

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“Jnana Sangama”, Belgaum – 590 018



A PROJECT REPORT ON

“STUDENT PERFORMANCE PREDICTION IN CORPORATE WORLD”

Submitted in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

BY

Tabrez Ahmed Khan N (1CR16IS115)

Saurav Praveen (1CR16IS094)

Shivam Kumar Agrawal (1CR16IS100)

Sulabh Nand Tiwary (1CR16IS111)

Under the guidance of

Dr. S. Geetha

(Associate Professor, Dept. of ISE, CMRIT)



DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CMR INSTITUTE OF TECHNOLOGY,

BANGALORE-37

**CMR INSTITUTE OF TECHNOLOGY
BANGALORE-37**



Department of **Information Science & Engineering**

Certificate

This is to certify that the project entitled “**STUDENT PERFORMANCE PREDICTION IN CORPORATE WORLD**” has been successfully completed by **Tabrez Ahmed Khan N (1CR16IS115)**, **Saurav Praveen (1CR16IS094)**, **Shivam Kumar Agrawal (1CR16IS100)** and **Sulabh Nand Tiwary (1CR16IS111)**, bonafide students of CMR Institute of Technology in partial fulfillment of the requirements for the award of degree of Bachelor of Engineering in **Information Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the academic year **2019-2020**. It is certified that all the corrections/suggestions indicated for Internal Assessment have been incorporated in the project report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Name & Signature of
Guide (Dr. S. Geetha)

Name & Signature of HOD
(Mrs. Farida Begum)

Signature of Principal
(Dr.Sanjay Jain)

External Viva

Name of the Examiners

Signature with date

- 1.
- 2.

**CMR INSTITUTE OF TECHNOLOGY
BANGALORE-37**



Department of Information Science & Engineering

DECLARATION

We, **Tabrez Ahmed Khan N (1CR16IS115)**, **Saurav Praveen (1CR16IS094)**, **Shivam Kumar Agrawal (1CR16IS100)** and **Sulabh Nand Tiwary (1CR16IS111)**, bonafide students of CMR Institute of Technology, Bangalore, hereby declare that the dissertation entitled, “**STUDENT PERFORMANCE PREDICTION IN CORPORATE WORLD**” has been carried out by us under the guidance of **Dr. S. Geetha, Associate Professor**, CMRIT, Bangalore, in partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in **Information Science and Engineering**, of the Visvesvaraya Technological University, Belgaum during the academic year 2019-2020. The work done in this dissertation report is original and it has not been submitted for any other degree in any university.

Tabrez Ahmed Khan N (1CR16IS115)
Saurav Praveen (1CR16IS094)
Shivam Kumar Agrawal (1CR16IS100)
Sulabh Nand Tiwary (1CR16IS111)

ABSTRACT

In the nowadays-competitive race of finding a suitable talented, qualified, bright and potential personnel to fulfill the needed spot of a vacancy in an industry, and with the beginning of the fourth industrial revolution, employers are taking the hiring process to the digital world. An upcoming challenge is raised where if the new candidates for a vacancy will give the expected performance based on the hiring criteria's or not, and how to hire a candidate that will while dealing with the hiring process? Employers are concerned with the performance evaluation of their current employees. This study is proposing a follow-up conceptual model of using Artificial Intelligent (AI) in the hiring process with the using of performance management and social screening to predict the new candidate expected performance by analyzing historical performances and conditions of employees. An upcoming challenge is raised where if the new candidates for a vacancy will give the expected performance based on the hiring criteria's or not, and how to hire a candidate that will while dealing with the hiring process? Employers are concerned with the performance evaluation of their current employees, but it is a challenge knowing the performance of new candidates before hiring. This study is proposing a follow-up conceptual model of using Artificial Intelligent (AI) in the hiring process with the using of performance management and social screening to predict the new candidate expected performance by analyzing historical performances and conditions of employees. This method will give an additional parameter that assists the decision makers in the hiring process.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance.

So, it is with gratitude that we acknowledge all those whose guidance and encouragement served as beacon of light and crowned our effort with success.

We would like to thank **Dr. Sanjay Jain**, Principal, CMRIT, Bangalore, for providing an excellent academic environment in the college and his never-ending support for the B.E program.

We would like to express our gratitude towards **Dr. Farida Begum**, Assoc. Professor and HOD, Department of Information Science and Engineering CMRIT, Bangalore, who provided guidance and gave valuable suggestions regarding the project.

We consider it a privilege and honor to express our sincere gratitude to our internal guide **Dr. S. Geetha**, Associate Professor, Department of Information Science & Engineering for their valuable guidance throughout the tenure of this project work.

We would also like to thank all the faculty members who have always been very Co-operative and generous. Conclusively, we also thank all the non-teaching staff and all others who have done immense help directly or indirectly during our project.

Tabrez Ahmed Khan N
Saurav Praveen
Shivam Kumar Agrawal
Sulabh Nand Tiwary

Table of Contents

CHAPTER 1

1.PREAMBLE.....	1
------------------------	----------

1.1Introduction.....	1
-----------------------------	----------

CHAPTER 2

2. LITERATURE SURVEY	4
-----------------------------------	----------

CHAPTER 3

3. SYSTEM REQUIREMENT SPECIFICATION.....	8
---	----------

3.1 Functional Requirement.....	8
--	----------

3.2 Non-Functional Requirement.....	9
--	----------

3.2.1 Product Requirement.....	9
---------------------------------------	----------

CHAPTER 4

4. SYSTEM DESIGN	11
-------------------------------	-----------

4.1 System development methodology.....	11
--	-----------

CHAPTER 5

5. IMPLEMENTATION	19
--------------------------------	-----------

CHAPTER 6

6. FUTURE SCOPE AND CONCLUSION	23
---	-----------

REFERENCES.....	24
------------------------	-----------

Chapter 1

PREAMBLE

1.1 Introduction

The Hiring Process of a company is a very crucial thing as it defines what kind of talent will be taken into the company. The recruiters need to be clear about the requirements of the company. Our application intends to bring a bit of science to the hiring process by evaluating candidates based on a vast number of attributes. We take into account the aptitude tests, the scores awarded to the candidate after the interview by the recruiting team, the behavior of the candidate as seen by the college faculty, his abilities to work in a team, etc. All these values will be quantified on a particular scale and then given as input to our application. ONE of the most concerns to any business industry is the human's capital, which focuses on profits maximization and cost minimization (WallerStone, 1980). Human resource is a critical process in the capitalization, finding peoples with effective ideas to benefit the financial impact of the industry and avoid hiring people with non-effective ideas who can cost the industry financially. More employers than ever are struggling to fill open jobs. Forty-five percent say they cannot find the skills they need, and for large organizations (250+ employees), it is even higher with 67% reporting talent shortages in 2018. Accordingly, employers are experiencing a financial loss due to bad hiring decisions, according to a survey the average cost of one bad hire is nearly \$15,000; average cost of losing a good hire is nearly \$30,000(CareerBuilder, 2017). Another task of the Human resources is performance management, the process where HR is keeping track of the employee's performances and ensures the motion of productivity is meeting the industry's goals.

