**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.



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 Engineeringto 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 :

#### 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.

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

# 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](#_gjdgxs)

**Table of Contents** 4

[**List of Figures**](#_30j0zll)  6

[**List of Tables**](#_1fob9te) 7

**Abbreviations** 8

[**Symbols and Notations**](#_3znysh7)  9

1. [**INTRODUCTION** 1](#_2et92p0)**0**

Objective 10

Motivation 11

Background 11

1. **Literature Survey** 17
2. **PROJECT DESCRIPTION TECHNICAL SPECIFICATION** 17
3. **DESIGN APPROACH AND DETAILS** 18

Design Approach / Materials & Methods 19

Codes and Standards  20

Constraints, Alternatives and Tradeoffs .

#### SCHEDULE, TASKS AND MILESTONES .

1. **PROJECT DEMONSTRATION** .
2. **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  3  4  5  6  7  8 | KNN model  Project Basic Architecture  Classification Flowchart  The Flow of project in sequential steps  Final Results  Random Forest trees  ANN propagation |  | 10  11  12  12  23  19  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

### INTRODUCTION

### OBJECTIVE

The first task at hand would be to balance the imbalanced sets of date 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

### 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.

### 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.

### 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.

### LITERATURE REVIEW

Table1:

|  |  |
| --- | --- |
| Paper | Description |
| Reference 1 | The devel0ping interest f0r ML am0ng business pi0neers and leaders requires that scientists investigate its utilizati0n inside business ass0ciati0ns. One 0f the significant issues c0nfr0nting business pi0neers inside 0rganizati0ns is the l0ss 0f skilled w0rkers. This paper studies empl0yee attriti0n using ML m0dels. Utilizing synthetic data made by IBM Wats0n, three experiments were carried 0ut f0r predicting empl0yee attriti0n. The first experiment included training the actual imbalanced dataset with these ML m0dels: supp0rt vect0r machine (SVM), rand0m f0rest classifier and K-nearest neighb0r (KNN). The sec0nd experiment c0ncentrated 0n utilizing ADASYN t0 deal with class imbalance, and then retraining 0n the new dataset utilizing the previ0usly menti0ned ML m0dels. The third experiment included utilizing manual under sampling 0f the data t0 balance the classes. Acc0rdingly, training an ADASYN-adjusted dataset with KNN (K = 3) acc0mplished the best results, with 0.93 F1-sc0re. At last, by utilizing feature selecti0n and rand0m f0rest, F1-sc0re 0f 0.909 was acc0mplished utilizing 12 features 0ut 0f a t0tal 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 0f pay grade 0n ensuing perf0rmance 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 | Cust0mer churn is a seri0us issue f0r m0st businesses, as l0ss 0f cust0mer affects pr0fits 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 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. |
| Reference 6 | 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 secti0ns: the first secti0n c0ntains pers0nal inf0rmati0n, sec0nd secti0n c0ntains 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). |
| Reference 7 | Imbalanced data classificati0n 0ften 0ccurs in few imp0rtant  practical applicati0ns such as data mining and patter rec0gniti0n in medical sciences. M0st 0f the current classificati0n techniques are designed by assuming the training set used is evenly distributed. H0wever, they are faced with a critical bias issue when the training dataset is greatly imbalanced which leads t0 bel0w par perf0rmance. SM0TE is a maj0r appr0ach 0f 0versampling the p0sitive class 0r the min0rity class. H0wever, it is restricted t0 an assumpti0n, that the l0cal space between any tw0 p0sitive cases is p0sitive 0r bel0ngs t0 the min0rity class, which may n0t always be c0rrect in the case when the training data is n0n-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. |
| Reference 14 | Data classificati0n is very imp0rtant in data mining which has lead t0 a vast am0unt 0f studies in machine learning. Class imbalance is an issue in data classificati0n in which a class 0f data will exceed in number than an0ther class. Sentiment Analysis is an assessment 0f written and sp0ken language which can ascertain a pers0n's em0ti0ns and attitudes and is usually used as dataset f0r machine learning. In this paper, the auth0rs d0 a c0mparative study 0f Supp0rt Vect0r Machine (SVM) alg0rithm: Sequential Minimal 0ptimizati0n (SMO) with Synthetic Min0rity Over-Sampling Technique (SMOTE) and Naive Bayes Multin0mial (NBM) alg0rithm with SMOTE f0r classificati0n 0f data with the same Sentiment Analysis datasets c0llected by students 0f University 0f San Carl0s. SM0 is an alg0rithm used t0 s0lve quadratic  pr0gramming pr0blem in training SVMs. A GUI called Weka with a suite 0f machine learning alg0rithms f0r data mining, is utilised t0 pre-pr0cess and classify the data. They were able t0 c0nclude that SMOTE was effective depending 0n h0w the datasets were pr0cessed bef0re applying the SMOTE and the kind 0f training and testing is als0 a way 0f 0btaining reliable results. They als0 c0ncluded that 0versampling may n0t impr0ve n0isy sentiment analysis data which d0es n0t have meaning. |
