**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 i

[Executive Summary ii](#_TOC_250004)

Table of Contents iii

[List of Figures ix](#_TOC_250003)

[List of Tables xiv](#_TOC_250002)

Abbreviations xvi

[Symbols and Notations xix](#_TOC_250001)

1. [INTRODUCTION 1](#_TOC_250000)0
   1. Objective 10
   2. Motivation 11
   3. Background 11
2. Literature Survey 17
3. PROJECT DESCRIPTION TECHNICAL SPECIFICATION 17
4. **DESIGN APPROACH AND DETAILS** 18
   1. Design Approach / Materials & Methods 19
   2. Codes and Standards 20
   3. Constraints, Alternatives and Tradeoffs .

#### SCHEDULE, TASKS AND MILESTONES .

1. **PROJECT DEMONSTRATION** .
2. **COST ANALYSIS / RESULT & DISCUSSION** .

|  |  |  |
| --- | --- | --- |
| 8 | **SUMMARY** | . |
| 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

### Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Paper** | **Description** | **Pros** | **Cons** |
| ***Machine learned job recommendation AUTHORS: IoannisPaparrizos,B.BarlaCambazoglu,Aristides Gionis*** | The highly competitive and dynamic nature of the job market as well as personal preferences and goals lead individuals to change their jobs frequently in their lives. They trained a machine learning model using a large number of job transitions extracted from the publicly available employee profiles in the Web. The results of their experiments demonstrate that the transition of an employee to an institution can be quite accurately predicted, significantly improving over a baseline predictor that always predicts the most frequent institution in the data. The results indicate that the most important feature in predicting a job transition is the current institution of the employee. | Uses a Decision table/Naive Bayes hybrid classifier. | Does not take into account the social aspects that could affect job transitions. |
| Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear ***Embedding AUTHORS: Juanjuan Wang; Mantao Xu; Hui Wang; Jiwu Zhang*** | Imbalanced data classification often arises in many practical applications in the context of medical pattern recognition and data mining. MOst of the existing classification approaches are well developed by assuming the underlying training set is evenly distributed. HOwever, they are faced with a severe bias problem when the training set is a highly imbalanced distribution thus leading to poor performance. SMOTE is an important approach by oversampling the positive class or the minority class. | Use of LLE algorithm along with SMOTE so that oversampling works on datasets that are non-linearly separable.  This paper experiments on 3 different datasets using 3 different classifiers. |  |
| ***Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm AUTHORS: PiyasakJeatrakul, KokWai Wong, and Chun Che Fung*** | In recent years, many research groups have found that an imbalanced data set could be one of the obstacles for many Machine Learning algorithms In this paper, in order to re-balance the class distribution, the combined approaches of two techniques, Complementary Neural Network (CMTNN) and SMOTE, are proposed. While CMTNN is applied as an under-sampling technique, SMOTE is used as an over-sampling technique. CMTNN is used because of its special feature of predicting not only the "truth" classified data but also the "false" data. SMOTE is applied because it can create new instances rather than replicate the existing instances. | The proposed method is implemented on 4 different datasets, each with different number of attributes, to check for consistency.  Two different CMTNN variations are used. | In some cases, the CMTNN removes too many instances and the remaining instances are not enough for learning algorithms to obtain correct results. |
| ***Combination approach of SMOTE and biased-SVM for Imbalanced datasets AUTHORS: He-Yong Wang*** | Imbalanced data learning is problematic as traditional machine learning approaches fail to provide satisfactory results due to skewed class distribution. The cost of misclassifying an abnormal example into a normal example is often much higher than that of the reverse error | The proposed method is implemented on 4 different datasets, each with different number of attributes, to check for consistency. |  |
| ***A hybrid classifier combining SMOTE with PSo to estimate 5- year survivability of breast cancer patients AUTHORS : Kung-JengWanga, BunjiraMakonda, Kun-Huang Chena, Kung-Min Wang*** | Data mining is a process to discover useful information through a large amount of data. This process is widely applied in medical, social science, management, engineering, and many other fields. In recent years, data mining is used for health care management to classify/justify disease prevalence and medical diagnosis. HOwever, data mining problems are challenging in health care due to large, complex, heterogeneous, hierarchical time series data. The annual number of deaths caused by cancers is around million worldwide and breast cancer is one of the five most life threatening types of cancer. It is essential to know the survivability of the patients and to ease the decision making process regarding medical treatment and financial preparation. Meanwhile, false classification will cause wasted money and/or inappropriate treatments to cure the breast cancer. | This paper tries implementing and comparing different combinations of algorithms and gives the best combination to be SMOTE + PSo+ C5 decision tree. |  |