The concerning thing is, Will the new candidates for a vacancy give the expected performance based on the hiring criteria's or not? In addition, how to hire a candidate that will meet the performance expectation while dealing with the hiring process? Employers are concerned with the performance evaluation of their current employees, but it is a challenge knowing the performance of new candidates before hiring. In the hiring process, finding an approximate estimation of the new candidate performance will assist the recruiters with the

decision of the hire. Normal techniques of performance predicting can be sophisticated and timely cost, knowing who is going to perform best in the required job role. Methods such as knowledge tests, cognitive tests, personality tests, reference checks, structured/unstructured interviews, work samples, and integrity tests and others are used in the normal hiring process (Warton, people Analytic). Majority of students in higher education join a course for securing a good job. Therefore taking a wise career decision regarding the placement after completing a particular course is crucial in a student's life. An educational institution contains a large number of student records.

Therefore finding patterns and characteristics in this large amount of data is not a difficult task. Higher Education is categorized into professional and non-professional education. Professional education provides professional knowledge to students so that they can make their stand in corporate sector. Professional education may be technology oriented or it may be totally concentrating on improving managerial skills of candidate. We apply data mining techniques using Decision tree and Naïve Bayes classifier to interpret potential and useful knowledge. In addition, artificial intelligence (AI) machine learning software on social media or called screening a process whereby selectors browse social media profiles of applicants in order to find positive or negative information that may help them decide if the candidate is suitable for an open position. Social media screening is very valuable to recruiters, hiring managers, it humanizes the candidate and makes them more than just the words on their resume. There are various tools for social media screening such as LinkedIn Profile API, FAMA, and Meltwater. With the use of performance management and Screening techniques, a procedure that can help decision-makers with finding and hiring the best candidate by predicting his\her performance according to generated performance patterns using Machine-learning techniques based on historical performance. Machine learning is one type of artificial intelligence methods; it is a computational program goaled to optimize a process performance based on learned information extracted from training on a data set.

1.2 Data Mining

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or 'mining' knowledge from large amounts of data. Data mining techniques are used to

operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Data mining is: Discovering the methods and patterns in large databases to guide decisions about future activities. It is expected that data mining tools to get the model with minimal input from the user to recognize. The model presented can be useful to understand the unexpected and provide an analysis of data followed by other tools to put decision-making are examined and it ultimately leads to strategic decisions and business intelligence. The simplest word for knowledge extraction and exploration of volume data is very high and the more appropriate name for this term is "Exploring the knowledge of database". A database is knowledge of discovery process. This process includes the preparation and interpretation of results. Classification is the most commonly applied data mining technique, which employs a set of pre-classified attributes to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data sets. The classifier-training algorithm uses these pre-classified attributes to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Knowing the factors for placement of student can help the teachers and administrators to take necessary actions so that the success percentage of placement can be improved. Predicting the placement of a student needs a lot of parameters to be considered. Prediction models that include all personal, social, psychological and other environmental variables are necessitated for the effective prediction of the placement of the students.

Chapter 2

LITERATURE SURVEY

Machine learning is one type of artificial intelligence methods; it is a computational program goaled to optimize a process performance based on learned information extracted from training on a data set. One use of this ML learned information is predictions in the future or description of a status. In the proposed module by (AlRadaideh and Alnagi 2012), the data set was a table of personal information such as Age, Gender, Marital status, education degree...etc. collected among three IT companies using a survey. This information assumed to have an effect on the performance of an employee, applying ID3, C4.5 and Naïve Bayes algorithms using J4.8 in WEKA to generate a decision tree, which returned highest information gain for the attribute "WorkRating". Apply the ID3 and C4.5 to generate a decision tree and Naïve Bayes algorithms in each phase to build the classification model. The proposed enhancement will result a prediction not only for the performance evaluation as a total but on the level of the skill by dividing the historical evaluation to be depending on the skills areas, this addition to the attributes will raise the specificity and accuracy of the prediction process. In order to have the maximum accuracy, the performance evaluation processed on areas separately, as an example; the employee evaluated on the level of Ethics and then evaluated on another area of skills separately.

When the evaluation process finishes, the evaluation results categorized to areas in such how the employee performed in each area in that period of evaluation, then stored in the historical evaluations. These evaluation areas are bounded to the employees' statuses in the same period of evaluation, these values are processed together to initiate the bounded values and stored in the knowledge base. More the evaluations and more the statuses, more the values and new related information generated and stored in the knowledge base. This information is collected from multiple sources and maintained in the data set. The regular surveys at the time of evaluations, social media screening and areas performance evaluations are the key sources of this data set. Certain patterns of performances learned by relating these combined evaluations and collected data. When a new candidate is applying for a vacancy,

his\her information collected and bounded to the job role of application, these inputs compared with the results in the knowledge base. Based on these learned patterns, a prediction of a new candidate performance evaluation is derived, comparing a single area in the new evaluation with a similar area in previous evaluations a prediction in this certain area is also can be predicted. Data mining techniques has evolved its research very well in the field of education in a massive amount. This tremendous growth is mainly because it contributes much to the educational systems to analyze and improve the performance of students as well as the pattern of education. Various works had been done by a large number of scientists to explore the best mining technique for performance monitoring and placement. Few of the related works are listed down to have a better understanding of what should be carried on in the past for further growth. Han and Kamber [4] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. Bhardwaj and Pal [5] conducted study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance. Tongshan Chang, & Ed.D [6] introduces a real project to assist higher education institutions in achieving enrollment goals using data mining techniques Furthermore, the results also provide evidence that data mining is an effective technology for college recruitment. It can help higher education institutions mange enrollment more effectively. Pandey and Pal [7] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not. Hijazi and Naqvi [8] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and

mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance. Khan [9] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to

establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socioeconomic status had relatively higher academic achievement in general. Z. J. Kovacic [10] presented a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. The algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively. Galit [11] gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams. Yadav, Bhardwaj and Pal [12] conducted study on the student retention based by selecting 398 students from MCA course of VBS Purvanchal University, Jaunpur, India. By means of classification they show that student's graduation stream and grade in graduation play important role in retention. Al-Radaideh, et al [13] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the NaïveBayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models. Sudheep Elayidom , Sumam Mary Idikkula & Joseph Alexander [14] proved that the technology named data mining can be very effectively applied to the domain called employment prediction, which helps the students to choose a good branch that may fetch them placement. A generalized framework for similar problems has been proposed. Baradwaj and Pal [15] obtained the university