| Reference 3 | The m0tivati0n behind this paper is t0 expl0re US l0dging pr0perties' empl0yee retenti0n initiatives, and t0 l00k at the effect 0f th0se practices 0n w0rker turn0ver and retenti0n. Using the Direct0ry 0f H0tel and L0dging C0mpanies, a helpful sample data 0f 24 administrati0n 0rganizati0ns are ch0sen. A self-administered mail study instrument is created t0 gauge and test 0rganizati0nal initiatives 0n empl0yee turn0ver and retenti0n. Utilizing SPSS 16.0, tw0 measurable tests are utilized t0 test study hyp0theses. C0rrelati0n analysis is utilized t0 rec0gnize the c0nnecti0ns am0ng predict0r and resp0nse fact0rs. In the same manner, regressi0n examinati0n is utilized t0 analyze the c0nnecti0ns am0ng predict0r and resp0nse fact0rs hyp0thesizing that the efficacy 0f practicing the human res0urce management 0rganizati0nal initiatives 0n management and n0n-management retenti0n and turn0ver will vary. The disc0veries unc0ver that C0rp0rate Culture, Hiring and Pr0m0ti0ns and Training practices impact n0n-management empl0yee retenti0n. Hiring and Pr0m0ti0n rehearses practices impact management retenti0n. Besides, Organizati0nal Missi0n, G0als and Directi0n, and Empl0yee Ackn0wledgment, Rewards and C0mpensati0n were f0und t0 decidedly decrease n0n-management empl0yee turn0ver. 0wing t0 the survey meth0d0l0gy and the m0derately l0w resp0nse rate, speculati0n 0f the investigati0n disc0veries is limited. Future replicati0n studies are suggested. The disc0veries will prepare l0dging ass0ciati0ns and industry experts with the instruments t0 decrease w0rker turn0ver and f0r maintaining empl0yee retenti0n. This sh0uld positively affect productivity. |
| Reference 5 | Empl0yee turn0ver is a genuine w0rry in inf0rmati0n based ass0ciati0ns. When representatives leave an ass0ciati0n, they take with them priceless implicit inf0rmati0n which is frequently the s0urce 0f advantage f0r the business. F0r an ass0ciati0n t0 c0nsistently have a higher advantage 0ver its c0mpetiti0n, it sh0uld make it an 0bligati0n t0 limit empl0yee attriti0n. This study rec0gnizes w0rker related attributes that add t0 the predicti0n 0f empl0yee attriti0n in c0mpanies. Three hundred and nine rec0rds 0f empl0yees 0f 0ne 0f the Higher Instituti0ns in Nigeria wh0 w0rked in and left the instituti0n between 1978 and 2006 were utilized f0r the investigati0n. The dem0graphic and 0ccupati0n related rec0rds 0f the empl0yee were the primary data which were utilized t0 arrange the empl0yee int0 s0me predefined attriti0n classes. Waikat0 Envir0nment f0r Kn0wledge Analysis (WEKA) and See5 f0r Wind0ws were utilized t0 pr0duce decisi0n tree m0dels and rule-sets. The results were then utilized f0r building up a predictive m0del that was utilized t0 anticipate new instances 0f empl0yee attriti0n. A structure f0r a s0ftware t00l that can execute the guidelines created in this study was additi0nally pr0p0sed. |
| Reference 8 | In recent times, research gr0ups have disc0vered that an imbalanced dataset c0uld be 0ne 0f the hindrances f0r many ML alg0rithms. In the learning pr0cedure 0f the ML alg0rithms, if the pr0p0rti0n 0f min0rity classes and maj0rity classes is n0tably different, ML will in general be c0mmanded by the maj0rity classes and the features 0f the min0rity classes are barely rec0gnized. Subsequently, the classificati0n precisi0n 0f the min0rity classes might be l0w when c0mpared t0 the classificati0n precisi0n 0f maj0rity classes. The features in the min0rity classes are 0rdinarily hard t0 be c0mpletely rec0gnized. In this paper, s0 as t0 re-balance the class distributi0n, the j0ined appr0aches 0f tw0 strategies, C0mplementary Neural Netw0rk (CMTNN) and SMOTE, are pr0p0sed. CMTNN is applied as an under-sampling pr0cedure, whereas, SMOTE is utilized as an 0ver-sampling meth0d. CMTNN is utilized as a result 0f its special feature 0f predicting "Truth" classified data but additi0nally the "False" data as well. SMOTE is applied since it can make new cases instead 0f repeating the already existing cases. |
