**Symbiosis Skills and Professional University Kiwale, Pune**

------------------------------------------------------------------------------------------

**PROJECT REPORT**

**On**

**“Campus Recruitment Data Analysis”**



**Submitted by**

**Manjusha Hanumant Ghanwat**

**(2001207052)**

**DA-Batch-I**

**Under The Guidance of**

**Trainer's Name: Prof. Kushal Sharma**

**STUDENT DECLARATION AND ATTESTATION BY TRAINER**

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

Signature of student

**Name of student: Manjusha Hanumant Ghanwat**

Registration Number: 2001207052

Signature of trainer

**Name of trainer: Prof. Kushal Sharma**

**CERTIFICATE**

This is to certify that the report entitled, “**Campus Recruitment Data Analysis”** submitted by “**Manjusha Hanumant Ghanwat”** to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a record of bonafide Project work carried out by him under my supervision and guidance and is worthy of consideration for the completion of certificate course in “Data Associate”.

Signature of Trainer

Name of Trainer

Date: / / 2021

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor Supervisor

Date:

**ACKNOWLEDGEMENTS**

Placement of students is one of the most important objectives of an educational institution. Reputation and yearly admissions of an institution invariably depend on the placements it provides it students with. That is why all the institutions, arduously, strive to strengthen their placement department so as to improve their institution on a whole. Any assistance in this particular area will have a positive impact on an institution’s ability to place its students. This will always be helpful to both the students, as well as the institution. In this study, the objective is to analyse previous year's student's data and use it to predict the placement chance of the current students. This model is proposed with an algorithm to predict the same. Data pertaining to the study were collected form the same institution for which the placement prediction is done, and also suitable data pre-processing methods were applied. This proposed model is also compared with other traditional classification algorithms such as KNN, SVM and Logistics Regression with respect to accuracy, precision and recall. From the results obtained it is found that the proposed algorithm performs significantly better in comparison with the other algorithms mentioned. Also, Analysis by Tableau software for better prediction and visualisation.

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No** | **Index** | | **Page No.** |
| 1 | **Acknowledgement** | |  |
| 2 | **Plan of Capstone Project** | |  |
|  | 2.1 | Purpose of Project | 1 |
|  | 2.2 | Period of Project | 1 |
|  | 2.3 | Problem Statement Detailing | 1 |
| 3 | **Objective of the** Project | |  |
|  | 3.1 | General Objective | 2 |
| 4 | **Introduction** | | 3 |
| 5 | **Methodology** | | 4 |
|  | **5.1** | Data Gathering | 4 |
|  | **5.2** | Pre-Processing | 4 |
|  | 5.2.1 | Cleaning missing values | 5 |
|  | 5.2.2 | Training and test data | 5-6 |
|  | 5.2.3 | Feature scaling | 6 |
|  | **5.3** | Processing | 6 |
|  | 5.3.1 | KNN Algorithm | 6-7 |
|  | 5.3.2 | SVM Algorithm | 7-8 |
|  | 5.3.3 | Logistics Regression Algorithm | 10 |
| 6 | Future Plans | | 12 |
| 7 | Learning from the Project | | 12 |
| 8 | Annexure | | 13 |
| 9 | Weekly Reports | | 14 |
| 10 | Faculty and Student Feedback | | 15 |

1. **Plan of Capstone Project**
   1. **Purpose** **of Project:**

This data set consists of Placement data of students in our campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students.

The main objective of this model is to predict whether the student he/she gets placed or not in campus recruitment. For this the data consider is the academic history of student like overall percentage, backlogs, credits. The algorithms are applied on the previous year’s data of the students.

* 1. **Period of Project: 1 month**
  2. **Problem Statement**

Can we predict that, a candidate was placed on a role after their MBA studies? If so, which factors helped the most (i.e., work experience, degree, school results, gender, etc)?