students data like attendance, class test, seminar and assignment marks from the students' previous database, to predict the performance at the end of the semester. Ayesha, Mustafa, Sattar and Khan [16] describe the use of k-means clustering algorithm to predict student's learning activities. The information generated after the implementation of data mining technique may be helpful for instructor as well as for students. Pal and Pal [17] conducted study on the student performance based by selecting 200 students from BCA course. By means of ID3, c4.5 and Bagging they find that SSG, HSG, Focc, Fqual and FAIn were highly correlated with the student academic performance. Bray [18], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socioeconomic conditions. Yadav, Bhardwaj and Pal [19] obtained the university students data like attendance, class test, seminar and assignment marks from the students' database, to predict the performance at the end of the semester using three algorithms ID3, C4.5 and CART and shows that CART is the best algorithm for classification of data.

Chapter 3

SYSTEM REQUIREMENT SPECIFICATION

A System Requirement Specification (SRS) is basically an organization's understanding of a customer or potential client's system requirements and dependencies at a particular point prior to any actual design or development work. The information gathered during the analysis is translated into a document that defines a set of requirements. It gives the brief description of the services that the system should provide and also the constraints under which, the system should operate. Generally, SRS is a document that completely describes what the proposed software should do without describing how the software will do it. It's a two-way insurance policy that assures that both the client and the organization understand the other's requirements from that perspective at a given point in time.

SRS document itself states in precise and explicit language those functions and capabilities a software system (i.e., a software application, an ecommerce website and so on) must provide, as well as states any required constraints by which the system must abide. SRS also functions as a blueprint for completing a project with as little cost growth as possible. SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it. Requirement is a condition or capability to which the system must conform. Requirement Management is a systematic approach towards eliciting, organizing and documenting the requirements of the system clearly along with the applicable attributes. The elusive difficulties of requirements are not always obvious and can come from any number of sources.

3.1 Functional Requirement

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality.

The functional requirements for this application are:

- It should provide good UI for an individual to work upon.
- It should predict the performance of the student with the maximum accuracy
- There should minimum deviation between predictions for relatively similar inputs
- All invalid inputs should be handled correctly
- The UI should be user friendly

3.2 Non Functional Requirement

Non functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviors. They may relate to emergent system properties such as reliability, response time and store occupancy. Non functional requirements arise through the user needs, because of budget constraints, organizational policies, the need for interoperability with other software and hardware systems or because of external factors such as:-

- Product Requirements
- Basic Operational Requirements

3.2.1 Product Requirements

Platform Independency: Standalone executables for embedded systems can be created so the algorithm developed using available products could be downloaded on the actual hardware and executed without any dependency to the development and modeling platform.

Correctness: It followed a well-defined set of procedures and rules to compute and also rigorous testing is performed to confirm the correctness of the data.

Ease of Use: Model Coder provides an interface which allows the user to interact in an easy manner.

Modularity: The complete product is broken up into many modules and well-defined interfaces are developed to explore the benefit of flexibility of the product.

Robustness: This software is being developed in such a way that the overall performance is optimized and the user can expect the results within a limited time with utmost relevancy and correctness.

Non-functional requirements are also called the qualities of a system. These qualities can be divided into execution quality & evolution quality. Execution qualities are security & usability of the system which are observed during run time, whereas evolution quality involves testability, maintainability, extensibility or scalability.

3.3 Hardware Requirements

- Pentium IV or higher, (PIV-3.0 GHz recommended)
- 256 MB RAM
- 256 MB hard free drive space

3.4 Software Requirements

- Python
- Jupyter Notebook
- Operating System: Windows 10 / Windows XP

Chapter 4

SYSTEM DESIGN

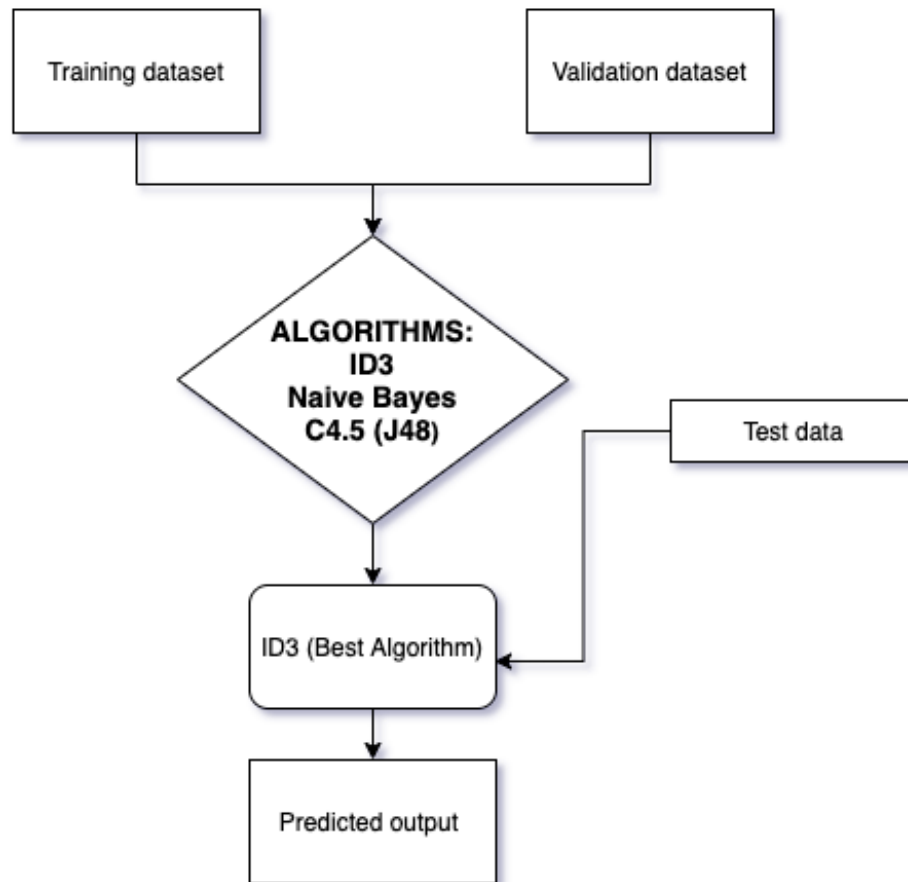
Design is a meaningful engineering representation of something that is to be built. It is the most crucial phase in the developments of a system. Software design is a process through which the requirements are translated into a representation of software. Design is a place where design is fostered in software Engineering. Based on the user requirements and the detailed analysis of the existing system, the new system must be designed. This is the phase of system designing. Design is the perfect way to accurately translate a customer's requirement in the finished software product. Design creates a representation or model, provides details about software data structure, architecture, interfaces and components that are necessary to implement a system. The logical system design arrived at as a result of systems analysis is converted into physical system design.