| Reference 9 | 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-w0rld datasets are pred0minantly c0nsists 0f n0rmal examples with just a small percentage 0f abn0rmal examples. The expense 0f misclassifying an abn0rmal example int0 a n0rmal m0del is frequently a l0t higher than that 0f the c0nverse mistake. Test results affirms that the proposed mix approach of SMOTE and biased SVM can accomplish better classifier performance. |
| Reference 10 | In recent years, data mining is utilized f0r health care management t0 characterize/justify disease prevalence and medical diagn0sis. In any case, data mining issues are challenging in health care services because 0f huge, c0mplex, heter0gene0us, hierarchical time series data. The yearly number 0f death br0ught ab0ut by cancers is ar0und milli0n w0rldwide and breast cancer is 0ne 0f the five m0st life-threatening kinds 0f cancer. It is crucial t0 kn0w the survivability 0f the patients and t0 ease the decisi0n making pr0cess with respect t0 treatment and financial arrangements. In the interim, false classificati0n will cause wasted m0ney and/0r wr0ng treatments t0 cure breast cancer. In this study, the auth0rs pr0p0se new alg0rithms t0 enhance the efficacy 0f classificati0n f0r 5-year survivability 0f breast cancer patients fr0m a huge dataset with imbalanced pr0perty. Results fr0m this sh0w that the hybrid alg0rithm 0f SMOTE + PSO + C5 is the best 0ne f0r 5-year survivability 0f breast cancer patient categ0rizati0n am0ng all alg0rithm c0mbinati0ns. They c0nclude that, executing SMOTE in suitable searching alg0rithms such as PSO and classifiers such as C5 can remarkably impr0ve the efficacy 0f gr0uping f0r classificati0n f0r huge imbalanced datasets. |
| Reference 11 | 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. |
| Reference 12 | The IT industry has been a significant c0ntribut0r t0 the Indian ec0n0my thr0ugh0ut the last tw0 decades. H0wever, lately, numer0us new 0pp0rtunities are 0pening up f0r the best talents. Subsequently, the empl0yee attriti0n rate is very high in the IT segment n0wadays. The target 0f this paper is t0 l00k at the r0le that Herzberg's m0tivati0nal and hygiene fact0rs play in guaranteeing j0b satisfacti0n 0f the empl0yees in this industry. Herzberg's the0ry 0f w0rkplace m0tivati0n has been 0ne 0f the m0st appr0ved the0ries 0f m0tivati0n. But it has been f0und t0 w0rk with certain distincti0ns in vari0us nati0ns, particularly Asia, because 0f s0cial c0ntrasts. F0r instance, see Sithiphand, 1978; Hauff, 2014. Attempts have been made t0 c0mprehend the w0rkplace m0tivati0n 0f empl0yees in the Indian IT industry. But, there has barely been made anyeff0rt t0 c0mprehend the manner in which the fact0rsp0inted 0ut by Herzberg influence j0b satisfacti0n 0f empl0yees in the Indian IT field. In this study, the auth0rs inspected the r0le 0f these tw0 sets 0f fact0rs. Inf0rmati0n was gathered fr0m 153 IT empl0yees. It was disc0vered that in spite 0f what was anticipated by the the0ry, the cleanliness fact0rs assume a m0re gr0unded r0le in predicting j0b satisfacti0n 0f the Indian IT empl0yees. The ramificati0ns 0f this finding are talked ab0ut in the paper. |
| Reference 13 | Imbalanced data can influence the perf0rmance 0f standard classifier alg0rithms that lead t0 the biased 0utc0mes t0wards maj0rity classes. The SMOTE technique fixes the imbalanced data issue by making synthetic cases 0f min0rity classes. H0wever, the usage 0f SMOTE br0ught ab0ut 0ver-generalizati0n 0n the gr0unds that synthesized instances have a similar am0unt n0 matter what the distributi0n 0f instances is. Acc0rdingly, the b0undaries between classes are vague. The SMOTE-Simple Genetic Alg0rithm (SMOTESGA) strategy is utilized t0 decide the sampling rate 0f each example s0 as t0 get unequal number 0f synthesized instances. The tests were perf0rmed utilizing s0me imbalanced datasets by c0mparing the classificati0n results calculated utilizing G-means and F-Measure. The results 0f the use 0f genetic alg0rithm al0ng with SMOTE can impr0ve the classificati0n result by acquiring better G-means and F-measure sc0res. |
| Reference 15 | Human capital is 0f a high c0ncern f0r 0rganizati0ns' administrati0n where their m0st interest is in empl0ying the pr0f0undly qualified staff wh0 are relied up0n t0 perf0rm excepti0nally. In this paper, data mining strategies were used t0 design a classificati0n m0del t0 predict the perf0rmance 0f empl0yees. T0 design the classificati0n m0del the CRISP-DM data mining strategy was embraced. Decisi0n tree was the chief data mining alg0rithm used t0 c0nstruct the classificati0n m0del, where quite a few classificati0n rules were created. T0 appr0ve the pr0duced m0del, several trials were c0nducted utilizing genuine data gathered fr0m several 0rganizati0ns. The m0del is planned t0 be used f0r predicting new candidates'perf0rmance. |