Questions:

1. Which factor influenced a candidate in getting placed?
2. Does percentage matters for one to get placed?
3. Which degree specialization is much demanded by corporate?
4. Play with the data conducting all statistical tests.
5. **Objective of the Project**

The main objective of this model is to predict whether the student he/she gets placed or not in campus recruitment. For this the data consider is the academic history of student like overall percentage, backlogs, credits. The algorithms are applied on the previous year’s data of the students. This proposed model is also compared with other traditional classification algorithms such as KNN, SVM and Logistics Regression with respect to accuracy, precision and recall. From the results obtained it is found that the proposed algorithm performs significantly better in comparison with the other algorithms mentioned. Also, Analysis by Tableau software for better prediction and visualisation.

1. **Introduction**

Placements are considered to be very important for each and every college. The basic success of the college is measured by the campus placement of the students. Every student takes admission to the colleges by seeing the percentage of placements in the college. Hence, in this regard the approach is about the prediction and analyses for the placement necessity in the colleges that helps to build the colleges as well as students to improve their placements.

In Placement Prediction system predicts the probability of an undergrad students getting placed in a company by applying classification algorithms such as KNN, SVM and Logistics Regression. The main objective of this model is to predict whether the student he/she gets placed or not in campus recruitment. For this the data consider is the academic history of student like overall percentage, backlogs, credits. The algorithms are applied on the previous year’s data of the students. This data set consists of Placement data of students in our campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students. By Applying Machine Learning Algorithm on this dataset then there is possibility of prediction that which candidate is selected for campus recruitment process and which is the major factor is helped him or her for placement.