4.1 System development methodology

System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps which is used for product progress.

- Test dataset and training set were run against various classification algorithms to get the algorithm which gives the most accurate output in Weka software.
- Algorithms give the most accurate output (best prediction) depending upon the dataset.
- For our Dataset , ID3 is the best.
- Decision Tree generated by running ID3 was used as the logic for the backend code.
- User is asked to enter his details (attributes) in the application.
- These details are then run through the code developed using the above logic to predict his performance.

The Result is displayed in the frontend to the User.



Algorithms Used

ID3:

INTRODUCTION

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan[1] used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains.

Algorithm:

The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy $H(S)$ or the information gain $IG(S)$ of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split or partitioned by the selected attribute to produce subsets of the data.

Pseudo Code:

ID3 (Examples, Target_Attribute, Attributes)

 Create a root node for the tree

 If all examples are positive, Return the single-node tree Root, with label = +.

 If all examples are negative, Return the single-node tree Root, with label = -.

 If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.

 Otherwise Begin

$A \leftarrow$ The Attribute that best classifies examples.

 Decision Tree attribute for Root = A .

 For each possible value, v_i , of A ,

 Add a new tree branch below Root, corresponding to the test $A = v_i$.

 Let Examples(v_i) be the subset of examples that have the value v_i for A

 If Examples(v_i) is empty

 Then below this new branch add a leaf node with label = most common target value in the examples

 Else below this new branch add the subtree ID3 (Examples(v_i), Target_Attribute, Attributes – { A })

 End

 Return Root

C4.5:

INTRODUCTION

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

ALGO:

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s(1), s(2), \dots$ of already classified samples. Each sample $s(i)$ consists of a p -dimensional vector $(x(1, i), x(2, i), \dots, x(p, i))$, where the $x(j)$ represent attribute values or features of the sample, as well as the class in which $s(i)$ falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the partitioned sublists.

Pseudo code:

In pseudocode, the general algorithm for building decision trees is:

1. Check for the above base cases.
2. For each attribute a , find the normalized information gain ratio from splitting on a .
3. Let a_{best} be the attribute with the highest normalized information gain.
4. Create a decision node that splits on a_{best} .
5. Recur on the sublists obtained by splitting on a_{best} , and add those nodes as children of node.

Naïve Bayes:

In machine learning we are often interested in selecting the best hypothesis (h) given data (d). In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).

One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.

Bayes' Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

Where:

$P(h|d)$ is the probability of hypothesis h given the data d . This is called the posterior probability.

$P(d|h)$ is the probability of data d given that the hypothesis h was true.

$P(h)$ is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h .

$P(d)$ is the probability of the data (regardless of the hypothesis).

You can see that we are interested in calculating the posterior probability of $P(h|d)$ from the prior probability $p(h)$ with $P(D)$ and $P(d|h)$.

After calculating the posterior probability for a number of different hypotheses, you can select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis.

This can be written as:

$$\text{MAP}(h) = \max(P(h|d))$$

or

$$\text{MAP}(h) = \max((P(d|h) * P(h)) / P(d))$$

or

$$\text{MAP}(h) = \max(P(d|h) * P(h))$$

The $P(d)$ is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize.

Back to classification, if we have an even number of instances in each class in our training data, then the probability of each class (e.g. $P(h)$) will be equal. Again, this would be a constant term in our equation and we could drop it so that we end up with:

$$\text{MAP}(h) = \max(P(d|h))$$

This is a useful exercise, because when reading up further on Naive Bayes you may see all of these forms of the theorem.

Algorithm:

Input: Training dataset T,

$F = (f_1, f_2, f_3, \dots, f_n)$ //value of the predictor variable in testing dataset.

Output:

A class of testing dataset.

Step:

1. Read the training set data T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat
Calculate the probability of f_i using the gauss density equation in each class;
Until the probability of all predictor variables ($f_1, f_2, f_3, \dots, f_n$) has been calculated.
4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

K-fold Cross validation

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

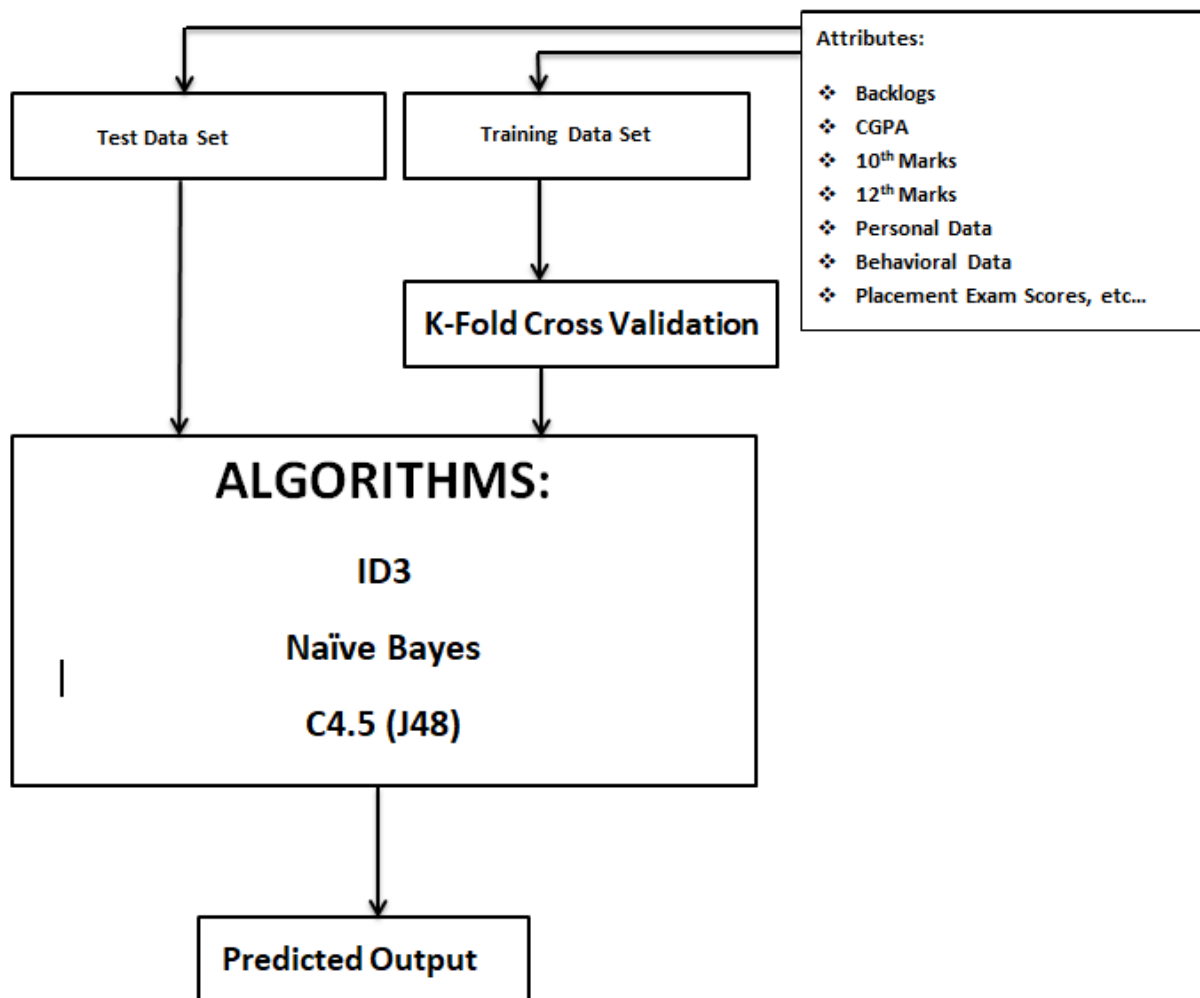
In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