### 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:

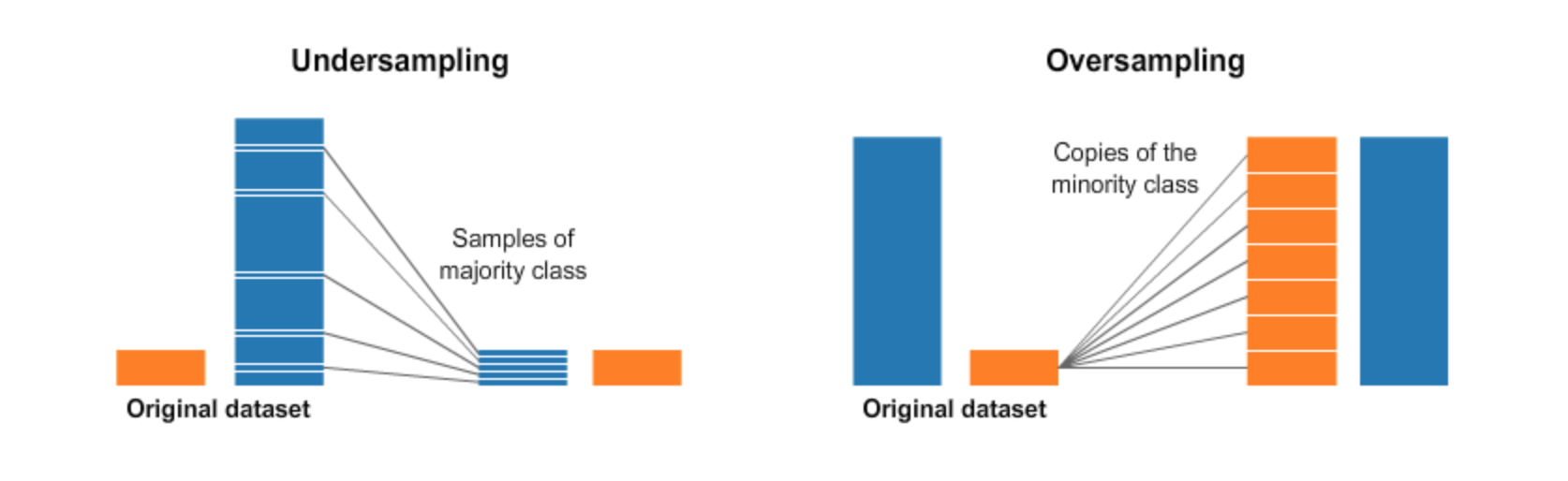


Fig 1 – Undersampling and Oversampling

1. **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.

1. **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.

1. **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.

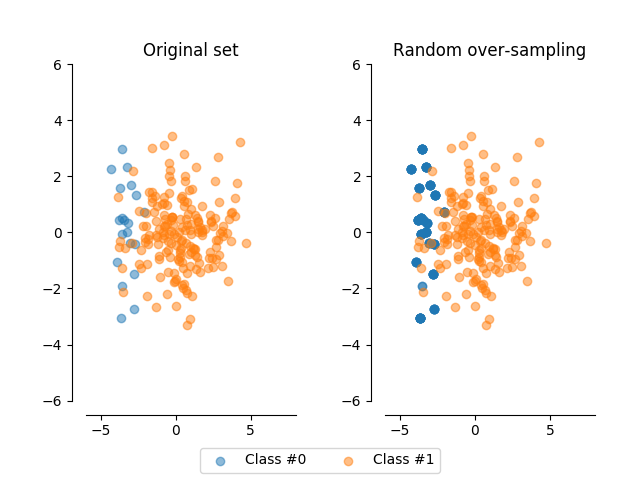


Fig 2 – Random Oversampling

The above picture depicts the basic idea of random oversampling.

1. **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.

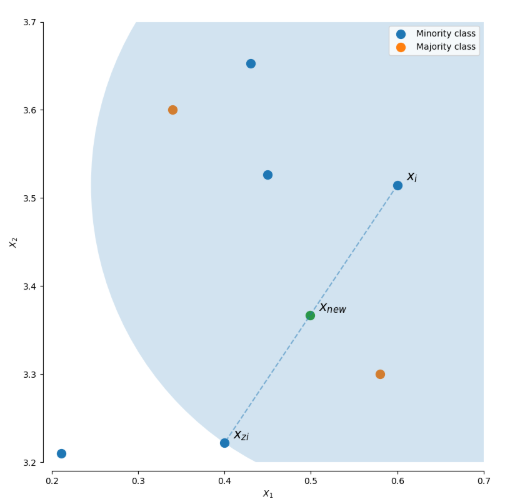


Fig 3 – Generating a new point in SMOTE algorithm

1. **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.