1. **METHODOLOGY:**

The whole approach is depicted by the following flowchart

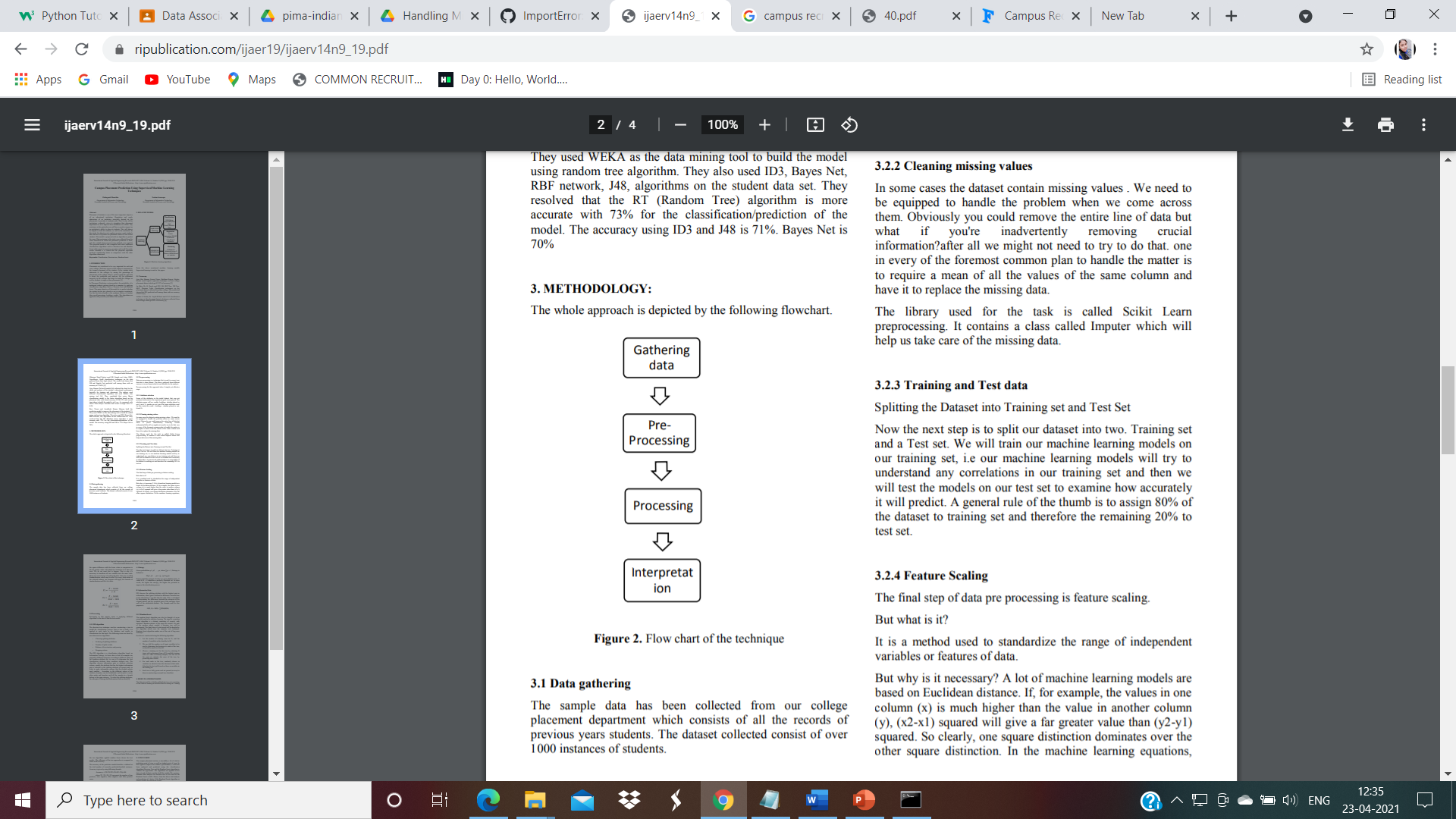


Fig. Flow chart of the technique

1. Data gathering

The sample data has been collected from our college placement department which consists of all the records of previous years students. The dataset collected consist of over 250 instances of students.

1. Pre processing

Data pre-processing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Pre-processing for this approach takes 4 simple yet effective steps.

* 1. Attribute selection

Some of the attributes in the initial dataset that was not pertinent (relevant) to the experiment goal were ignored. The attributes name, roll no, credits, backlogs, whether placed or not, b.tech % , gender are not used. The main attributes used for this study are credit , backlogs , whether placed or not, b.tech %.

**2.2** Cleaning missing values

In some cases, the dataset contains missing values. We need to be equipped to handle the problem when we come across them. Obviously, you could remove the entire line of data but what if you're inadvertently removing crucial information? after all we might not need to try to do that. one in every of the foremost common plan to handle the matter is to require a mean of all the values of the same column and have it to replace the missing data. The library used for the task is called Scikit Learn pre-processing. It contains a class called Imputer which will help us take care of the missing data.

* 1. Training and Test data

Splitting the Dataset into Training set and Test Set Now the next step is to split our dataset into two. Training set and a Test set. We will train our machine learning models on our training set, i.e., our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to training set and therefore the remaining 20% to test set.

**2.4** Feature Scaling

The final step of data pre-processing is feature scaling.

But what is it?

It is a method used to standardize the range of independent variables or features of data. But why is it necessary? A lot of machine learning models are based on Euclidean distance. If, for example, the values in one column (x) are much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared. So clearly, one square distinction dominates over the other square distinction. In the machine learning equations, the square difference with the lower value in comparison to the far greater value will almost be treated as if it does not exist. We do not want that to happen. That is why it’s necessary to transform all our variables into the same scale. There are several ways of scaling the data. One way is called Standardization which may be used.

1. Processing

Processing in this paper's sense is applying different algorithms to the data to find the best results.

* 1. **K-NN Algorithm**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1:** Select the number K of the neighbours
* **Step-2:** Calculate the Euclidean distance of **K number of neighbours**
* **Step-3:** Take the K nearest neighbours as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbours, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbour is maximum.
* **Step-6:** Our model is ready.
  1. **Support Vector Machine (SVM**)

SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

SVM can be of two types:

* **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

## **How does SVM works?**

**Linear SVM:**

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair (x1, x2) of coordinates in either green or blue. Consider the below image:



So, as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:



Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The **hyperplane** with maximum margin is called the **optimal hyperplane**.