For classification problems, one typically uses stratified k-fold cross-validation, in which the

folds are selected so that each fold contains roughly the same proportions of class labels.

In repeated cross-validation, the cross-validation procedure is repeated n times, yielding n random partitions of the original sample. The n results are again averaged (or otherwise combined) to produce a single estimation.

Data Flow Diagram



Gathering the Data Set:

We will be gathering information to train the model from various surveys, forms and also historical data provided by companies. As we need to train the model to high accuracy levels, we would be using a large amount of data sets.

Training and Testing the Model:

Once the data set has been obtained, the data set is divided at random into two different sets. One will be called as the training set and the other the testing set. The training data set is used to train the model by “teaching” the different type of outputs for the various inputs that can occur.

The testing data set is used after the model has been “trained” with the training data set. The values from the test data set are provided as input to the model. The outputs are recorded.

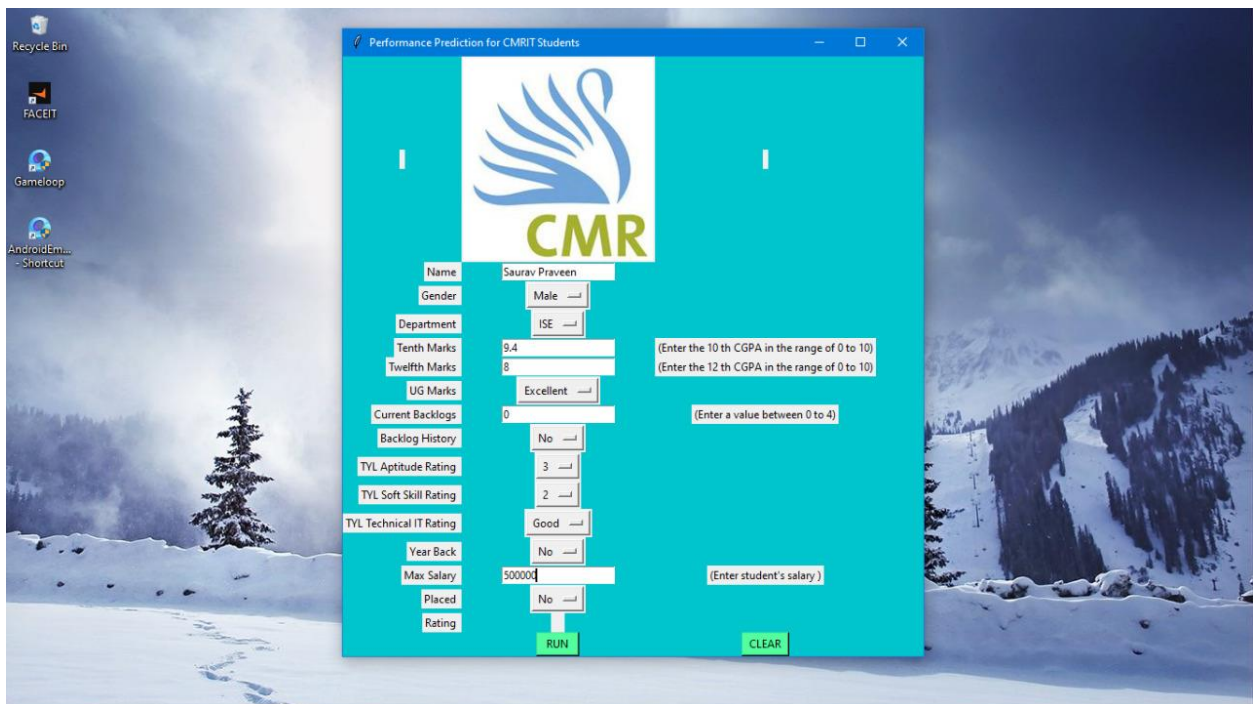
Calculating the accuracy of the results:

The outputs of the model after providing test data set values as input are recorded and then compared with the correct values in the test data set. By taking average of a large set of values, we can calculate the accuracy with which the application is able to predict the performance of the fresher.