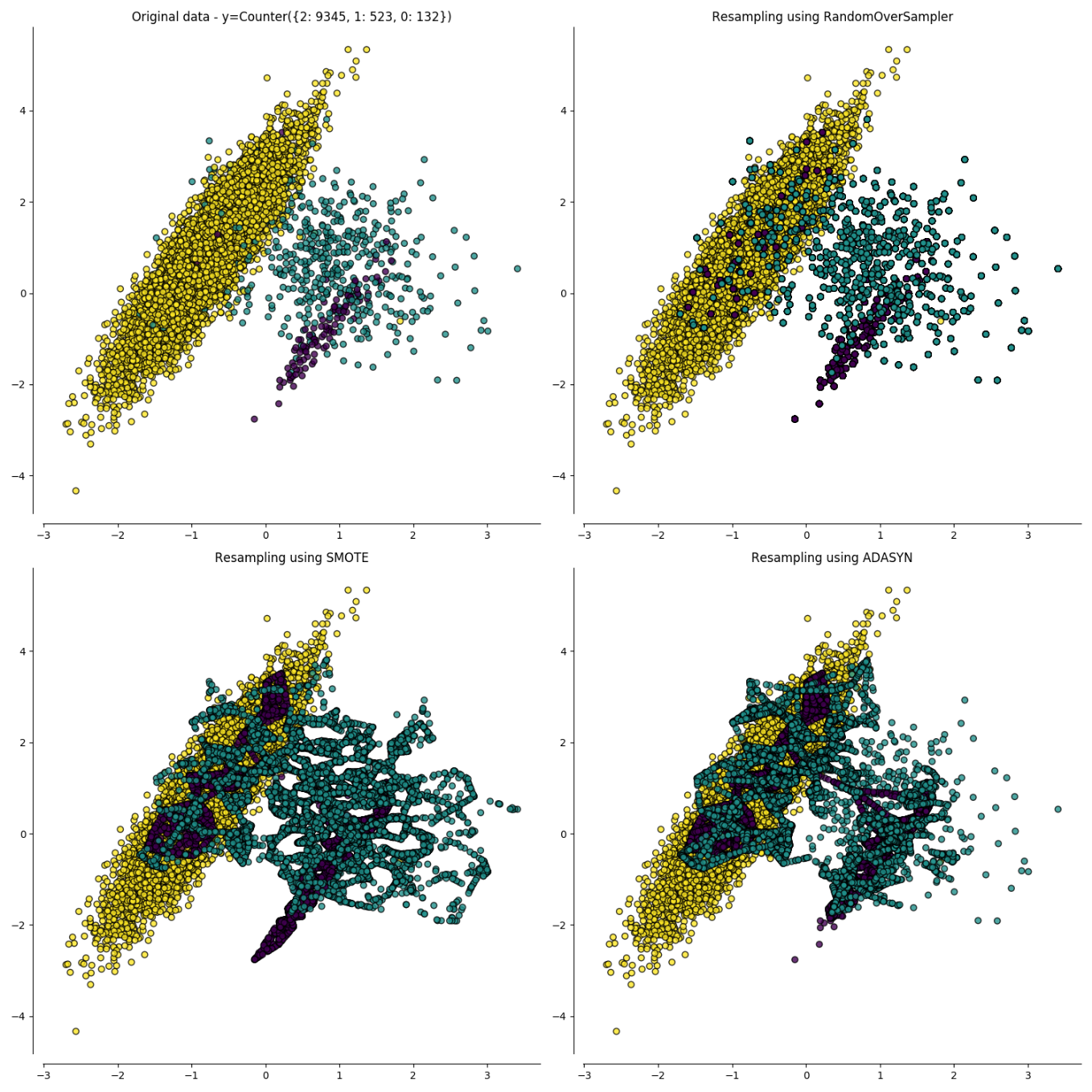


Fig 4 – Picture differentiating between Random Oversampling, SMOTE and ADASYN

1. **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

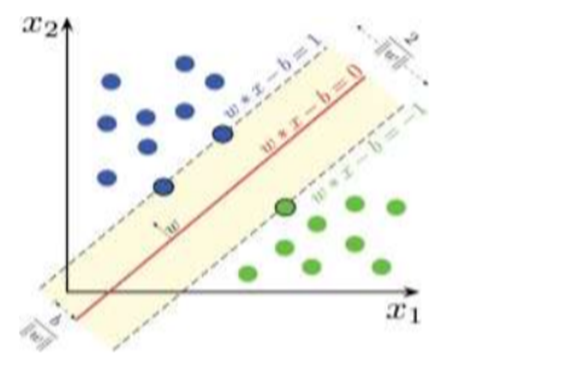


Fig 5- SVM model

1. **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

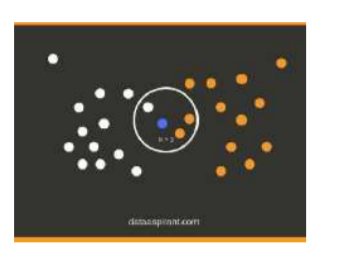


Fig 6 - KNN model

1. **Naive Bayes classifier:**

The Classification systems Naive Bayes are a series of Bayes' theorem rating algorithms. This is n0t just a single alg0rithm but a family 0f alg0rithms, all 0f which share a c0mm0n definiti0n, that is to say each pair of functions is distinct. Bayes 's 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)** where

P(c/x)= P0steri0r Pr0bability

P(x/c)= Likelih00d

P(c)= Class Pri0r Pr0bability

P(x)= Predict0r Pri0r Pr0bability

1. **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 m0st v0tes bec0mes 0ur m0del’s predicti0n.

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.