| ***Predicting employee attrition using machine learning algorithms and analyzing reasons for attrition*** | The major aim  of the paper is to analyze the employee dataset of an  organization and find the reasons, why the best and most  experienced employees leave the company prematurely  and also try to predict which valuable employees are  probable to leave the organization subsequently | Predicting possible employee attrition to avoid loss of talented employees and avoid a gap in the project and the company as a whole | This project has been executed based on the small  dataset, but the real company dataset can be expected to  be huge and with a greater number of attributes |
| ***Motivators, Hygiene Factors and Job Satisfaction of Employees in IT Sector in India*** | The objective of this paper is to examine the role that Herzberg's motivational and hygiene factors play in ensuring job satisfaction of the employees in this sector | This project identifies the factors that define the particular individuals motivation to work and gives us an idea on factors that defines the employee attrition | The dataset which has been used for this paper is quite small and it is also not reliable hence the results aren’t consistent |
| ***The Implementation of Genetic Algorithm in Smote (Synthetic Minority oversampling Technique) for Handling Imbalanced Dataset Problem*** | The objective of the paper is to implement an oversampling SMOTE algorithm which overcomes the imbalanced masses by creating synthetic instances of minority classes | This paper presents a genetic algorithm for handling imbalance in datasets. This algorithm generates synthetic minority data points based on the difficulty in learning a sample point and the performance improvement achieved by oversampling it | other heuristics which can lead to better results have not been considered. A promising avenue of research is to investigate the effectiveness of GenSample by combining it with ensemble methods. |
| ***A Review of Class Imbalance Problem*** | This paper proposes a general survey for class imbalance problem solutions and the most significant investigations recently introduced by researchers. | This paper summarizes various solutions for dealing with class imbalance problems. This paper also helps in solving the imbalance problem and have proposed various approaches. | May remove significant patterns and cause loss of useful information. Introducing additional computational cost It may lead to over-fitting |
| ***Employee churn prediction*** | Employee churn prediction which is closely related to customer churn prediction is a major issue of the companies. Despite the importance of the issue, there is few attention in the literature about. In this paper, they applied well-known classification methods including, Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes methods on the HR data and then analysed the results by calculating the accuracy, precision, recall, and F-measure values of the results. | The results of this study demonstrate that data mining algorithms can be used to build reliable and accurate predictive models for employee churn. The problem of churn prediction is not just to identify churners from no churners | The results of this paper are based upon quite a few assumptions and are also drawn out based on few inputs. |
| *Analyzing employee attrition using decision tree algorithms* | This study identifies employee related attributes that contribute to the prediction of employees’ attrition in organizations | Results obtained from the study shows that employee Salary and Length of service were determining factors for predicting employee attrition in the institution whose data was used for the case study. | The study has a picked up a case study based which might not co-relate to all the scenarios around the world. So the results may be inconsistent |
| *An exploratory study of us lodging properties' organizational practices on employee turnover and retention* | The purpose of this paper is to investigate US lodging properties’ organizational employee‐retention initiatives and practices, and to examine the impact of those initiatives on employee turnover and retention. | The findings reveal that Corporate Culture, Hiring and Promotions and Training practices influence non‐management employee retention. At the same time, Hiring and Promotion practices impact management retention, as well. | he findings will equip lodging organizations and industry professionals with the contemporary tools to proactively reduce employee turnover and for maintaining employee retention. This should have a |

