**PROFESSIONAL TRAINING REPORT**

**at**

**Sathyabama Institute of Science and Technology (Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

**Karumanchi.divya sruthi**

**(REG. NO. 39110468)**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF COMPUTING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**

**CHENNAI – 600119, TAMILNADU**

**APRIL 2022**

|  |  |  |
| --- | --- | --- |
|  | **SATHYABAMA**  **INSTITUTE OF SCIENCE AND TECHNOLOGY** (DEEMED TO BE UNIVERSITY) **Accredited with Grade “A” by NAAC**  (Established under Section 3 of UGC Act, 1956)  JEPPIAAR NAGAR, RAJIV GANDHI SALAI  CHENNAI– 600119  [**www.sathyabama.ac.in**](http://www.sathyabama.ac.in) |  |

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **k.divya sruthi(Reg. No: 39110468)** who carried out the project entitled “**Attrition analysis and prediction in an IT organization**” under my supervision from December 2022 to April 2021.

**Internal Guide**

**MS.SREE Krishna**

**Head of the Department**

**Dr.S.VIGNESHWARI, M.E., ph.D.,**

**Dr.L.LAKSHMANAN, M.E., ph.D.,**



**Submitted for Viva voce Examination held on**

**InternalExaminer ExternalExaminer**

# DECLARATION

I, **k.divya Sruthi** hereby declare that the project report entitled **Attrition analysis and prediction in an It organization** done by me under the guidance of **MS.Sree Krishna** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

|  |  |
| --- | --- |
| **DATE:** 09-04-2022 | k.divya sruthi |
| **PLACE**: CHENNAI | **SIGNATURE OF THE CANDIDATE** |

# ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D**, **Dean**, School of Computing, **Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **MS.Sree Krishna,** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

# Table of Contents

[Section 1: ABSTRACT 1](#_Toc100413129)

[Section 2: INTRODUCTION 2](#_Toc100413130)

[Section 3: DATASET 3](#_Toc100413131)

[Section 3(i) : Screenshot of DataSet 3](#_Toc100413132)

[Section 4: EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHMS USED 4](#_Toc100413134)

[Section 4 : METHODOLOGY 4](#_Toc100413135)

[Section 4(i) : Naïve Bayes 5](#_Toc100413136)

[Section 4(ii): Logistic Regression 6](#_Toc100413137)

[Section 4(iii): Multi-layer perceptron(MLP)classifier 6](#_Toc100413138)

Section 4(iv)K'-Nearest Neighbors(KNN)……………………………………………………6

[ALGORITHM 7](#_Toc100413139)

[Section 5 : Summary of Logistic Regression 7](#_Toc100413140)

[Section 6: Summary of Decision tree learning 8](#_Toc100413141)

Section 7:Summary of Random Forest…………………………………………………………8

[Section 8: RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS 9](#_Toc100413143)

[Section 8(i): RESULT 9](#_Toc100413144)

[Section 8(ii): OBSERVATIONS AND CONCLUSIONS 9](#_Toc100413145)

[Section 9: REFERENCES 9](#_Toc100413146)

[APPENDIX 10](#_Toc100413147)

[Section 10: SOURCE CODE 10](#_Toc100413148)

# Section 1: ABSTRACT

The project aim is to finding the attrition analysis and prediction in an It organization. In this I used linear regression model to predict the value.

In today world many industries and specially it(information technology) are experiencing high attrition rate .managers and HR department are aware of some common causes of attrition in their organization

The impact of employee leaving voluntary is not good for organization or to project in which they are working.hence HR and senior managers and the policy makers of any industry are working together to reduce this voluntary exit

A good leader senses and understands employee needs and work with them and HR to fix the issues.however not all attrition causes are know to managers and when it actually happens it turns out as a surprise to managers then they are not able to do much.

In-spite of many efforts taken attrition is still a problem in many industries some amount of attrition is certain and bound to happen like employee retiring or death of employee hence the scope of this work is only restricted to voluntary exit

On many occasion HR department has felt that if they would have known earlier,or they could have picked the sign of exit,they might have prevented good employes leaving

These tools not only predict but also show some clear pattern in attrition many organaziation today use cots attrition prediction tool or build their own-in -house prediction tools the scope of this work is to discuss clearly attrition,its impact

# Section 2: INTRODUCTION

Employee resignations are a reality for any business.

However, if the situation isn't handled properly, key staff

members' departures can lead to a downturn in productivity. The

organization may have to employ new people and train them on

the tool that is being used, which is time consuming.

Most organizations are interested in knowing which of their

employees are at the risk of leaving.

This paper discusses the application of the k-Nearest

Neighbours (KNN) algorithm as a method of predicting

employee attrition.

This is done by using data from Kaggle and

treating the problem as a classification task. The conclusion is

reached by comparing the performance of the KNN classifier

against other techniques.

This paper is structured as follows. Section II discusses the

attrition problem, and lists the work done by others using

machine learning algorithms to solve the problem. Section III

explores 4 different machine learning algorithms, including

KNN, that this paper compares

Section IV outlines the

experimental method employed in terms of the features used,

pre-processing, and the metrics used to compare the algorithms.

Section V presents the results of the comparison and a discussion

of the same, and possible future work. Section VI concludes the

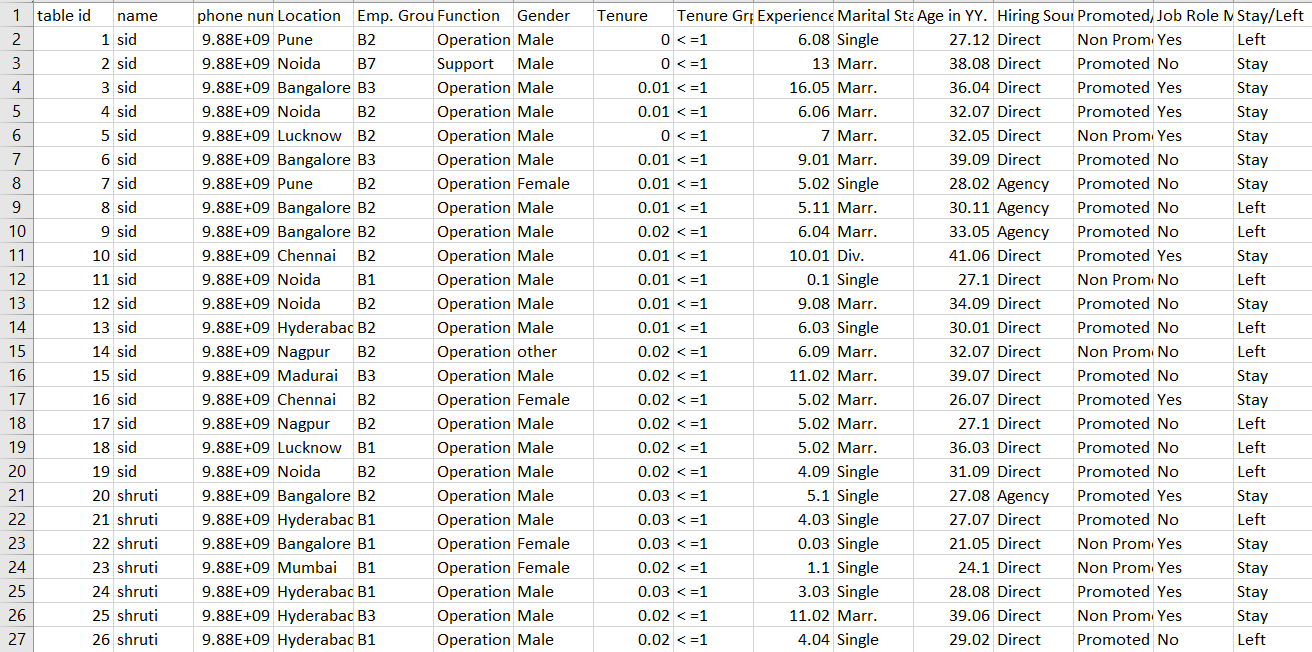
paper by recommending the KNN classifier as an approach to

solving the employee attrition prediction problem.

# Section 3: DATASET

# Hence to proceed further we have taken a Dataset of Employee Details like phone number,Location,Gender,Experience,Marital Status,Age……etc and we go through the given Dataset and in the Dataset we have many colums we select some random Data from the whole Dataset by using machine learning as shown in code snapshots.\

## Section 3(i) : Screenshot of DataSet



# Section 4: EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHMS USED

## Section 4: METHODOLOGY

This paper discusses supervised learning methods of

classification, since we know of the existence of two classes—

working and left.

This section outlines the theory behind each

machine learning algorithm.

## Section 4(i):

**Naive Bayes :**

Naive Bayes is a classification technique that has gained

popularity due to its simplicity . The Naive Bayes algorithm

makes use of the assumption that all the variables are

independent of each other, and then calculates probabilities, that

are used for classification.

The algorithm works as follows: to get an output function Y

given a set of input variables X, the algorithm estimates the

values of P(X|Y) and P(Y), and then uses Bayes’ rule to compute

P(Y|X), which is the required output, for each of the new samples.

In this paper, we use the Gaussian Naive Bayes algorithm, which assumes that the values associated with each class are distributed according to a Gaussian distribution.

## Section 4(ii):

**Logistic Regression:**

Logistic regression is a regression model that fits the to the logistic function. It is useful when the dependent is categorical . The general form of the model is

values variable



Logistic regression is often used with regularization techniques to prevent overfitting. An L2 regularized model .

## Section 4(iii)

## Multi-layer Perceptron:

MLP is an artificial neural network (ANN) model that consists of multiple layers of nodes, each fully connected to the next. The algorithm uses backpropagation for training the modell6). The input is transformed using a learned non-linear transformation, which projects the input data into a space where it becomes linearly separable. This intermediate layer is called a hidden layer.

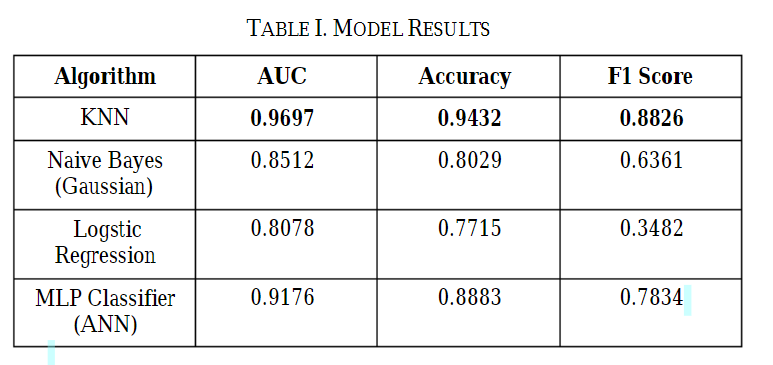
## Section 4(iv)

**K-Nearest Neighbors (KNN):**

The KNN algorithm classifies new data based on the class of the k nearest neighbors. This paper uses the value of k as 6. The distance from neighbors can be calculated using various distance metrics, such as Euclidean distance, Manhattan distance (used in this paper), Minkowski distance, etc. The class of the new data may be decided by majority vote or by an inverse proportion to the distance computed. KNN is a non-generalizing method, since the algorithm keeps all of its training data in memory, possibly transformed into a fast indexing structure such as a ball KD tree

The Manhattan distance is computed using the formula





**ALGORITHM**

***Section 5:* Summary of** [**Logistic regression**](https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/logistic-regression/)

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

# Section 6: Summary of decision tree learning:

**Decision tree learning** or **induction of decision trees** is one of the predictive modelling approaches used in [statistics](https://en.m.wikipedia.org/wiki/Statistics), [data mining](https://en.m.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.m.wikipedia.org/wiki/Machine_learning). It uses a [decision tree](https://en.m.wikipedia.org/wiki/Decision_tree) (as a [predictive model](https://en.m.wikipedia.org/wiki/Predictive_modelling)) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called [**classification**](https://en.m.wikipedia.org/wiki/Classification)[**trees**](https://en.m.wikipedia.org/wiki/Decision_tree); in these tree structures, [leaves](https://en.m.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.m.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically [real numbers](https://en.m.wikipedia.org/wiki/Real_numbers)) are called [**regression**](https://en.m.wikipedia.org/wiki/Regression_analysis)[**trees**](https://en.m.wikipedia.org/wiki/Decision_tree). Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

**Section 7:Summary of random forest**

Random forest is a ***Supervised Machine Learning Algorithm*** that is ***used widely in Classification and Regression problems***. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing ***continuous variables*** as in the case of regression and ***categorical variables*** as in the case of classification. It performs better results for classification problems.

# Section 8: RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

## Section 8(i): RESULT

## The output depends on the chosen model. For instance, ‘logistic model’ produces scorecards for employees based on their predicted ‘attrition risk’ parameters; while the classification model catalogues the employees into wider parameters, such as-more likely or less likely to quit, high risk or low risk, etc.

## Section 8(ii): OBSERVATIONS AND CONCLUSIONS

## Predictive Attrition Model helps in not only taking preventive measures but also into making better hiring decisions. Deriving trends in the candidate’s

## performance out of past data is important in order to predict the future trends, as well as to board new employees. Moreover, HR can use the employee data to predict attrition, the possible reasons behind it and can take appropriate measures to prevent it.

The results of this research showed the superiority of the KNN classifier in terms of accuracy and predictive effectiveness, by means of the ROC curve. When used with its optimal configuration, it is a robust method that delivers accurate results in spite of the noise in the dataset, which is a major challenge for machine learning algorithms. The authors thus

recommend the use of the KNN classifier for accurately

predicting employee attrition in an organization, which enables

# Section 9: REFERENCES

* FLASK - From Python
* DATASET - From github
* Machine Learning,Python - FROM Github,kaggle

# APPENDIX

# Section 10 : SOURCE CODE

## 

## Section A(i): SCREENSHOTS

