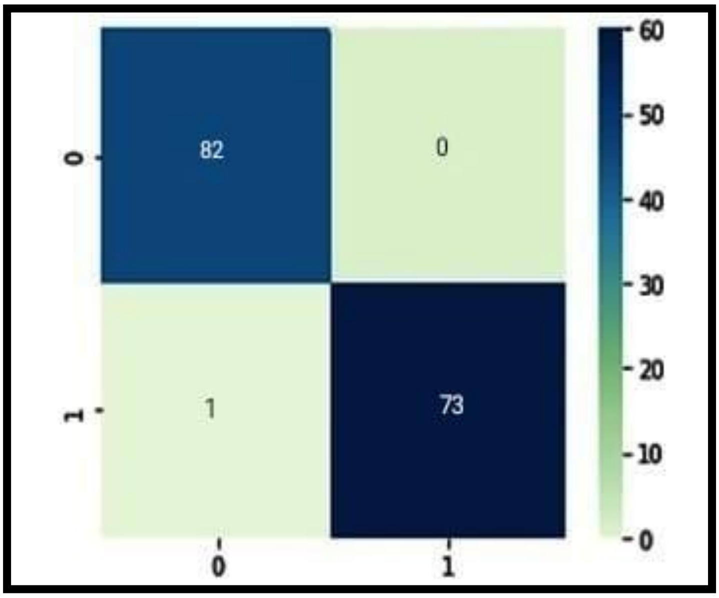


3.(Hotel management) Support Vector Machines:

Accuracy= 82+73/82+0+1+73
= 99.36%

Hence the accuracy obtained is 99.36%.

The dataset underwent training and testing using three algorithms: Support Vector Machines (SVM), Logistic Regression, and Naive Bayes. SVM outperformed the others with an accuracy of 99.36%. Logistic Regression followed with 88.33% accuracy, and Naive Bayes came in third.



SVM's superior performance underscores its effectiveness in handling the dataset. Its high accuracy indicates its ability to create precise models for classifying new data. Therefore, SVM was chosen for all future predictions due to its consistent results across various tests and potential to improve with more data.

This decision is based on SVM's reliability and its capacity to refine accuracy over time by incorporating new data. Using SVM for ongoing predictions is expected to enhance the model's precision as it gains access to more training data.

The study suggests adopting SVM for scenarios where high accuracy in classification is crucial. Its effectiveness makes it a valuable tool for predictive modeling, particularly for applications requiring precise and reliable classification outcomes.[9][13][15]

CHALLENGES AND LIMITATIONS:

1.Logistic regression challenges:

- **Straight line problem:** this method likes to draw straight lines (linear relationships) to separate different careers. But real-life career choices aren't always that simple and straight.
- **Overlapping features:** sometimes, the features (like skills or education level) that influence career paths overlap too much, confusing the model.
- **Too simple for complex data:** logistic regression can be too simple for complicated career data, missing out on hidden patterns.

2. SVM challenges:

- **Complex setup:** picking the right settings (like the kernel trick) for svm can be tricky. It's like trying to choose the right lens to view the data through, and sometimes it’s hard to know which one is best.

- **Slow with big data**: svm can get really slow when you have a lot of data, making it hard to scale up for big career datasets.
- **Hard to understand:** the way svm makes decisions can be complex, making it tough to explain why it recommends certain careers.

3.Naive Bayes challenges:

- **Assumption of independence:** naive bayes assumes each feature (like skills or interests) works independently to influence career paths. But in reality, many of these features affect each other.
- **Rare data problems:** if some career paths have very little data, naive bayes might struggle to make accurate predictions about them.
- **Simplicity vs. Accuracy:** naive bayes is simple and fast but might not catch all the complexities of career prediction, leading to less accurate suggestions.[10][11]

LIMITATIONS ACROSS ALL ALGORITHMS:

- **Fitting problems:** all three methods can either get too fixated on the training data (overfitting) or not learn enough from it (underfitting), leading to poor predictions.
- **Data bias:** if the data used to train these models has biases (like preferring certain careers for certain groups of people), the models might repeat those biases in their predictions.
- **Keeping up with changes:** the job market changes fast, with new careers popping up and old ones disappearing. These models need regular updates to stay relevant, which can be a challenge.

VI. CONCLUSION AND FUTURE WORK

In short, while these algorithms can be powerful tools for predicting career paths, they each have their own set of challenges, especially when dealing with the complex and ever-changing nature of the job market.[13]

REFERENCES

1. Nikita Gorad, Ishani Zalte,Career,“Censuring exercising Information Mining” Around the world Diaryof Inventive Inquire around in Computer and Communication Coordinating.Vol. 6,No. 18, January 2015.
2. Roshani Ade,Dr.P.R. Deshmukh, “ An incremental gathering of classifiers as a methodology for pine forof pupil’s career choice ”, To begin with Around the world Conference on Making Plans in Organizing andDevelopment,Vol. 5,No. 13, December 2015.
3. Rutvija Pandya, Jayati Pandya, “ C5.0 computation to Progressed Choice Tree with Highlightinstrument and dropped Botch Pruning “, Around the world Diary of Computer Applications(0975 – 8887),Volume 117,No. 16, May 2015.
4. Ali Daud, Naif Radi Aljohani, “ Predicting Understudy prosecution exercising Progressed LearningAnalytics ”, 2017 Around the world World Wide Web Conference Committee(IW3C2). Volume 13,No. 6,July 2016.
5. Anuj Karpatne, Gowtham Atluri, “ proposition- Guided Information Science A Unused Worldview for Strong Revelation from Data ”, IEEE Exchanges on Information and Information Coordinating,vol. 29,no.10, October 2017.
6. Bo Guo, Rui Zhang, “ Predicting Understudies prosecution in preceptors Information Mining ”, Allcomprehensive Symposium on preceptors Alter,Vol. 143,No. 8, Predominant 2016.
7. T. Pederson,S.Patwardhan,J. Michelizze, andS. Banerjee." Wordnetsimilarity", 2008.Available http//www.d.umn.edu/tped-erse/similarity.html.
8. public sodalities rankings- us news, 2010. Open http//colleges.usnews.rankingsandreviews.com/bestcolleges/national-universities-rankings.
9. Wordnet, 2006. Open http//wordnet.princeton.edu/.
10. SubahiA.,F.(2018). Information Collection for Career Way Figure Grounded on assaying Body ofInformation of Computer Science Degrees. Diary of Computer program, Volume 13.
11. UmarM.A.(2019). Understudy Quick prosecution Bear exercising Fake Neural Systems A CaseConsider. Around the world Diary of Computer Applications(0975 – 8887), Volume 178.
12. DawoodE.ABD.E., ElfakhranyE., MaghrabyF.A.(2017). Make strides sketching bank client’s gesteexercising machine literacy. IEEE Access.

13. RoyK.S., RoopkanthK., UdayV., BhavanaV., PriyankaJ.(2018). Understudy Career Figure exercisingProgressed Machine Learning Strategies. Around the world Diary of Building & Technology.
14. LiuY., ZhangL., NieL., YanY., RosenblumD.S.(2016). Fortune Teller foreknowing Your Career Way.
15. UhlerB.D., HurnJ.E.(2013). exercising Learning Analytics to Anticipate(and Development)Understudy Triumph A Staff Point of see. Diary of Data Online Learning, Volume 12.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING