

Predicting Employee Attrition using various ML Algorithms

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

In

Computer Science and Engineering

by

A Sai Kaushik 16BCE0527

P.D Sai Vardhan 16BCE0459

T.S Balagopal 16BCE2226

Under the guidance of

Dr. J.Sairabanu

SCOPE VIT,

Vellore.



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

May, 2020

DECLARATION

I hereby declare that the thesis entitled “Predicting Employee Attrition using various ML algorithms” submitted by me, for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** to VIT is a record of bona-fide work carried out by me under the supervision of J. Sairabanu.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

Date :

X

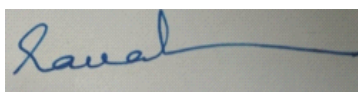
Sai Kaushik, Sai Vardhan, Balagopal

Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled “Predicting Employee Attrition using various ML algorithms” submitted by **A Sai Kaushik (16BCE0527), P.D Sai Vardhan(16BCE0459), Balagopal T.S(16BCE2226)**, VIT, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering*, is a record of bonafide work carried out by him / her under my supervision during the period, 01. 12. 2019 to 30.04.2020, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

X 

Saira Banu. J

Signed by: 1fb42809-0b6a-49cd-b7b5-56bc08070739

Place : Vellore

Date :

Signature of the Guide

Internal Examiner

External Examiner

V.Santhi

B.Tech – SCOPE

ACKNOWLEDGEMENTS

We are greatly indebted to our professors and students for their constant support in pointing us in the right directions and steer us from obstacles. We are especially grateful for the SCOPE department of the Vellore Institute of Technology, Vellore to allow us to work on our Capstone Project and present it during the widespread COVID-19 pandemic.

This project was done thanks to the expert tutelage of our professor, J.Sairabanu from the SCOPE department in the Vellore Institute of Technology, Vellore. A major thank you to the representatives of the different publications and websites from which research knowledge has been gathered.

Student Name

A Sai Kaushik	16BCE0527
P.D Sai Vardhan	16BCE0459

Executive Summary

Employee Attrition can be defined as the loss of talented employees in a company. This can occur due to various reasons like low-pay, working environment, goals of the company and the tasks given to the employees. This is a major problem for companies aiming for rapid growth, so we have tried to provide and suggest the best algorithm that predicts possible attrition and helps the company to avoid it.

For this project we have taken a sample IBM dataset, the dataset was initially imbalanced, so we balanced it using SMOTE- Synthetic Minority Over-sampling Technique and Random Sampling techniques individually. We then trained the balanced data with algorithms like Random-Forest, Artificial Neural Networks, K-Nearest Neighbors in each case. Thus the outputs for 6 cases of algorithm combinations have been generated and compared and the best possible one with the highest accuracy will be suggested for the use of the company.

	CONTENTS	Page No.
	Acknowledgement	2
	Executive Summary	3
	Table of Contents	4
	List of Figures	6
	List of Tables	7
	Abbreviations	8
	Symbols and Notations	9
1	INTRODUCTION	10
	Objective	10
	Motivation	11
	Background	11
2	Literature Survey	17
3	PROJECT DESCRIPTION TECHNICAL SPECIFICATION	17
4	DESIGN APPROACH AND DETAILS	18
	Design Approach / Materials & Methods	19
	Codes and Standards	20
	Constraints, Alternatives and Tradeoffs	.
5	SCHEDULE, TASKS AND MILESTONES	.
6	PROJECT DEMONSTRATION	.
7	COST ANALYSIS / RESULT & DISCUSSION	.

8	SUMMARY AND RESULTS	.
9	REFERENCES	.
	APPENDIX A	.

List of Figures

Figure No.	Title	Page No.
1	SVM model	9
2	KNN model	10
3	Project Basic Architecture	11
4	Classification Flowchart	12
5	The Flow of project in sequential steps	12
6	Final Results	23
7	Random Forest trees	19
8	ANN propagation	20

List of Tables

Table No.	Title	Page No.
1	Literature Review Analysis	2

List of Abbreviations

ML	Machine Learning
KNN	K – Nearest Neighbors
RF	Random Forest
SVM	Support Vector Machine
ANN	Artificial-Neural-Networks

1) INTRODUCTION

1.1. OBJECTIVE

The first task at hand would be to balance the imbalanced sets of data using SMOTE – Synthetic Minority Over Sampling Technique or Random Sampling techniques and then the balanced set will be trained using KNN, Random Forest, ANN algorithms. The machines will be trained in combination of the balancing techniques and the ML algorithms and then tested for accuracy of their prediction and the best combination will be suggested for real time use of a company

1.2. THEORETICAL BACKGROUND

Employee attrition can be characterized as the loss of employees due to any of the following reasons: individual reasons, low mental fulfillment, low pay, business environment, unethical practices by the company. Employee attrition can be sorted into two types: intentional and automatic attrition.

Automatic attrition happens when employees are fired by their manager for various reasons like , low employee output or business prerequisites. In deliberate attrition, then again, high-performing employees opt to leave the organization independently and do not withstand the organization's endeavors to hold them.

Intentional attrition can result from early retirement or employment propositions from different firms, for instance. Despite the fact that organizations understand the significance of their employees ordinarily put resources into their workforce by giving significant preparation and an incredible working condition, they also experience the ill effects of willful attrition and the loss of gifted employees.

Another issue, recruiting substitutions, forces significant expenses on the organization, including the expense of talking, recruiting and preparing candidates for the position of responsibility. This research studies employee job satisfaction using machine learning models. Using a synthetic data created by IBM Watson, three main experiments were conducted to predict employee job satisfaction. The third and final part of the experiment involved using manual under sampling of the data to balance between classes.

1.3. MOTIVATION

The next phase of the computer world would be the automation phase. Every organization is investing huge into research and development of the said techniques. Automation would require to implement machine learning algorithms and artificial intelligence techniques.

Organizations need to cover a huge array of tasks that would require a lot of human resources which can be replaced with artificial intelligent bots.

1.4. AIM OF THE PROPOSED WORK

To develop a software that predicts the attrition possibility of an employee using various parameters with the help of various machine learning algorithms and also compare the results.

2) LITERATURE REVIEW

Table1:

Paper	Description
Reference 1	The developing interest for ML among business pioneers and leaders requires that scientists investigate its utilization inside business associations. One of the significant issues confronting business pioneers inside organizations is the loss of skilled workers. This paper studies employee attrition using ML models. Utilizing synthetic data made by IBM Watson, three experiments were carried out for predicting employee attrition. The first experiment included training the actual imbalanced dataset with these ML models: support vector machine (SVM), random forest classifier and K-nearest neighbor (KNN). The second experiment concentrated on utilizing ADASYN to deal with class imbalance, and then retraining on the new dataset utilizing the previously mentioned ML models. The third experiment included utilizing manual under sampling of the data to balance the classes. Accordingly, training an ADASYN-adjusted dataset with KNN (K = 3) accomplished the best results, with 0.93 F1-score. At last, by utilizing feature selection and random forest, F1-score of 0.909 was accomplished utilizing 12 features out of a total of 29.
Reference 2	Enterprise managements presume that higher pay grades will maintain and increase effective operation of employees in the future. This paper inspects the effect of high pay grade as incentive for improved employee performance. The authors collected data from various sources in three stages in a span of 12 months, and assessed the effects of pay grade on ensuing performance and self-esteem. They hypothesized that pay grade effects on employees' self-esteem brought about improvement in performance. The hypothesis is based on the assumption that pay levels in a company conveys how much the company appreciates the efforts of the individual and thus affects employees' self-esteem and hence performs better.
Reference 4	Customer churn is a serious issue for most businesses, as loss of customer affects profits and bringing in new customers is not easy. Prediction models for customer churn can be beneficial in developing customer retention programs. Employee attrition has a similar effect on businesses, causing operational

	<p>disturbances, customer discontentment and effort and time wasted in finding replacements. This paper surveys and compares some machine learning algorithms that have been employed to design predictive customer churn models. The authors also carried out a case study designing and comparing predictive employee attrition models. They also propose a value model that can identify how many of the attrition affected employees were valuable.</p>
Reference 6	<p>In this paper, the authors design a supervised machine learning-based job recommendation system. This algorithm utilizes all previous job changes and data linked with organizations and employees to predict next job transition of an employee. They trained a machine learning model using large dataset of job transitions of approximately 5 million employees publicly available on the Internet. The data on each employee is divided into three sections: the first section contains personal information, second section contains professional background of the employee and third section contains educational background of the employee. Experiments conducted by the authors have proved that job transitions can be predicted accurately. The machine learning algorithm used is a decision tree + naive Bayes hybrid classifier (DTNB).</p>
Reference 7	<p>Imbalanced data classification often occurs in few important practical applications such as data mining and pattern recognition in medical sciences. Most of the current classification techniques are designed by assuming the training set used is evenly distributed. However, they are faced with a critical bias issue when the training dataset is greatly imbalanced which leads to below par performance. SMoTE is a major approach of oversampling the positive class or the minority class. However, it is restricted to an assumption, that the local space between any two positive cases is positive or belongs to the minority class, which may not always be correct in the case when the training data is non-linearly separable. However, plotting the training data into a more linearly separable space can fix this issue. In this paper, the authors have combined Locally Linear Embedding algorithm (LLE) and SMOTE so that oversampling can be done on datasets that are non-linearly separable. Experiments have shown that this technique yields better results than traditional SMOTE.</p>
Reference 14	<p>Data classification is very important in data mining which has lead to a vast amount of studies in machine learning. Class imbalance is an issue in data classification in which a class of data will exceed in number than another class. Sentiment Analysis is an assessment of written and spoken language which can ascertain a person's emotions and attitudes and is usually used as dataset for machine learning. In this paper, the authors do a comparative study of Support Vector Machine (SVM) algorithm: Sequential Minimal optimization (SMO) with Synthetic Minority Over-Sampling Technique (SMOTE) and Naive Bayes Multinomial (NBM) algorithm with SMOTE</p>

	<p>for classification of data with the same Sentiment Analysis datasets collected by students of University of San Carlos. SMO is an algorithm used to solve quadratic programming problem in training SVMs. A GUI called Weka with a suite of machine learning algorithms for data mining, is utilised to pre-process and classify the data. They were able to conclude that SMOTE was effective depending on how the datasets were processed before applying the SMOTE and the kind of training and testing is also a way of obtaining reliable results. They also concluded that oversampling may not improve noisy sentiment analysis data which does not have meaning.</p>
Reference 3	<p>The motivation behind this paper is to explore US lodging properties' employee retention initiatives, and to look at the effect of those practices on worker turnover and retention. Using the Directory of Hotel and Lodging Companies, a helpful sample data of 24 administration organizations are chosen. A self-administered mail study instrument is created to gauge and test organizational initiatives on employee turnover and retention. Utilizing SPSS 16.0, two measurable tests are utilized to test study hypotheses. Correlation analysis is utilized to recognize the connections among predictor and response factors. In the same manner, regression examination is utilized to analyze the connections among predictor and response factors hypothesizing that the efficacy of practicing the human resource management organizational initiatives on management and non-management retention and turnover will vary. The discoveries uncover that Corporate Culture, Hiring and Promotions and Training practices impact non-management employee retention. Hiring and Promotion rehearses practices impact management retention. Besides, Organizational Mission, Goals and Direction, and Employee Acknowledgment, Rewards and Compensation were found to decidedly decrease non-management employee turnover. owing to the survey methodology and the moderately low response rate, speculation of the investigation discoveries is limited. Future replication studies are suggested. The discoveries will prepare lodging associations and industry experts with the instruments to decrease worker turnover and for maintaining employee retention. This should positively affect productivity.</p>
Reference 5	<p>Employee turnover is a genuine worry in information based associations. When representatives leave an association, they take with them priceless implicit information which is frequently the source of advantage for the business. For an association to consistently have a higher advantage over its competition, it should make it an obligation to limit employee attrition. This study recognizes worker related attributes that add to the prediction of employee attrition in companies. Three hundred and nine records of employees of one of the Higher Institutions in Nigeria who worked in and left the institution between 1978 and 2006 were utilized for the investigation. The</p>

	<p>demographic and occupation related records of the employee were the primary data which were utilized to arrange the employee into some predefined attrition classes. Waikato Environment for Knowledge Analysis (WEKA) and See5 for Windows were utilized to produce decision tree models and rule-sets. The results were then utilized for building up a predictive model that was utilized to anticipate new instances of employee attrition. A structure for a software tool that can execute the guidelines created in this study was additionally proposed.</p>
Reference 8	<p>In recent times, research groups have discovered that an imbalanced dataset could be one of the hindrances for many ML algorithms. In the learning procedure of the ML algorithms, if the proportion of minority classes and majority classes is notably different, ML will in general be commanded by the majority classes and the features of the minority classes are barely recognized. Subsequently, the classification precision of the minority classes might be low when compared to the classification precision of majority classes. The features in the minority classes are ordinarily hard to be completely recognized. In this paper, so as to re-balance the class distribution, the joined approaches of two strategies, Complementary Neural Network (CMTNN) and SMOTE, are proposed. CMTNN is applied as an under-sampling procedure, whereas, SMOTE is utilized as an over-sampling method. CMTNN is utilized as a result of its special feature of predicting "Truth" classified data but additionally the "False" data as well. SMOTE is applied since it can make new cases instead of repeating the already existing cases.</p>
Reference 9	<p>Imbalanced data learning is risky as conventional ML approaches fail to give satisfactory outcomes because of skewed class distribution. Rather than the two usual solutions to this problem, undersampling and oversampling, a new approach to develop the classifiers from imbalanced datasets is proposed in this paper by joining SMOTE and BiasedSVM approaches. Often, real-world datasets are predominantly consists of normal examples with just a small percentage of abnormal examples. The expense of misclassifying an abnormal example into a normal model is frequently a lot higher than that of the converse mistake. Test results affirms that the proposed mix approach of SMOTE and biased SVM can accomplish better classifier performance.</p>
Reference 10	<p>In recent years, data mining is utilized for health care management to characterize/justify disease prevalence and medical diagnosis. In any case, data mining issues are challenging in health care services because of huge, complex, heterogeneous, hierarchical time series data. The yearly number of death brought about by cancers is around million worldwide and breast cancer is one of the five most life-threatening kinds of cancer. It is crucial to know the survivability of the patients and to ease the decision making process with respect to</p>

	<p>treatment and financial arrangements. In the interim, false classification will cause wasted money and/or wrong treatments to cure breast cancer. In this study, the authors propose new algorithms to enhance the efficacy of classification for 5-year survivability of breast cancer patients from a huge dataset with imbalanced property. Results from this show that the hybrid algorithm of SMOTE + PSO + C5 is the best one for 5-year survivability of breast cancer patient categorization among all algorithm combinations. They conclude that, executing SMOTE in suitable searching algorithms such as PSO and classifiers such as C5 can remarkably improve the efficacy of grouping for classification for huge imbalanced datasets.</p>
Reference 11	<p>The Employee turnover has consistently been an important matter of worry for organizations. In the present period of globalization there are plentiful opportunities for talented individuals in this world, therefore, workers always move from one organization to another. Due to this organizations are facing the issue of employee attrition. A huge degree of worker turnover is profoundly damaging to both the association and the employees. The most effective method to decrease employee attrition is a definitive test for HR executives. This article presents a comprehensive perspective of attrition and retention of workers in this competitive scenario regarding Retail Industry. Alongside other factors, Job Satisfaction has been considered as a significant source of attrition and retention. The research is based on their literature review and also from the data accessible on the web.</p>
Reference 12	<p>The IT industry has been a significant contributor to the Indian economy throughout the last two decades. However, lately, numerous new opportunities are opening up for the best talents. Subsequently, the employee attrition rate is very high in the IT segment nowadays. The target of this paper is to look at the role that Herzberg's motivational and hygiene factors play in guaranteeing job satisfaction of the employees in this industry. Herzberg's theory of workplace motivation has been one of the most approved theories of motivation. But it has been found to work with certain distinctions in various nations, particularly Asia, because of social contrasts. For instance, see Sithiphand, 1978; Hauff, 2014. Attempts have been made to comprehend the workplace motivation of employees in the Indian IT industry. But, there has barely been made any effort to comprehend the manner in which the factors pointed out by Herzberg influence job satisfaction of employees in the Indian IT field. In this study, the authors inspected the role of these two sets of factors. Information was gathered from 153 IT employees. It was discovered that in spite of what was anticipated by the theory, the cleanliness factors assume a more grounded role in predicting job satisfaction of the Indian IT employees. The ramifications of this finding are talked about in the paper.</p>

Reference 13	Imbalanced data can influence the performance of standard classifier algorithms that lead to the biased outcomes towards majority classes. The SMOTE technique fixes the imbalanced data issue by making synthetic cases of minority classes. However, the usage of SMOTE brought about over-generalization on the grounds that synthesized instances have a similar amount no matter what the distribution of instances is. Accordingly, the boundaries between classes are vague. The SMOTE-Simple Genetic Algorithm (SMOTESGA) strategy is utilized to decide the sampling rate of each example so as to get unequal number of synthesized instances. The tests were performed utilizing some imbalanced datasets by comparing the classification results calculated utilizing G-means and F-Measure. The results of the use of genetic algorithm along with SMOTE can improve the classification result by acquiring better G-means and F-measure scores.
Reference 15	Human capital is of a high concern for organizations' administration where their most interest is in employing the profoundly qualified staff who are relied upon to perform exceptionally. In this paper, data mining strategies were used to design a classification model to predict the performance of employees. To design the classification model the CRISP-DM data mining strategy was embraced. Decision tree was the chief data mining algorithm used to construct the classification model, where quite a few classification rules were created. To approve the produced model, several trials were conducted utilizing genuine data gathered from several organizations. The model is planned to be used for predicting new candidates' performance.

3) OVERVIEW OF THE PROPOSED SYSTEMS

3.1 Introduction to Related Concepts

In this project the model is prepared in two steps. First the imbalanced data is converted to balanced data using SMOTE and Random Sample. Then the data is trained using various Classifier models like SVM, KNN, Random Forest, Naïve Bayes and Artificial Neural Network(ANN).

Then their performance is compared in terms of accuracy, precision and F1 score. The existing paper uses ADASYN for sampling but we are using random sample and SMOTE instead. For classification purpose we are using ANN and Naïve bayes for better comparison. But let us first understand why we cannot proceed with the raw data and the various ways in which we can deal with imbalanced data.

The primary inspiration driving the need to preprocess imbalanced data before we feed them into a classifier is that commonly classifiers are typically more sensitive in distinguishing the

majority share class and less when it comes to the minority class. Accordingly, on the off chance that we don't deal with the issue, the yield will be one-sided, so in most cases the output will be favorable towards the majority class even if it is not the case. A lot of methods were tried and tested in order to overcome this issue of imbalanced data.

The two variations in which one can convert this imbalanced data is by using two types of methods:

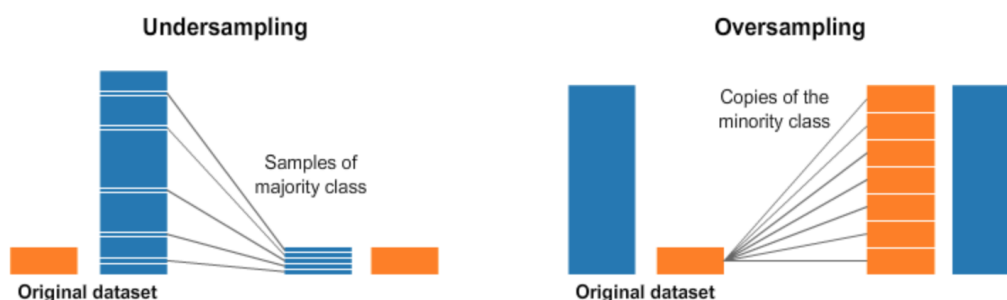


Fig 1 – Undersampling and Oversampling

a) Undersampling:

Under sampling alludes to a group of procedures intended to adjust the class distribution for a dataset that involves classification properties and also has a majority and minority class issue. A polarity induced class distribution will have at least one class with hardly any models (the minority classes) and at least one class with numerous models (the majority classes). It is best comprehended with regards to a binary (two-class) order issue where class 0 is the class with bulk attached to it and class 1 is the class with less representation. Under sampling procedures expel models from the preparation dataset that have a place with the majority class to more readily adjust the class conveyance, for example, diminishing the slant from a 1:100 to a 1:10, 1:2, or even a 1:1 class circulation.

This is not quite the same as oversampling that includes adding guides to the minority class with an end goal to diminish the slant in the class conveyance. Under sampling techniques can be utilized legitimately on a preparation dataset that can at that point, thusly, be utilized to fit an AI model.

The least difficult under sampling strategy includes haphazardly choosing models from the majority class and erasing them from the preparation dataset. This is alluded to as irregular under sampling. Albeit basic and viable, a restriction of this strategy is that models are expelled with no worry for how helpful or significant they may be in deciding the choice limit between the classes. This implies it is conceivable, or even likely, that helpful data will be erased.

b) Oversampling:

Oversampling refers to the idea of generating more data values of the minority class by fabricating new set of data that governs the plot of the minority class. Different techniques are used to implement oversampling.

c) Random Over Sampling:

Random Oversampling involves the idea of enhancing the training data with numerous duplicates of a portion of the minority classes. Oversampling should be possible more than once (2x, 3x, 5x, 10x, and so on.) This is one of the most punctual and proposed strategies, that is additionally demonstrated to be strong. Instead of copying each example in the minority class, some of them might be arbitrarily picked with substitution.

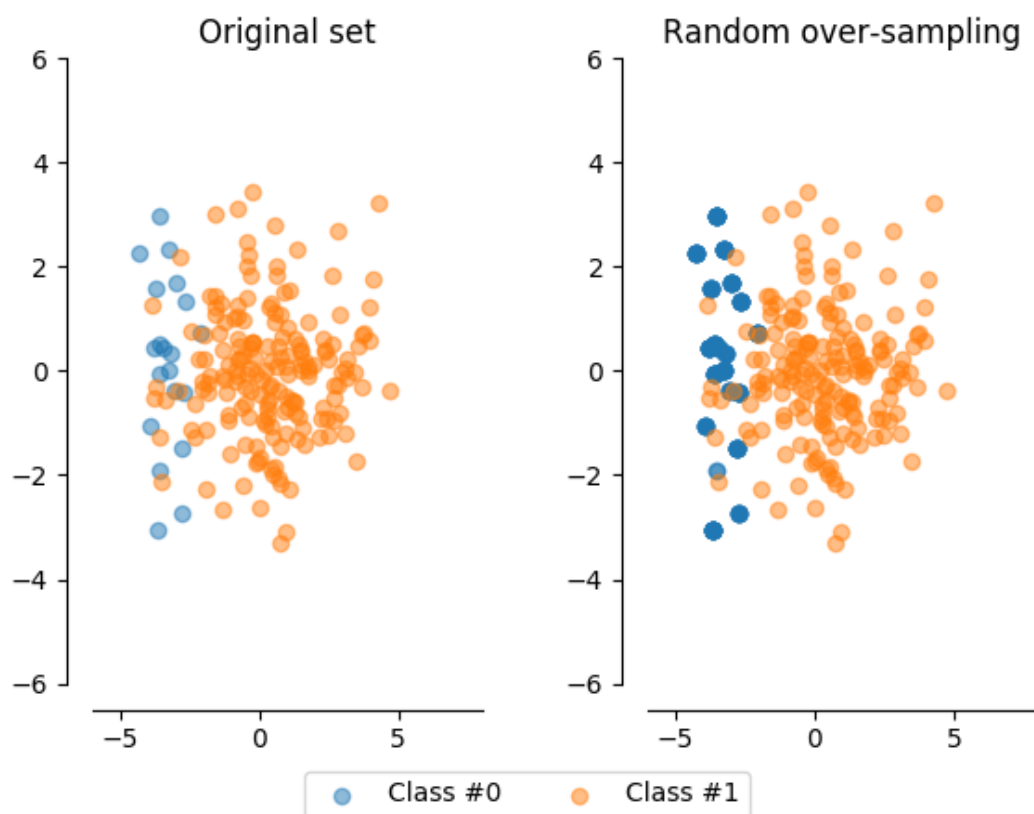


Fig 2 – Random Oversampling

The above picture depicts the basic idea of random oversampling.

d) SMOTE:

The SMOTE algorithm is one of the first and still the most mainstream algorithmic way to remove the imbalance in datasets between majority and the minority classes. The algorithm

was designed for the same intricate purpose and was developed in the year 2002. It works by using under sampling method to generate new synthetic points that build up the size of the minority class.

The SMOTE algorithm is parameterized with k-neighbors (the quantity of closest neighbors it will consider) and the quantity of new focuses you wish to make. Each progression of the algorithm will:

- First randomly select a minority point
- Secondly select any of its k-neighbors nearest neighbors randomly which also belong to the same class.
- Now randomly select an alpha value which ranges between the values 0 and 1, inclusive.
- Now generate a new synthetic point on the vector between the two points which is located lambda percent of the way from the point originally considered.

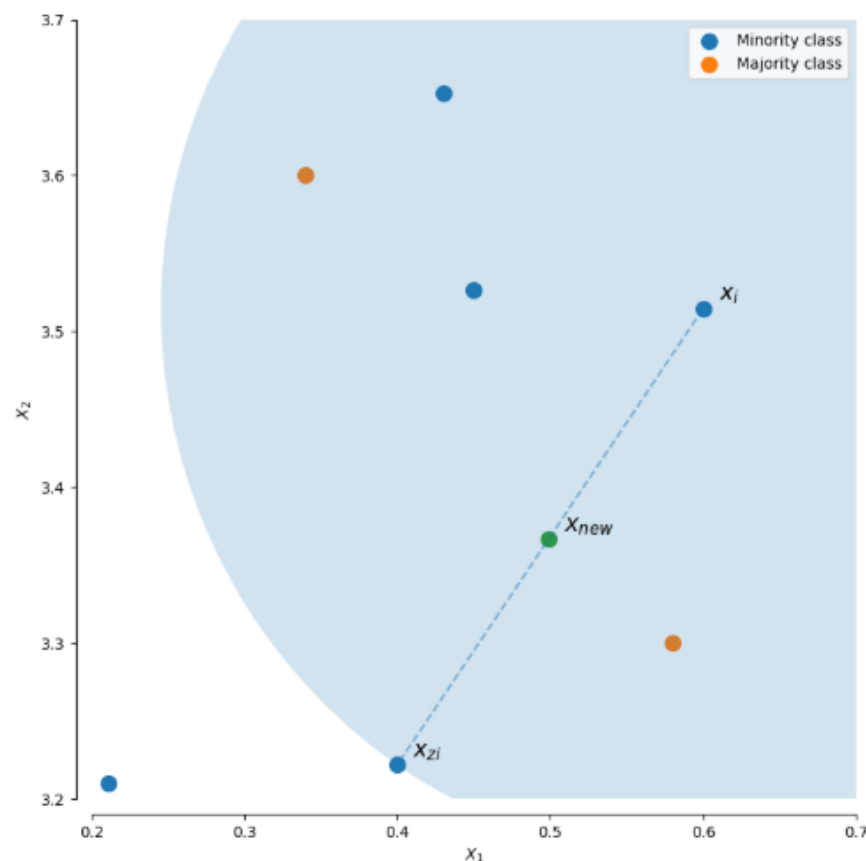


Fig 3 – Generating a new point in SMOTE algorithm

e) ADASYN:

ADASYN is similar SMOTE, and is based on it, with only one significant distinction. It will incline the sample space or in other words shows bias towards it (that is, the probability that a

specific point will be picked for duplicating) towards points which are not found in homogenous neighborhoods.

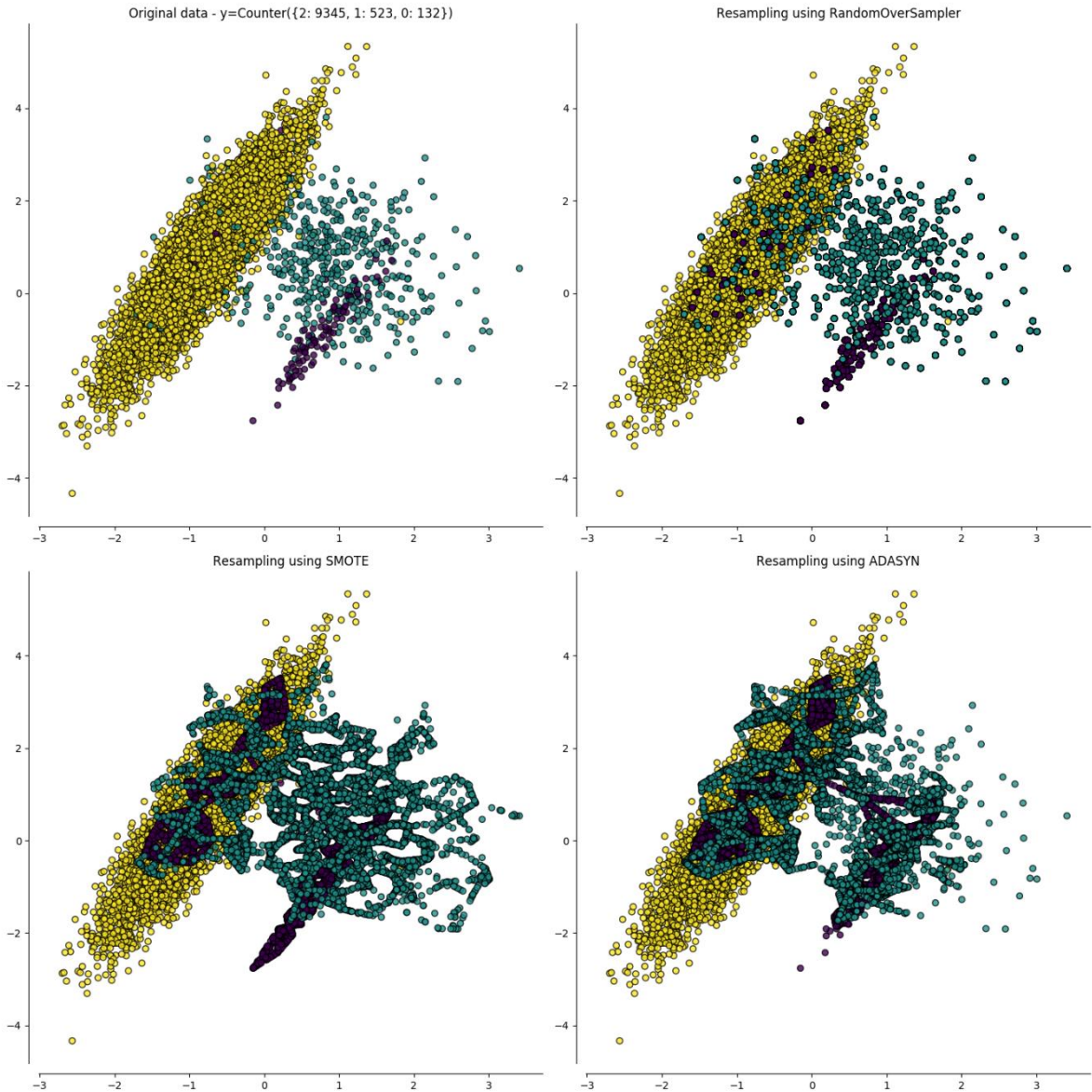


Fig 4 – Picture differentiating between Random Oversampling, SMOTE and ADASYN

f) Support Vector Machine (SVM):

Support Vector Machine (SVM) is one of the prominently used machine learning algorithms used to identify a pattern, spam filter and intrusion network anomaly. With the aid of class labels, SVM can learn the pattern. A virtual system is used to identify unknown samples with the model training dataset. The nearest data are support vectors and the features are declared by the expected class. Given the training dataset of n points of the form

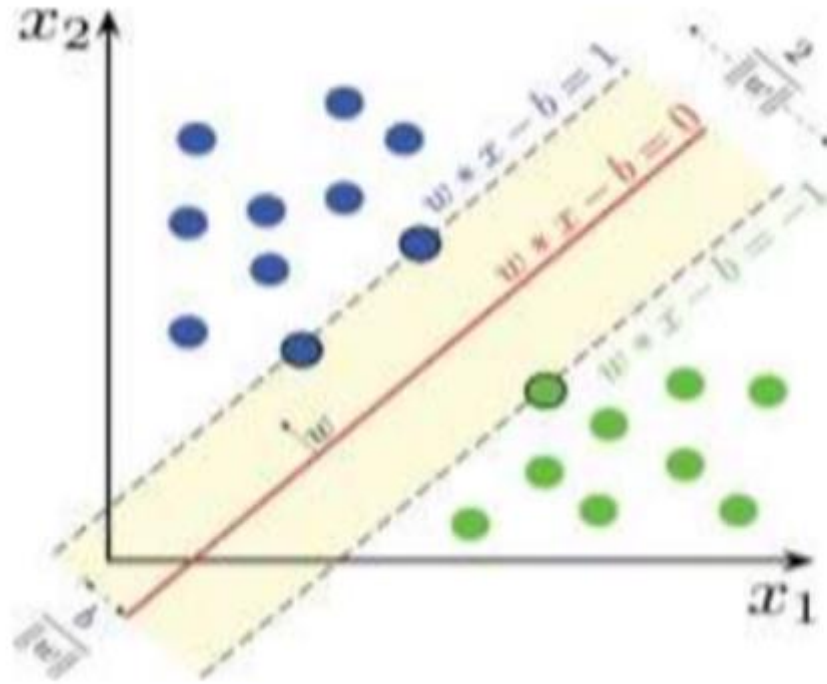


Fig 5- SVM model

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

K-Nearest Neighbors is one of the most foundational and vital machine learning classification algorithms. It comes under the supervised learning domain and is used most prominently in pattern identification, intrusion sensing and data collection. It is very widely available in real-life scenarios as it is non-parametric, which means that it does not make implicit assumptions unlike the most algorithms. For example, GMM assumes that the data given is distributed in Gaussian.

We have some earlier data that are grouped into attribute-identified classes, also known as training data. K nearest neighbor is a basic algorithm which reserves all cases that pre-exist and classifies the new cases introduced on the basis of an equivalence (e.g. function of distance). In 1970s, KNN was employed as a non-parametric technique in statistical estimation and pattern recognition.

F=Fig

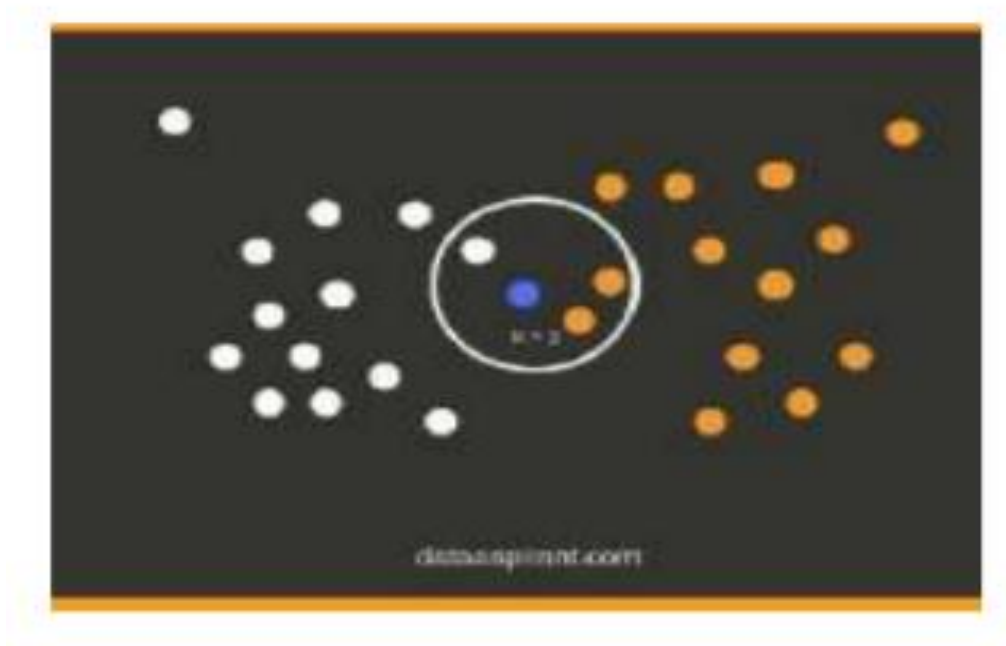


Fig 6 - KNN model

h) Naive Bayes classifier:

The Classification systems Naive Bayes are a series of Bayes' theorem rating algorithms. This is not just a single algorithm but a family of algorithms, all of which share a common definition, that is to say each pair of functions is distinct. Bayes' theorem determines the probability of a happening because of the possibility of another occurrence. Bayes' theorem equation :

$$P(c/x) = (P(x/c)P(c)) / P(x) \text{ where}$$

$P(c/x)$ = Posterior Probability

$P(x/c)$ = Likelihood

$P(c)$ = Class Prior Probability

$P(x)$ = Predictor Prior Probability

i) Random Forest Classifier:

Random forest is a type of algorithm which consists of a wide section of decision making trees which work in an assembled fashion. Each individual tree component makes out a decision. After all the trees make their decisions the one with the most votes becomes our model's prediction.

The basic root of this algorithm is **“A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.”** The correlation between models has to be low, which is a vital characteristic. The reason for this is that as long as all trees don’t make the same decision, each tree protects other trees. Albeit some of the trees make wrong prediction most of them make the right one which makes the herd take a right prediction. So the following characteristics ensures that random forest works well:

1. There should be some clear signals in our features so that the models built upon the features developed do a better job than arbitrary guesswork.
2. The prediction made by individual trees should be in low correlation.

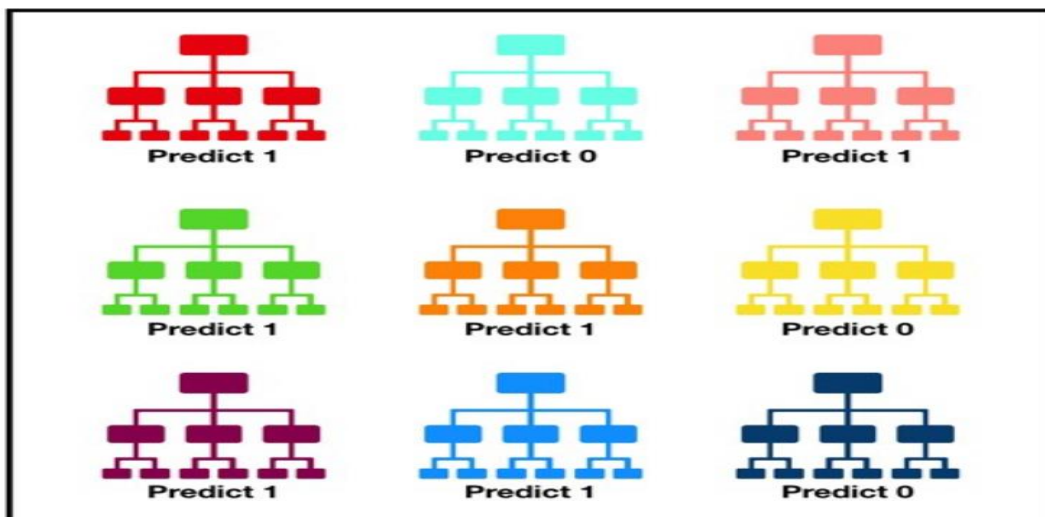


Fig 7 – Random Forest Trees

j) Artificial Neural Networks (ANN):

Through deep learning, a computer model learns how to recognize pictures, text, or sounds directly. Deep learning models can achieve high-tech precision and often superior performance compared to human level. Models are built using a broad variety of defined data and architectures of neural networks with several layers.

Neural networks need a trainer to explain what should be generated as an input response. The error value, also called cost function, is calculated and returned through the method based on the difference between the actual value and the expected value.

The cost function is measured for each layer in the network and used to change the weight and

thresholds for the next entry. Our goal is to reduce costs. The lower the cost function, the greater the actual value. This makes the error slightly smaller each time the network is able to analyze values. We return the resulting data through the entire neural network. As long as the actual value and expected value are different, those weights have to be modified. If we tweak them slightly and restart the neural network, we hope that we create a new Cost function smaller than the last one. This method has to be pursued until we optimize the costs to the lowest degree possible.

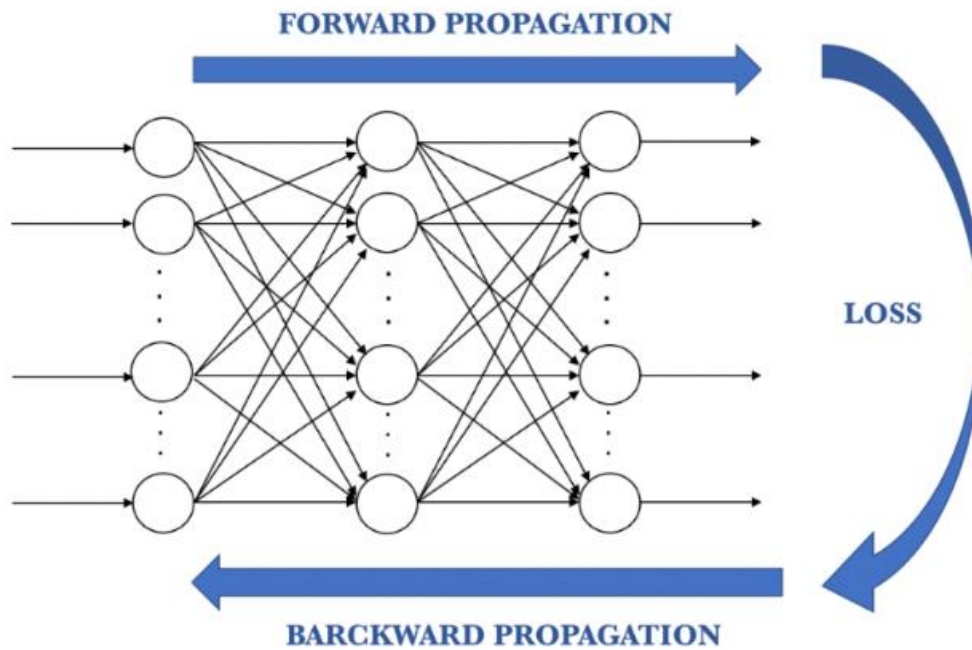


Fig 9- ANN propagation

3.3 Framework for the Proposed System

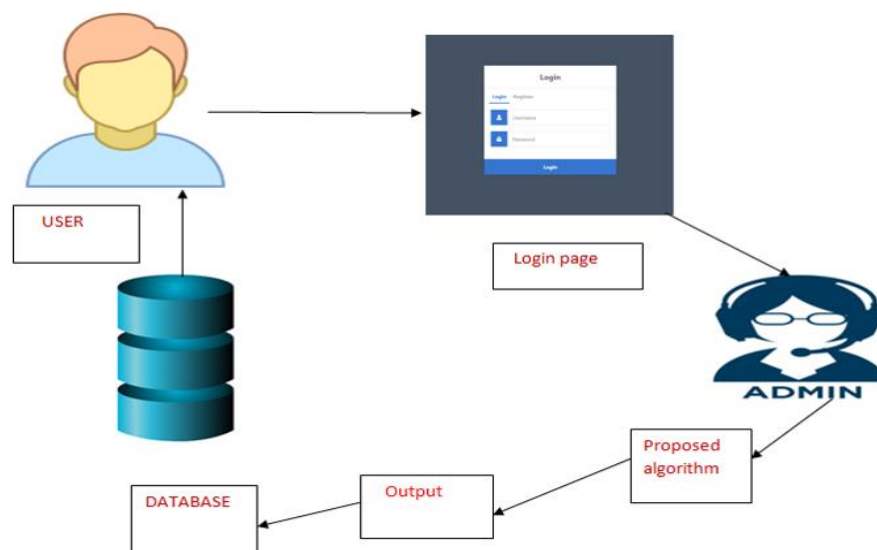


Fig 10 - Project Basic Architecture

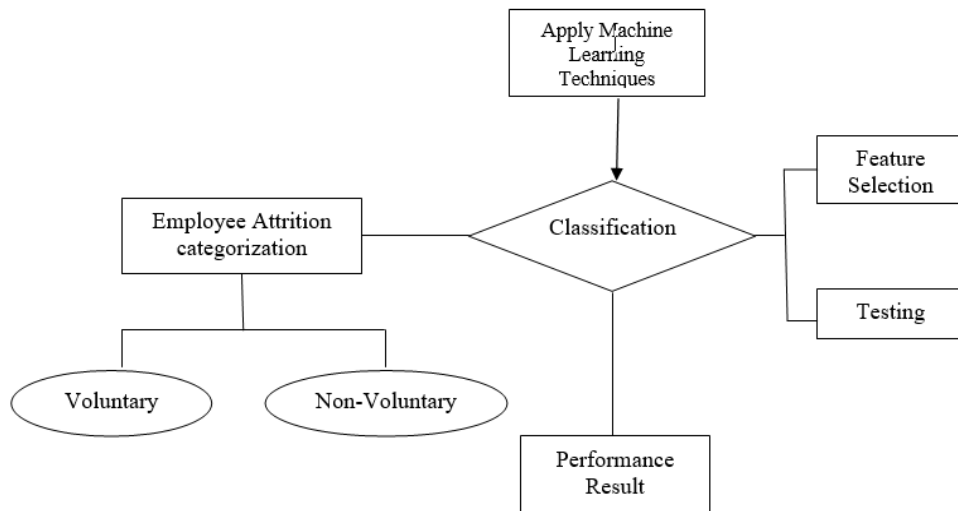


Fig 11 - Classification Flowchart

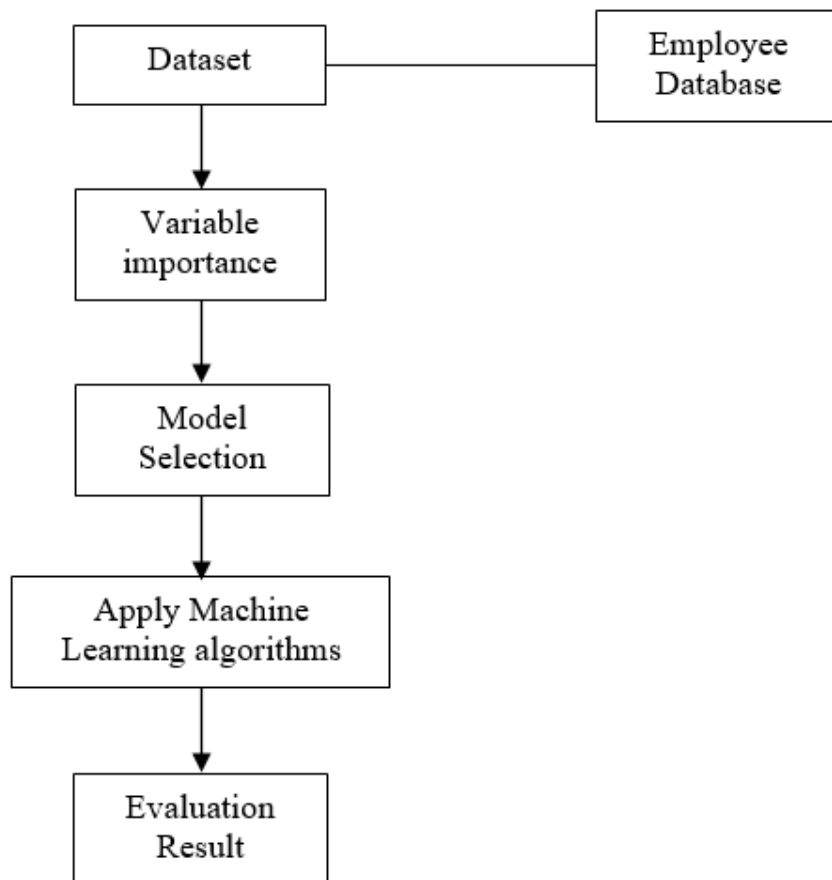


Fig 12 - The flow of project in sequential steps

3.4 Existing System Analysis:

The current existing method to predict employee attrition can be studied by the base paper we have considered.

In the paper we have considered, they have use three trials to predict attrition. In the first place, they have tried to anticipate employee attrition. For this purpose they utilized the first imbalanced dataset. In the subsequent trial, they have used the ADASYN algorithm to solve the class imbalance problem. In this experiment they used ADASYN oversampling technique to increase the samples of the minority class by creating synthetic data. In the third experiment that they have done, they used random oversampling technique in which they have selected an equal set of data from both the sides. Besides, each test included preparing and approving a lot of classifiers to anticipate the attrition.

The major drawbacks of ADASYN- Adaptive Synthetic Sampling are :

- For minority examples that are sparsely distributed, each neighborhood may only contain 1 minority example.
- Precision of ADASYN may suffer due to adaptability nature.

Since the accuracies of these algorithms greatly depend on the datasets being used we used the SMOTE algorithm on the dataset in order to try and overcome the drawbacks of the ADASYN algorithm and achieve a better accuracy on the predictions.

4) PROPOSED SYSTEM ANALYSIS

4.1 Proposed system methodology

Data Pre-processing:

The dataset is highly imbalanced and it is converted to balanced data by using up sampling methods like ADASYAN or SMOTE. After that the data is standardized or normalized to avoid over fitting also null values are replaced with zeros. Redundant columns are also removed in this step.

Data splitting:

Now 70% of the data is used for training and 30% is used for testing maintaining the class ratio. Training the model: As different models are to be used so k-fold cross validation is used for proper model selection. Then we fit our data to the models to evaluate the performance. Also we use different hyper parameters for choosing the best model.

4.2 Requirement Analysis

Employee attrition can be characterized as the loss of employees because of any of the accompanying reasons: individual reasons, low occupation fulfillment, low compensation, and a terrible business condition. Loss of employees can leave a company in a state of turmoil and uncertainty, hence prevention of employee attrition is essential for the sustainable growth of any company.

Especially in a country like India which is bound to become the technology hub of the world, loss of talented employees in the software field can hinder the growth speed of the country. Hence we are working on various Machine Learning algorithms and training a model data set in order to present a model to companies which will help them accurately predict any possible employee attrition and prevent it.

The initial requirements needed for this model to work is a sample dataset of the company and its employees and their work progress and various other factors which might affect the employee within the company. 70% of the dataset and all its factors will be used to test the models and the remaining 30% will be used to predict and test the employee attrition. Based on the training model and the algorithm used the accuracy of the test will vary and we can try out various algorithms depending on the dataset and provide the best one based on its acquired accuracy.

4.2.3.1 Hardware Requirements

Processor - I5 / I3

RAM - 4GB / 8GB

System Free Space - Minimum 15GB

4.2.3.2 Software Requirements

Programming Language - Python

IDE - Anaconda

5) RESULTS and SUMMARY

```
sns.countplot(data['Attrition'])# showing class imbalance
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fad1ac62860>
```

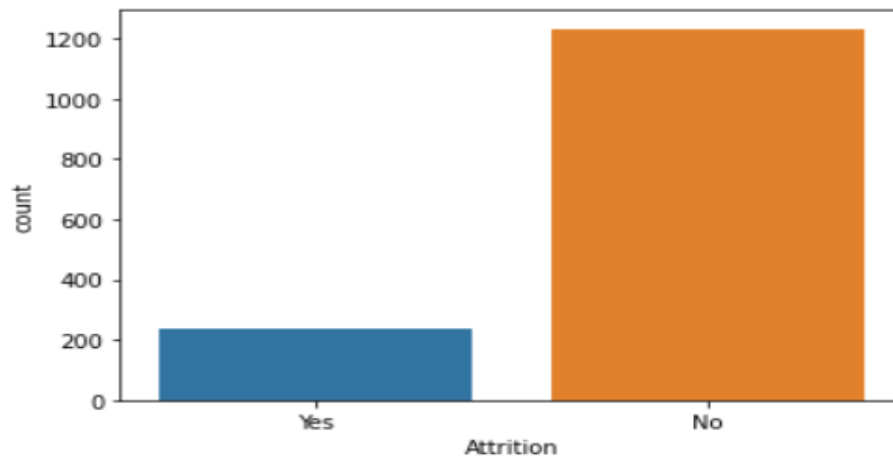


Fig 13 - Unbalanced Data

```
sns.countplot(y_sample)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fad1a8f79b0>
```

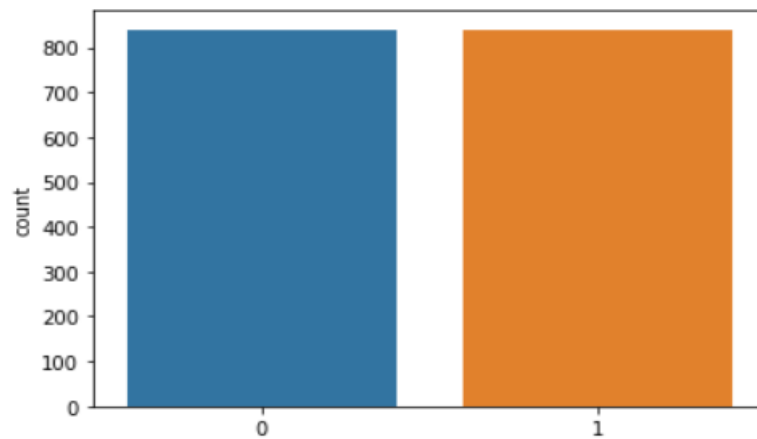


Fig 14 - Data after using Oversampling techniques

In the above figures we can identify that the data was quite imbalanced to start with, hence we used oversampling techniques to balance out the differences.

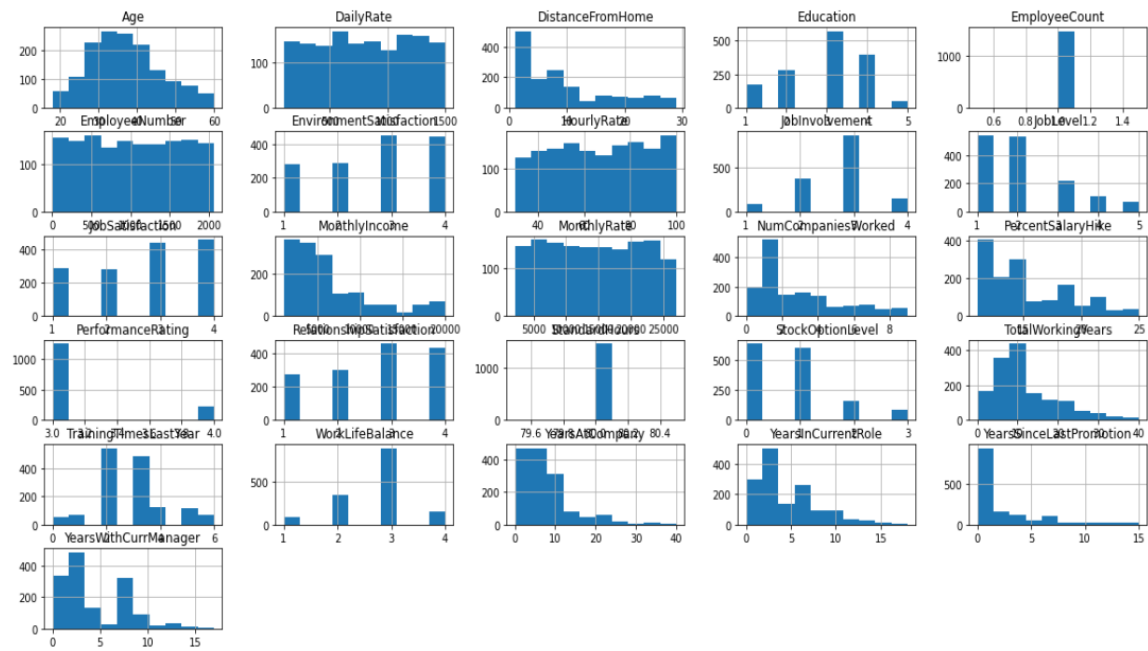


Fig 15 - Various parameters that affected the outcome

The above picture indicates the factors or the columns that have majorly affected the outcome we can see that few of them had major part to play than rest of the other attributes. Attributes such as age, wages, job satisfaction quotient, work life balance ha major parts to play than the others.

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvol
Age	1.000000	0.010661	-0.001686	0.208034	NaN	-0.010145	0.010146	0.024287	0
DailyRate	0.010661	1.000000	-0.004985	-0.016806	NaN	-0.050990	0.018355	0.023381	0
DistanceFromHome	-0.001686	-0.004985	1.000000	0.021042	NaN	0.032916	-0.016075	0.031131	0
Education	0.208034	-0.016806	0.021042	1.000000	NaN	0.042070	-0.027128	0.016775	0
EmployeeCount	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
EmployeeNumber	-0.010145	-0.050990	0.032916	0.042070	NaN	1.000000	0.017621	0.035179	-0
EnvironmentSatisfaction	0.010146	0.018355	-0.016075	-0.027128	NaN	0.017621	1.000000	-0.049857	-0
HourlyRate	0.024287	0.023381	0.031131	0.016775	NaN	0.035179	-0.049857	1.000000	0
JobInvolvement	0.029820	0.046135	0.008783	0.042438	NaN	-0.006888	-0.008278	0.042861	1
JobLevel	0.509604	0.002966	0.005303	0.101589	NaN	-0.018519	0.001212	-0.027853	-0
JobSatisfaction	-0.004892	0.030571	-0.003669	-0.011296	NaN	-0.046247	-0.006784	-0.071335	-0
MonthlyIncome	0.497855	0.007707	-0.017014	0.094961	NaN	-0.014829	-0.006259	-0.015794	-0
MonthlyRate	0.028051	-0.032182	0.027473	-0.026084	NaN	0.012648	0.037600	-0.015297	-0
NumCompaniesWorked	0.299635	0.038153	-0.029251	0.126317	NaN	-0.001251	0.012594	0.022157	0
PercentSalaryHike	0.003634	0.022704	0.040235	-0.011111	NaN	-0.012944	-0.031701	-0.009062	-0
PerformanceRating	0.001904	0.000473	0.027110	-0.024539	NaN	-0.020359	-0.029548	-0.002172	-0
RelationshipSatisfaction	0.053535	0.007846	0.006557	-0.009118	NaN	-0.069861	0.007665	0.001330	0
StandardHours	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
StockOptionLevel	0.037510	0.042143	0.044872	0.018422	NaN	0.062227	0.003432	0.050263	0
TotalWorkingYears	0.680381	0.014515	0.004628	0.148280	NaN	-0.014365	-0.002693	-0.002334	-0
TrainingTimesLastYear	-0.019621	0.002453	-0.036942	-0.025100	NaN	0.023603	-0.019359	-0.008548	-0

Fig 16 - Sample of attribute values weighing the outcome

Accuracy using smote

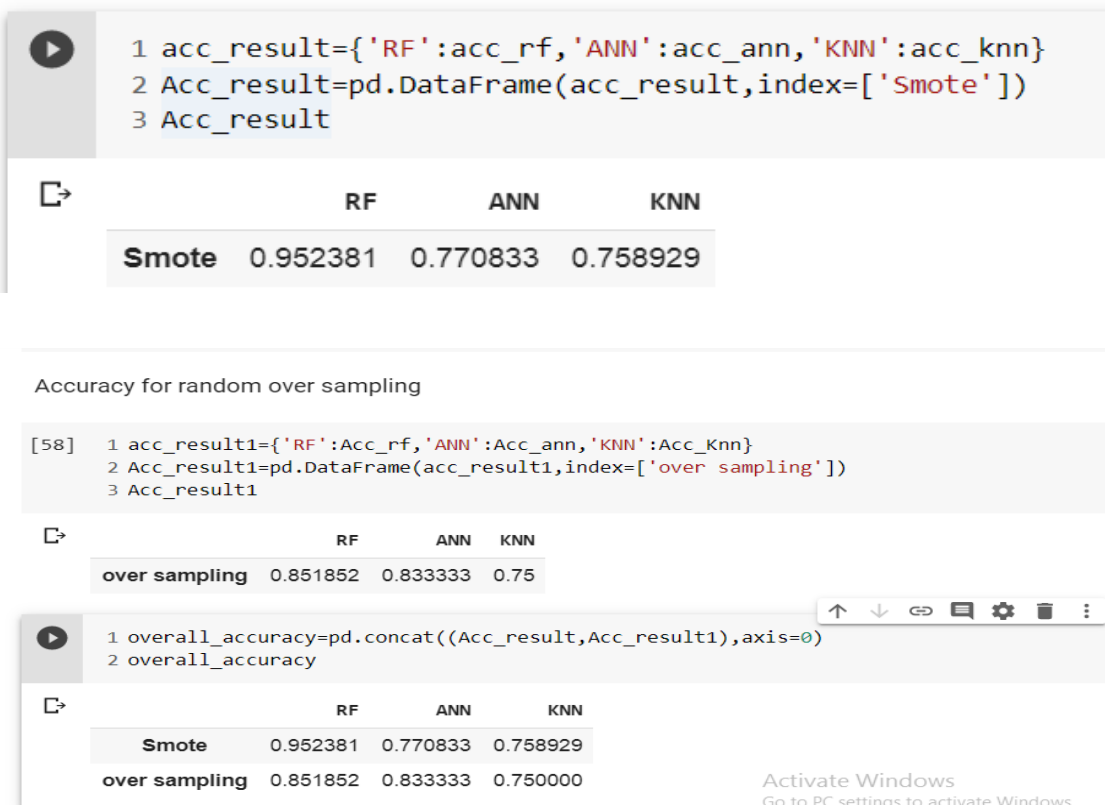


Fig.7 Accuracy of SMOTE and oversampling techniques

Now coming to our final result, comparing two outcomes where in one we used random oversampling technique for oversampling the data and in the other we used SMOTE for oversampling we obtain the following results.

So when we used **SMOTE**, we had the following accuracies:

Random Forest: 0.952381

ANN: 0.770833

KNN: 0.78929

And when we used **random sampling** technique we had the following accuracies:

Random Forest: 0.851852

ANN: 0.833333

KNN: 0.75

The results in the Base Paper after using **ADASYN** were as follows:

Random Forest : 92.6

KNN : 87.2

So from the above derived data we can conclude that:

- 1) Of all the techniques we used the combination of SMOTE and Random Forest has the highest accuracy.
- 2) Thus we can say that SMOTE used with Random Forest gives a slightly better accuracy than ADASYN with Random Forest

- 3) Compared with Random Oversampling, SMOTE was the better accurate oversampling technique even though it has less accuracy when combined with ANN.
- 4) Random Forest algorithm provided better results compared with ANN and KNN.

6) REFERENCES

1. TITLE: Predicting Employee Attrition using Machine Learning (Base Paper)

AUTHORS: Sarah S. Alduayj, Kashif Rajpoot

2. TITLE: The Effects Of Pay Level On Organization-Based Self-Esteem And Performance: A Field Study

AUTHORS: D. G. Gardner, L. V. Dyne and J. L. Pierce

3. TITLE: An Exploratory Study Of Us Lodging Properties' Organizational Practices On Employee Turnover And Retention

AUTHORS: E. Moncarz, J. Zhao and C. Kay

4. TITLE: Employee churn prediction

AUTHORS: G. K. P. V. Vijaya Saradhi

5. TITLE: ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS

AUTHORS: D. A. B. A. Alao

6. TITLE: Machine learned job recommendation

AUTHORS: Ioannis Paparrizos, B. Barla Cambazoglu, Aristides Gionis

7. TITLE: Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding

AUTHORS: Juanjuan Wang; Mantao Xu; Hui Wang; Jiwu Zhang

8. TITLE: Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm

AUTHORS: Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung

9. TITLE: Combination approach of SMOTE and biased-SVM for Imbalanced datasets

AUTHORS: He-Yong Wang

10. TITLE: A hybrid classifier combining SMOTE with PSO to estimate 5- year survivability of

breast cancer patients

AUTHORS : Kung-Jeng Wanga, Bunjira Makonda, Kun-Huang Chena,
Kung-Min Wang

11. TITLE: HISTOPATHOLOGICAL IMAGE ANALYSIS: A REVIEW

AUTHORS: S. Kaur and R. Vijay

12. TITLE: Motivators, Hygiene Factors and Job Satisfaction of Employees in IT Sector in India

AUTHORS: Akhil Gokuldas Warriar, Rajiv Prasad

13. TITLE: The Implementation of Genetic Algorithm in Smote (Synthetic Minority
Oversampling Technique) for Handling Imbalanced Dataset Problem

AUTHORS: Tince Etlin Tallo, Aina Musdholifah

14. TITLE: An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data
Set

AUTHORS: Tince Etlin Tallo, Aina Musdholifah

15. TITLE: USING DATA MINING TECHNIQUES TO BUILD A CLASSIFICATION
MODEL FOR PREDICTING EMPLOYEES PERFORMANCE

AUTHORS: A. Al-Radaideh and E. A. Nagi

16. <https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family>

Author: Rohit Walimbe

17. <https://medium.com/coinmonks/smote-and-adasync-handling-imbalanced-data-set-34f5223e167>

Author: Indresh Bhattacharya

18. Digi-key Corporation. <http://www.digikey.com/>

19. Alldatasheet. <http://www.alldatasheet.com/>