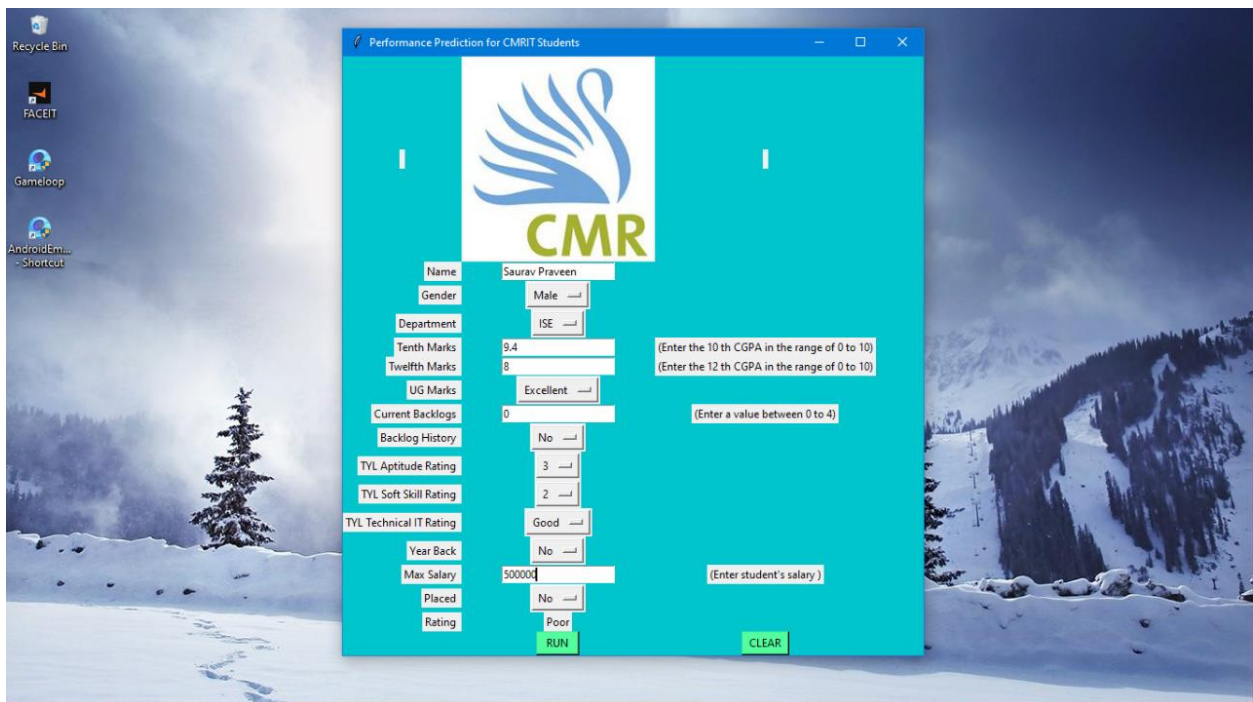
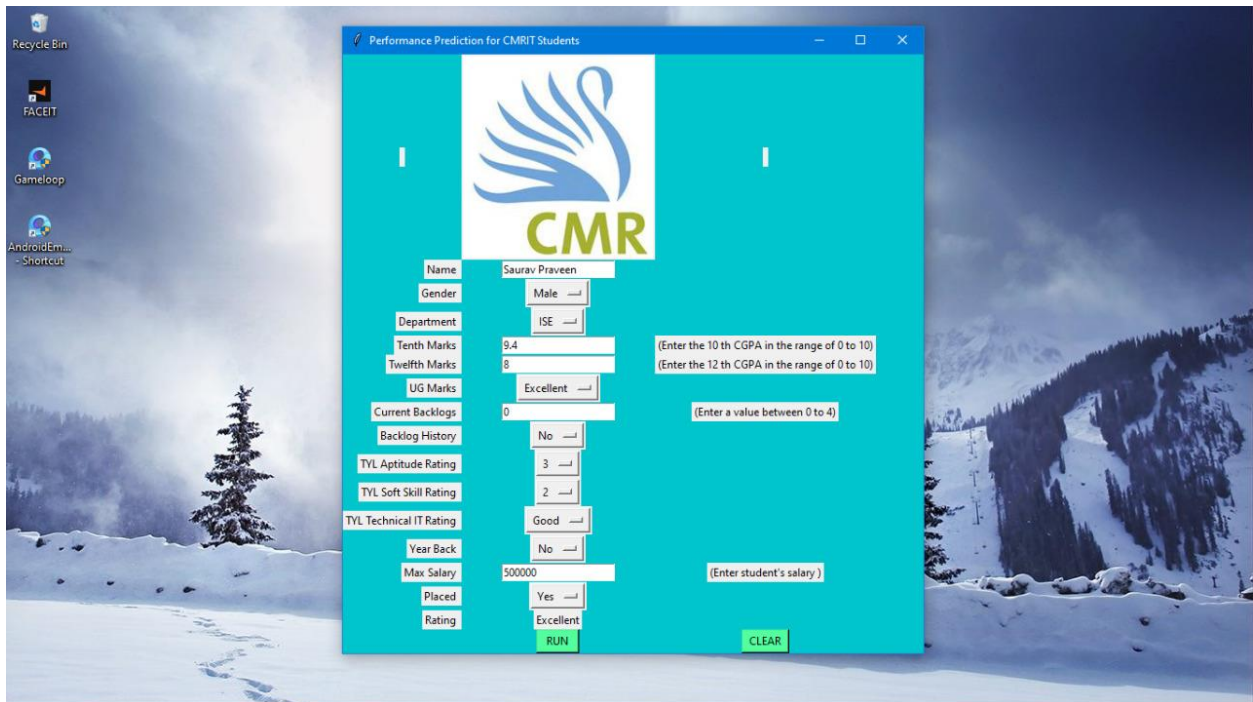
Chapter 5