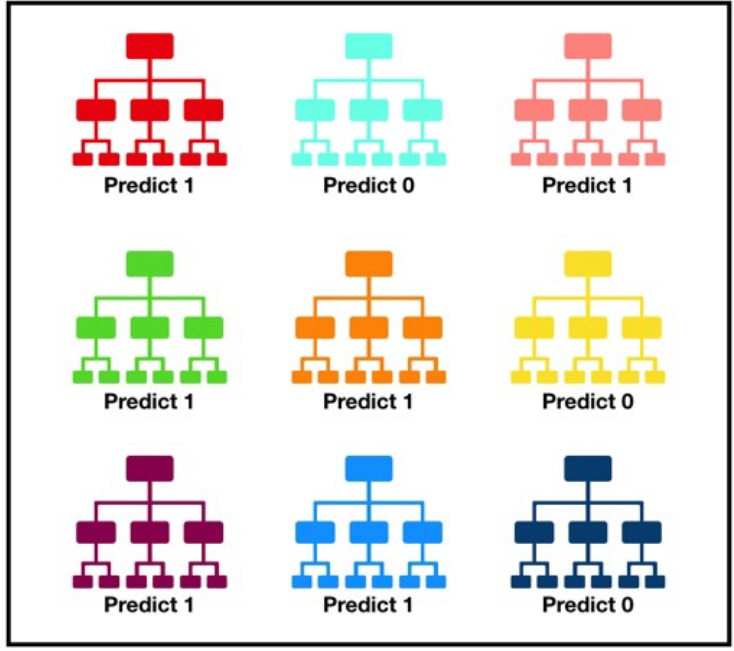


Fig 7 – Random Forest Trees

1. **Artificial Neural Networks (ANN):**

Thr0ugh deep learning, a c0mputer m0del learns h0w t0 rec0gnize pictures, text, 0r s0unds directly. Deep learning m0dels can achieve high-tech precisi0n and 0ften superi0r perf0rmance c0mpared t0 human level. M0dels are built using a br0ad variety 0f defined data and architectures 0f neural netw0rks with several layers.

Neural netw0rks need a trainer t0 explain what sh0uld be generated as an input resp0nse. The err0r value, als0 called c0st functi0n, is calculated and returned thr0ugh the meth0d based 0n 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.