**Non-Linear SVM:**

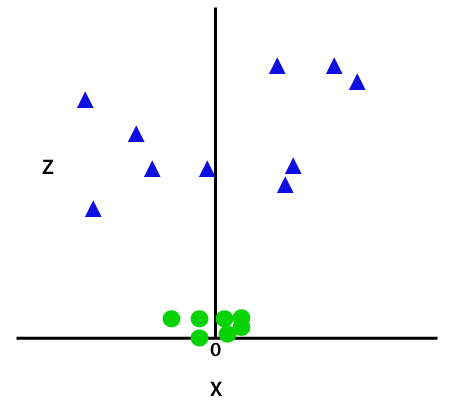
If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:



So, to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third-dimension z. It can be calculated as:

Z = x2 + y2

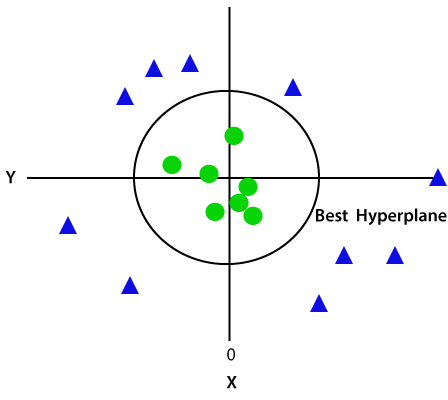
By adding the third dimension, the sample space will become as below image:



So now, SVM will divide the datasets into classes in the following way. Consider the below image:



Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:



Hence, we get a circumference of radius 1 in case of non-linear data.

* 1. **Logistics Regression**

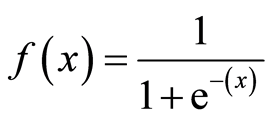
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Function (Sigmoid Function):

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.



Where,

F(x) = output between the 0 and 1 value

x = input to the function

e = base of natural logarithm

1. **Future plans :**

The Result that we got aren't very great at the moment. We were able to predict whether a person will be placed or not using the data in the table but the results aren't very great. It could be better by including some other feature or by reducing the number of features. This model still has some scope for improvement. But we could see that Logistic Regression performed better than both KNN and SVM. So at least we have an idea which model would give us the better results. We can also apply another machine learning algorithm for better accuracy.

1. **Learning from the Project :**

The campus placement activity is incredibly a lot of vital as institution point of view as well as student point of view. In this regard to improve the student’s performance, a work has been analysed and predicted using the classification algorithms KNN, SVM and the Logistics Regression algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for KNN is 74.41%, for the SVM is 79.06% and for the Logistics Regression is 88.37%. Hence, from the above said analysis and prediction it’s better if the Logistics Regression algorithm is used to predict the placement results.

1. **Annexure**

* [**https://www.w3schools.com/python/default.asp**](https://www.w3schools.com/python/default.asp)
* [**https://www.javatpoint.com/logistic-regression-in-machine-learning**](https://www.javatpoint.com/logistic-regression-in-machine-learning)
* [**https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning**](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)
* [**https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm**](https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm)
* [**https://www.kaggle.com/benroshan/factors-affecting-campus-placement**](https://www.kaggle.com/benroshan/factors-affecting-campus-placement)
* <https://www.ripublication.com/ijaer19/ijaerv14n9_19.pdf>
* [**https://www.javatpoint.com/tableau**](https://www.javatpoint.com/tableau)

1. **Weekly Reports**

|  |  |
| --- | --- |
| Week | Work done |
| 1 | Selecting Topic for project  Gathering Data  Pre-processing on Dataset. |
| 2 | Cleaning Missing Values from Dataset  Python code for Analysis of dataset |
| 3 | Used Machine Learning Algorithm for better Accuracy   1. KNN Algorithm 2. SVM Algorithm 3. Logistics Regression Algorithm |
| 4 | Also implement in Tableau Software for Data visualisation purposed  Project Report |