IMPLEMENTATION

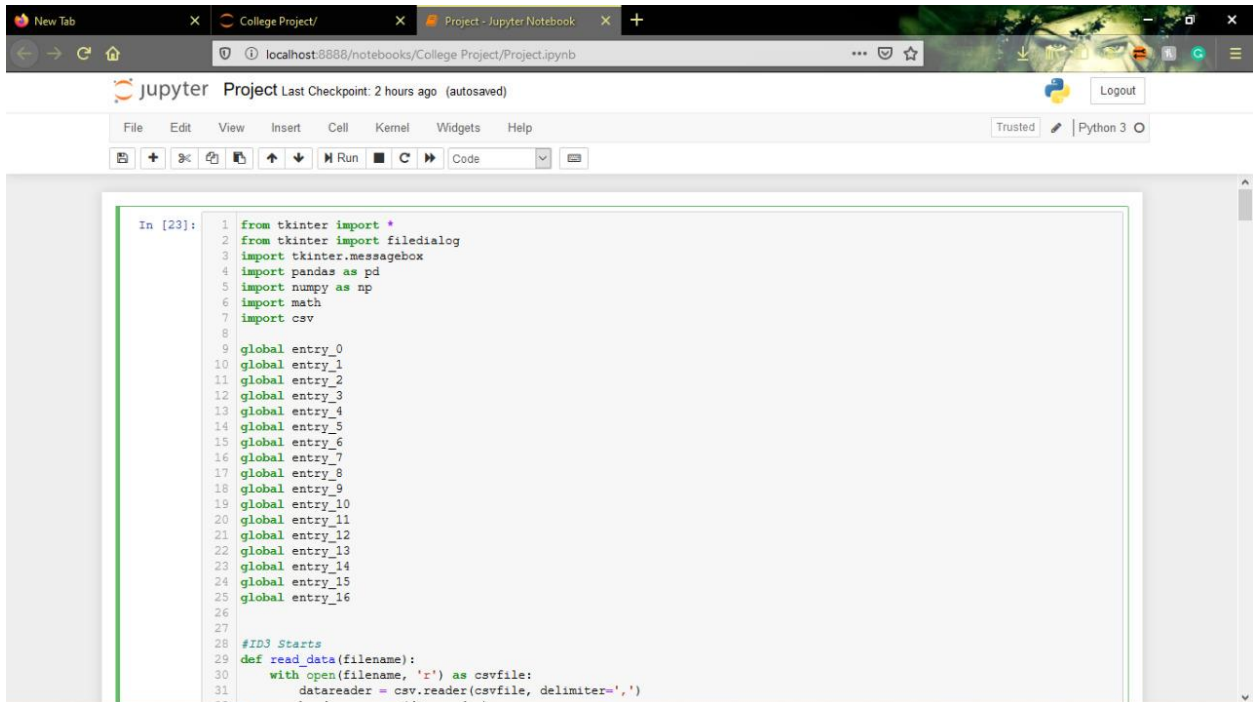
5.1 User Interface



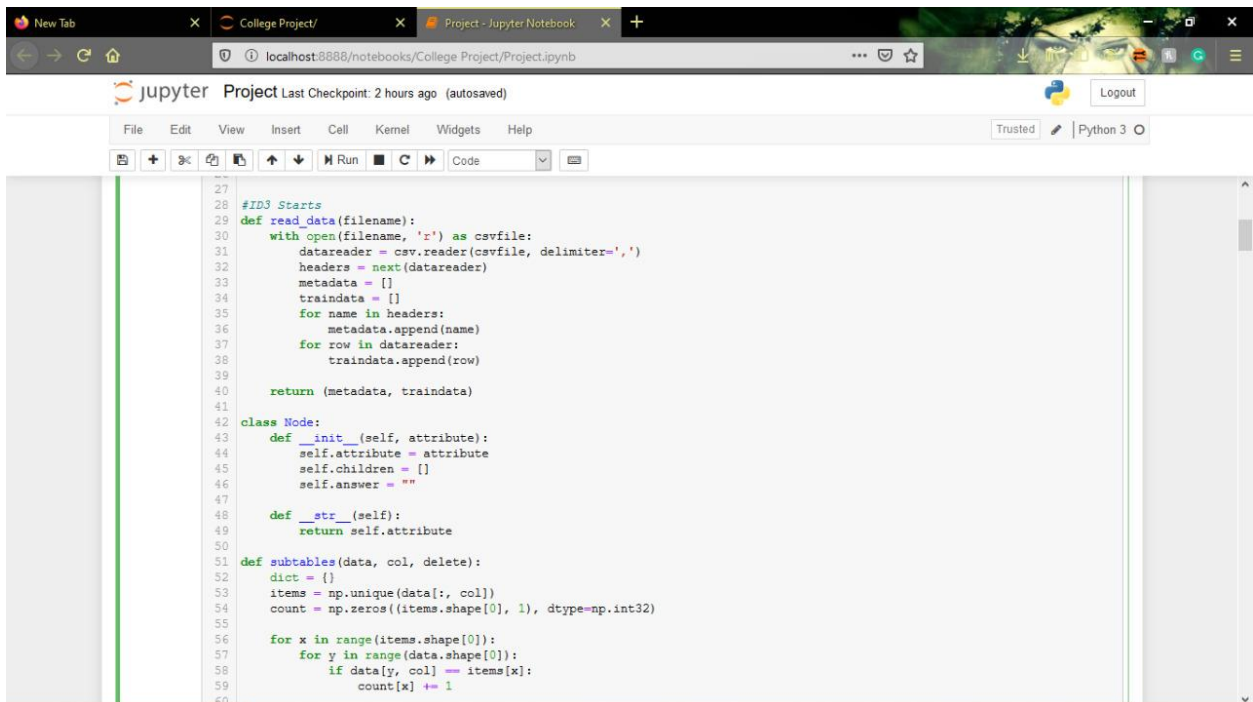
5.2 Output



5.3 Code



```
In [23]: 1 from tkinter import *
2 from tkinter import filedialog
3 import tkinter.messagebox
4 import pandas as pd
5 import numpy as np
6 import math
7 import csv
8
9 global entry_0
10 global entry_1
11 global entry_2
12 global entry_3
13 global entry_4
14 global entry_5
15 global entry_6
16 global entry_7
17 global entry_8
18 global entry_9
19 global entry_10
20 global entry_11
21 global entry_12
22 global entry_13
23 global entry_14
24 global entry_15
25 global entry_16
26
27
28 #ID3 Starts
29 def read_data(filename):
30     with open(filename, 'r') as csvfile:
31         datareader = csv.reader(csvfile, delimiter=',')
```



```
27
28 #ID3 Starts
29 def read_data(filename):
30     with open(filename, 'r') as csvfile:
31         datareader = csv.reader(csvfile, delimiter=',')
32         headers = next(datareader)
33         metadata = []
34         traindata = []
35         for name in headers:
36             metadata.append(name)
37         for row in datareader:
38             traindata.append(row)
39
40     return (metadata, traindata)
41
42 class Node:
43     def __init__(self, attribute):
44         self.attribute = attribute
45         self.children = []
46         self.answer = ""
47
48     def __str__(self):
49         return self.attribute
50
51 def subtables(data, col, delete):
52     dict = {}
53     items = np.unique(data[:, col])
54     count = np.zeros((items.shape[0], 1), dtype=np.int32)
55
56     for x in range(items.shape[0]):
57         for y in range(data.shape[0]):
58             if data[y, col] == items[x]:
59                 count[x] += 1
60
```

```

109 def create_node(data, metadata):
110     if (np.unique(data[:, -1])).shape[0] == 1:
111         node = Node("")
112         node.answer = np.unique(data[:, -1])[0]
113         return node
114
115     gains = np.zeros((data.shape[1] - 1, 1))
116
117     for col in range(data.shape[1] - 1):
118         gains[col] = gain_ratio(data, col)
119
120     split = np.argmax(gains)
121
122     node = Node(metadata[split])
123     metadata = np.delete(metadata, split, 0)
124
125     items, dict = subtables(data, split, delete=True)
126
127     for x in range(items.shape[0]):
128         child = create_node(dict[items[x]], metadata)
129         node.children.append((items[x], child))
130
131     return node
132
133 def empty(size):
134     s = ""
135     for x in range(size):
136         s += " "
137     return s
138
139 def print_tree(node, level):
140     if node.answer != "":
141         print(empty(level), node.answer)
142