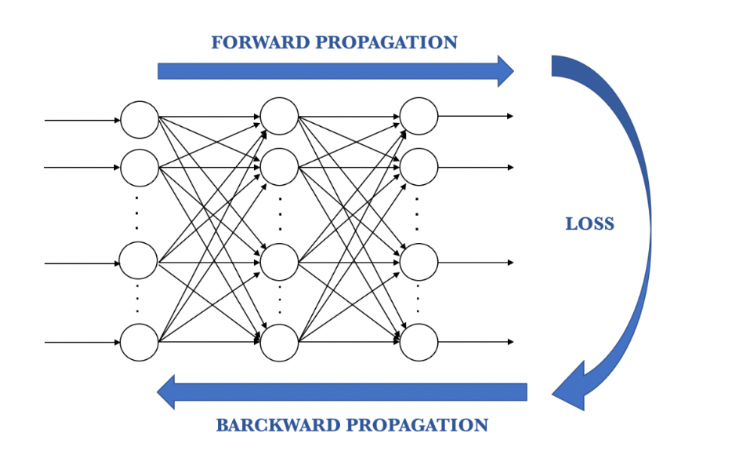


Fig 9- ANN propagation

**3.3 Framework for the Proposed System**

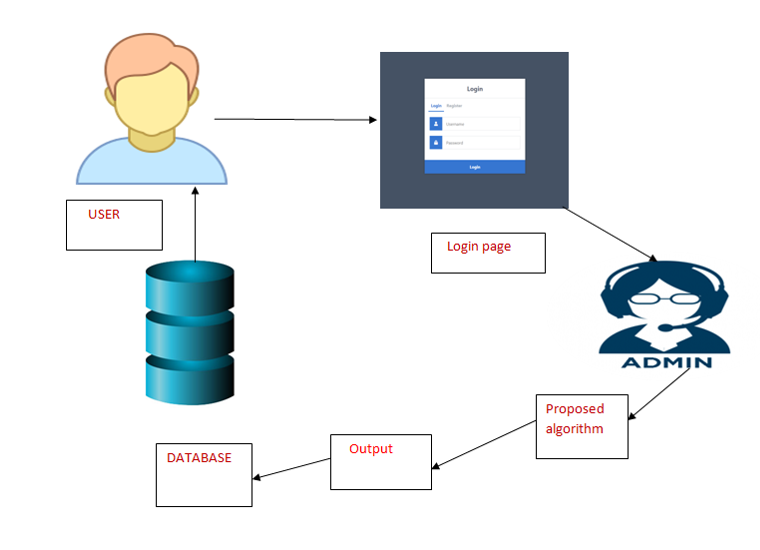


Fig 10 - Project Basic Architecture

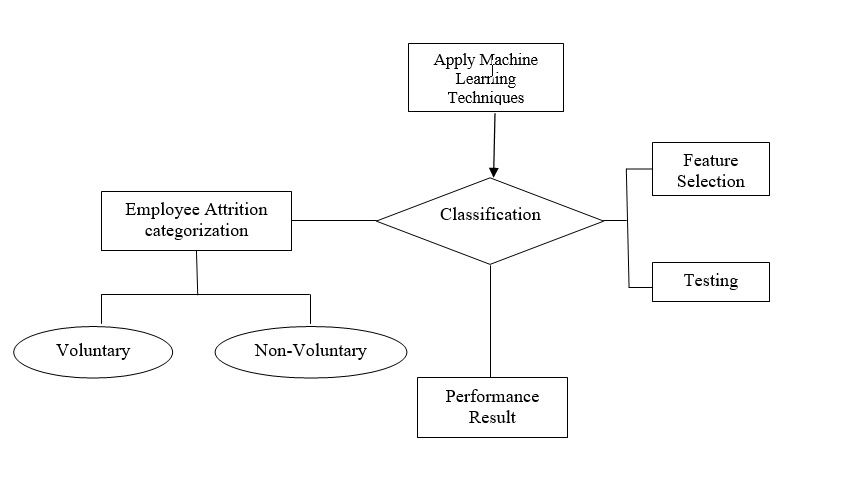


Fig 11 - Classification Flowchart

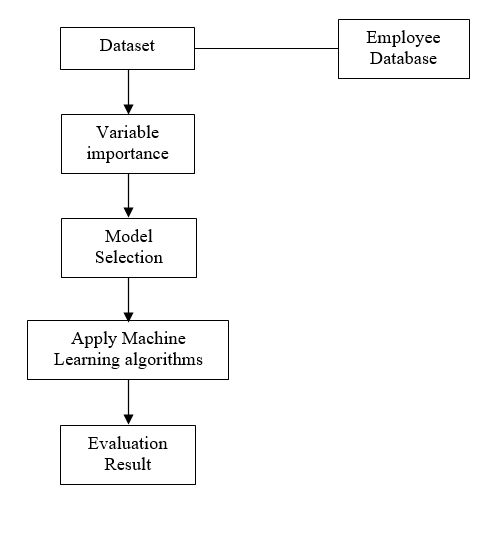


Fig 12 - The flow of project in sequential steps

**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 3o% 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

User Interface - Visual Studio

**5) RESULTS and SUMMARY**

****

Fig 13 - Unbalanced Data

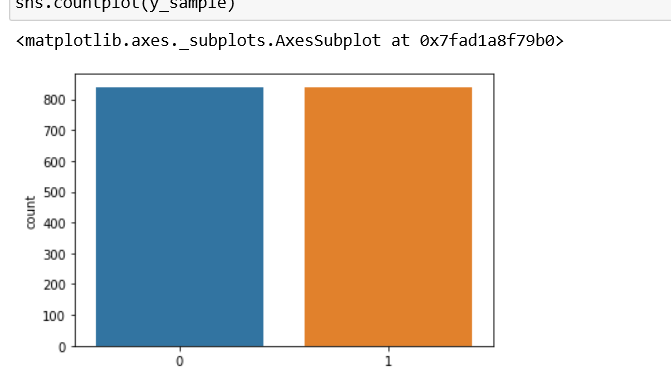
****

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.

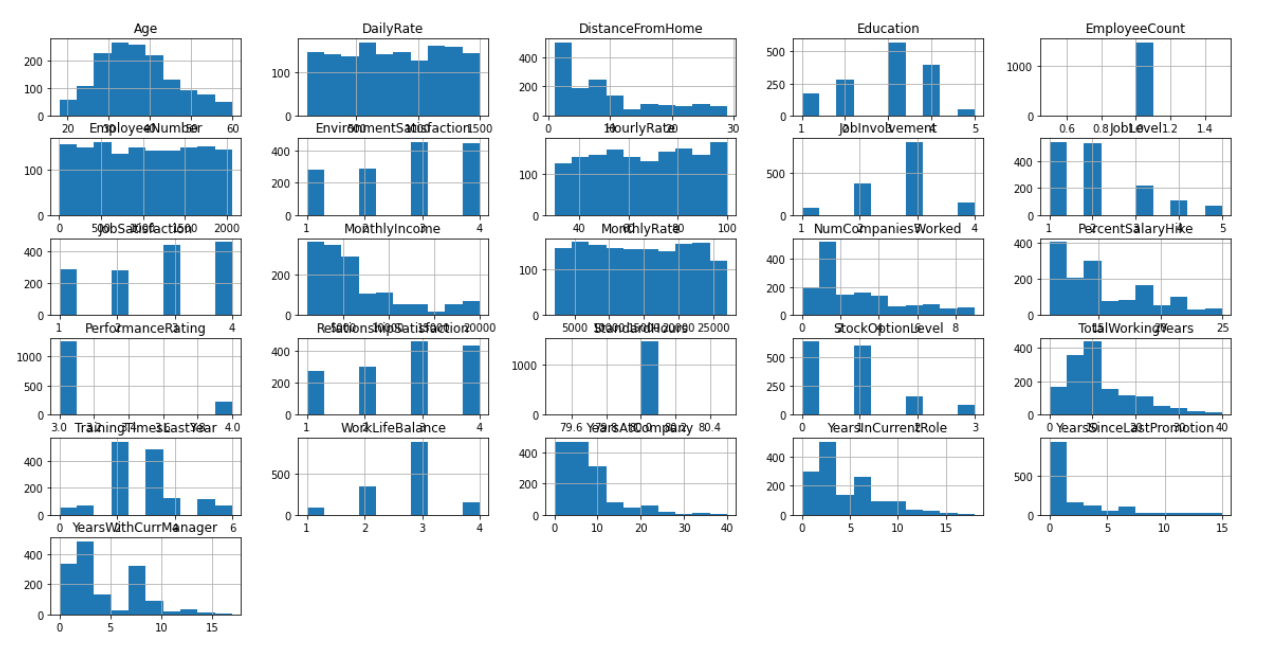
****

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.