```

```

288 def run_code():
289     np.seterr(divide='ignore', invalid='ignore')
290     metadata, traindata = read_data("Output6.csv")
291     data = np.array(traindata)
292     node = create_node(data, metadata)
293     #print_tree(node, 0)
294
295     Department = tkvar.get()
296     Tenth_Mark = entry_3.get()
297     Twelfth_Mark = entry_4.get()
298     Ug_Mark = tkvarUG.get()
299     Current_Backlogs = entry_6.get()
300     Backlog_History = tkvarBH.get()
301     TYL_Aptitude_Rating = str(tkvarR1.get())
302     TYL_Soft_Skills_Rating = str(tkvarR2.get())
303     TYL_Technical_IT_Rating = tkvarR3.get()
304     Year_Back = tkvarYB.get()
305     Max_Salary = entry_15.get()
306     Placed = tkvarP.get()
307
308     output = "None"
309
310     output = Work_Rating_Cal(Department, Tenth_Mark, Twelfth_Mark, Ug_Mark, Current_Backlogs, Backlog_History, TYL_Aptitude_Rat
311
312     if output == "None":
313         output = excpHandle(Department, Tenth_Mark, Twelfth_Mark, Ug_Mark, Current_Backlogs, Backlog_History, TYL_Aptitude_Rati
314
315     if output is None:
316         output = "Poor"
317
318     label_18.config(text=output)
319
320
321

```

Chapter 6

FUTURE SCOPE AND CONCLUSION

The application when run by a recruiter will have the values to attributes which the company has set according to their requirements. After the evaluation of the candidates in the hiring process has been completed, the values of the various attributes of the candidate and his performance evaluation are given to the application as inputs. The application takes these inputs and evaluates them against values provided by the recruiter based on their requirements. A score is awarded to the candidate. The score would range from 0 to 100. The score of the candidate indicates the rating of the candidate i.e. how well suited that particular candidate is for the role the recruiter is looking to hire for. Every candidate in the hiring process is awarded a score. The candidates with the highest scores can be recruited. This makes the job of the recruiters easier.

As a conclusion, we have met our objective. It is difficult to find a suitable talented personal to fit the job needed, normal hiring process is time-consuming and bad hiring can cause a financial loss. Moving to the digital evolution is a trend in the hiring process; a number of solutions such as INDEED, CAREER, Google for jobs and others or a 3rd party agencies such as ARYA's solutions are now in the market using the artificial intelligence (AI) techniques to help to find the correct choice of candidates to fulfill the needed roles. Still, the recruiters are depending on attitude hiring more than skills hiring. This paper shows that the more attributes considered in the calculation of predicting more prediction accuracy, although not much research has been worked on this idea clearly, so we worked in this model to narrow the gap between skills and attitude hiring by predicting the performance of a new candidate in multiple areas, areas that hard to be decided in the regular hiring process. In order to have the most accurate approximation for a new candidate performance, a huge amount of data needs to be processed. More the evaluations history and employee statuses more reasonable approximation outputted. In addition to the variance of employees natures and skills. With the deep search in the machine learning algorithms, and more about applying

the Multilayers Machine learning algorithms in prediction algorithms, this approach considered as future work to be implemented instead of decision trees. The best algorithm based on the placement data is Naïve Bayes Classification with an accuracy of 86.15% and the total time taken to build the model is at 0 seconds. Naïve Bayes classifier has the lowest average error at 0.28 compared to others. These results suggest that among the machine learning algorithm tested, Naïve Bayes classifier has the potential to significantly improve the conventional classification methods for use in placements.

REFERENCES

- [1] Transforming Human Resources into Human Capital, S.Zakaria,W.Yusoff, Information Management and Business review, Vol 2,No2,PP 48-54,Feb 2011
- [2] a survey by Harris Poll on behalf of CareerBuilder between August 16 and September 15, 2017 <http://press.careerbuilder.com/2017-1207-Nearly-Three-in-Four-Employers-Affected-by-a-Bad-HireAccording-to-a-Recent-CareerBuilder-Survey>
- [3] A Study on the Effectiveness of Performance Appraisal System and its Influence with the Socio-Demographic Factors of the Employees of a Manufacturing Industry in Tamil Nadu, International Journal of Research in Management & Business Studies (IJRMBS 2015), Vol. 2 Issue 1 Jan. - Mar. 2015
- [4] RECRUITMENT THROUGH ARTIFICIAL INTELLIGENCE: A CONCEPTUAL STUDY, International Journal of Mechanical Engineering and Technology (IJMET), Volume 9, Issue 7, July 2018, pp. 63–70, Article ID: IJMET_09_07_007
- [5] https://cdn2.hubspot.net/hubfs/1951043/Content%20Offers/Ai_Whitepaper/Artisan%20Talent%20AI%20and%20Bots%20in%20Recruiting%20Paper.pdf
- [6] Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance,Q.Alradaideh, Eman.Alnagi, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012
- [7] <https://ideal.com/top-recruiting-software/>
- [8] <https://goarya.com/solutions/>
- [9] The new age: artificial intelligence for human resource opportunities and functions, EY 2018

[https://www.ey.com/Publication/vwLUAssets/EY-the-new-ageartificial-intelligence-for-human-resource-opportunities-andfunctions/\\$FILE/EY-the-new-age-artificial-intelligence-for-humanresource-opportunities-and-functions.pdf](https://www.ey.com/Publication/vwLUAssets/EY-the-new-ageartificial-intelligence-for-human-resource-opportunities-andfunctions/$FILE/EY-the-new-age-artificial-intelligence-for-humanresource-opportunities-and-functions.pdf)

[10] Role of Artificial Intelligence in Recruitment, Anjana Raviprolu, International Journal of Engineering Technology, Management and Applied Sciences, April 2017, Volume 5 Issue 4, ISSN 2349-4476

[11] Effectiveness of Performance Appraisal System and its Effect on Employee Motivation, Idowu, Ayomikun O., Nile Journal of Business and Economics, (2017) 5: 15-39

[12] The History of Artificial Intelligence, Chris Smith, Brian McGuire, Ting Huang, Gary Yang, University of Washington December 2006

[13] Introduction To Machine Learning, 2nd edition , Ethem Alpaydın, The MIT Press Cambridge,2010

[14] <https://ieeexplore.ieee.org/document/7457238> (Hirability in the Wild: Analysis of Online Conversational Video Resumes)

[15] Should human resource managers use social media to screen job applicants? Managerial and legal issues in the USA - <https://www.emeraldinsight.com/doi/abs/10.1108/14636691211196941>

[16] The Writing on the (Facebook) Wall: The Use of Social Networking Sites in Hiring Decisions - <https://link.springer.com/article/10.1007/s10869-011-9221-x> [17] Dataset taken from “Network International” company in Jordan