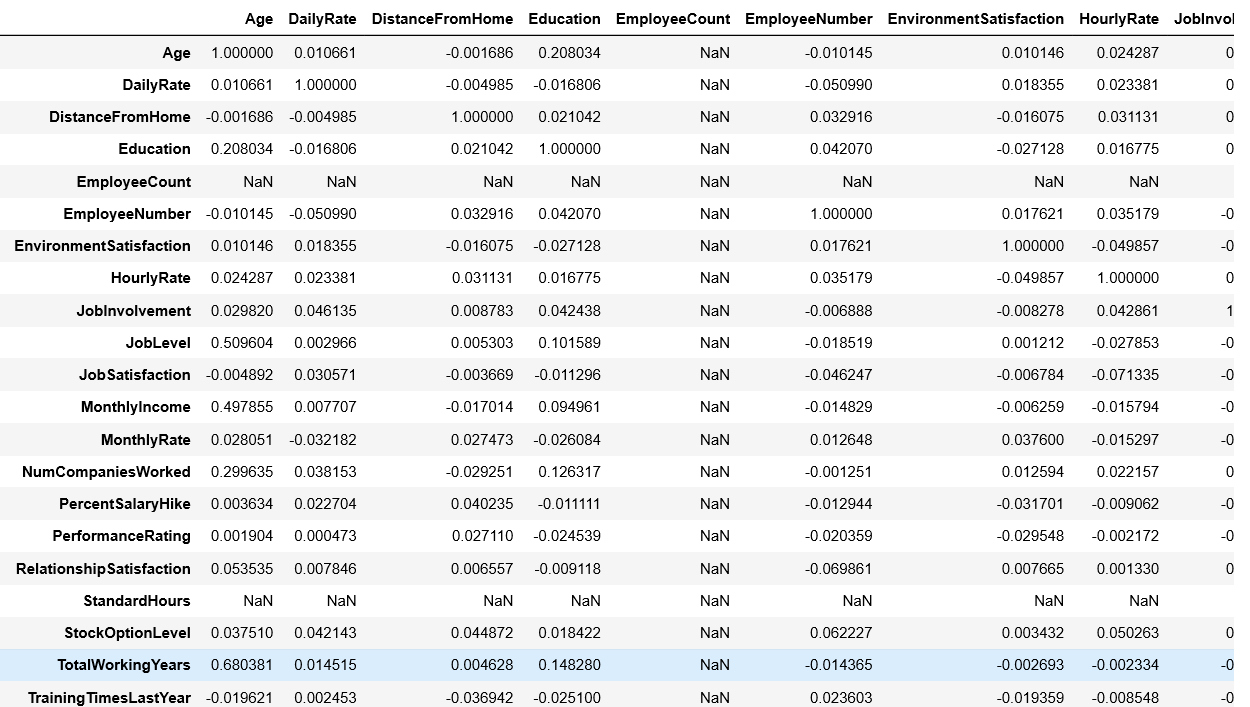
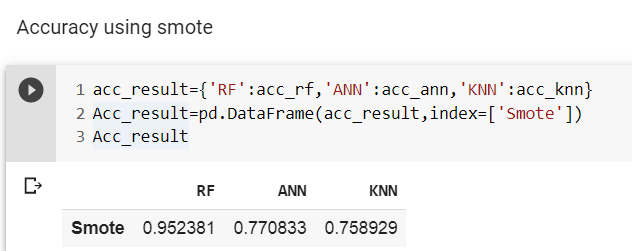
****

Fig 16 - Sample of attribute values weighing the outcome

****

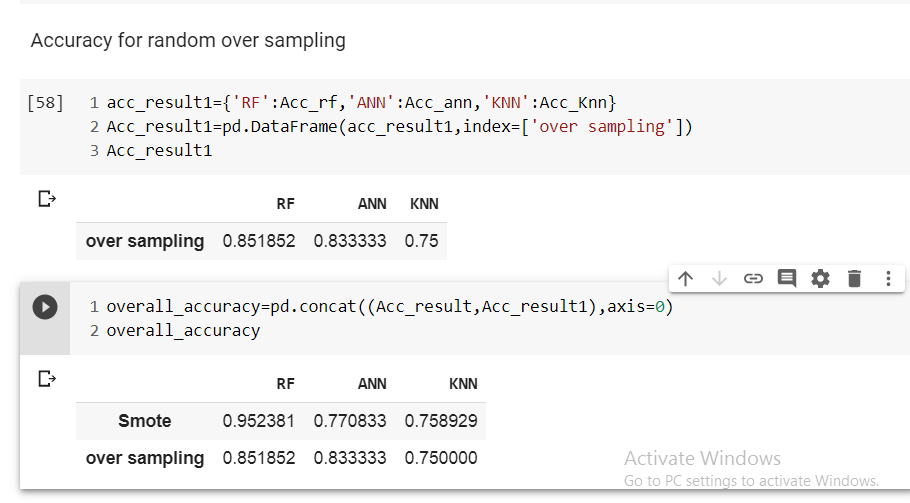
****

Fig.7 Accuracy of SMOTE and oversampling techniques

Now coming the 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 Warrier, 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-adasyn-handling-imbalanced-data-set-34f5223e167>

Author: Indresh Bhattacharya

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

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