### 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:

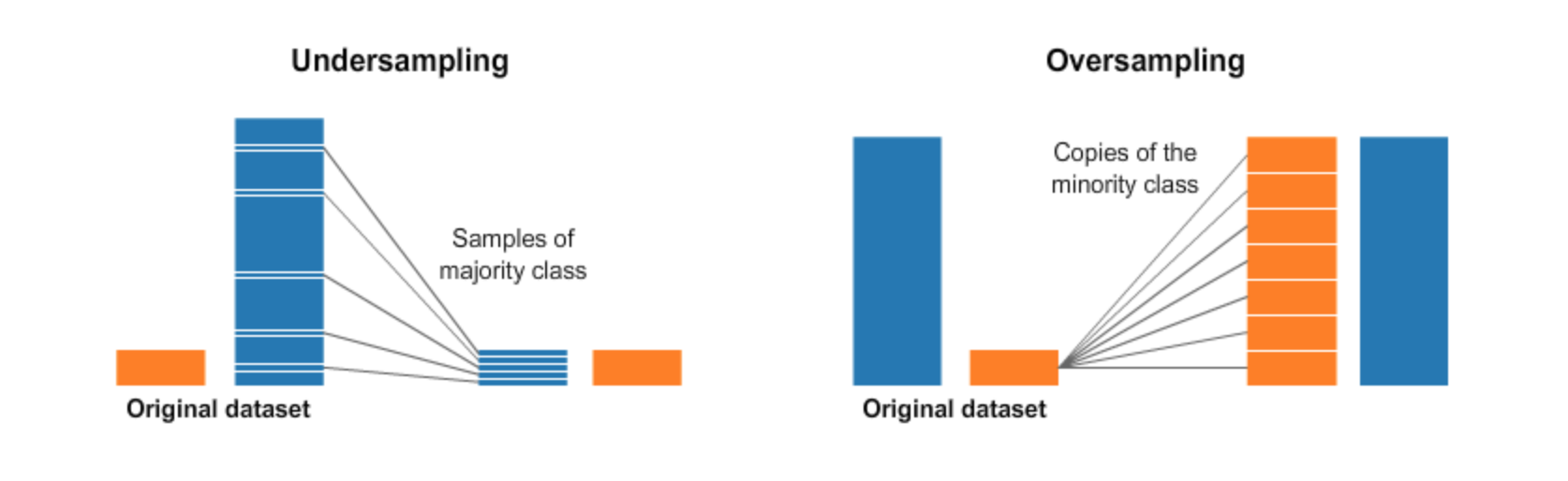


Fig 1 – Undersampling and Oversampling

1. **Undersampling:**

Under sampling alludes to a group of procedures intended to adjust the class distribution for a classification dataset that has a slanted class appropriation. An imbalanced 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 majority class and class 1 is the minority class. 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. Commonly, under sampling strategies are utilized related to an oversampling procedure for the minority class, and this mix frequently brings about preferred execution over utilizing oversampling or under sampling alone on the preparation dataset.

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.

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.

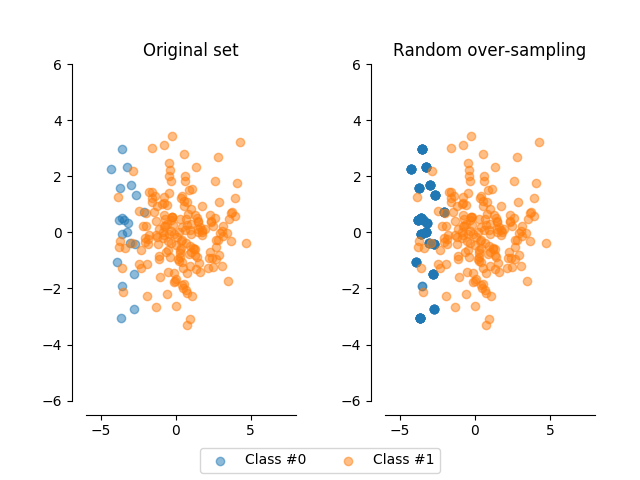


Fig 2 – Random Oversampling

The above picture depicts the basic idea of random oversampling.

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.

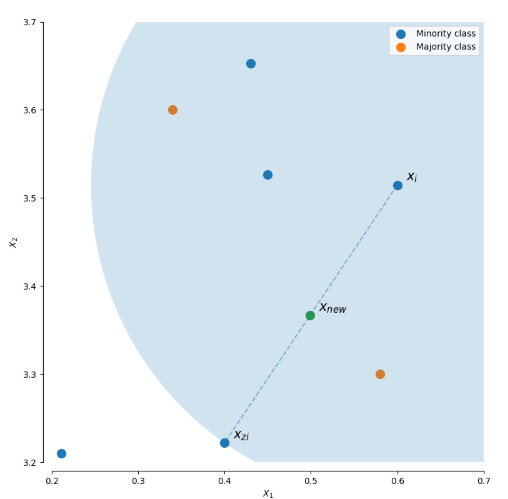


Fig 3 – Generating a new point in SMOTE algorithm

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.

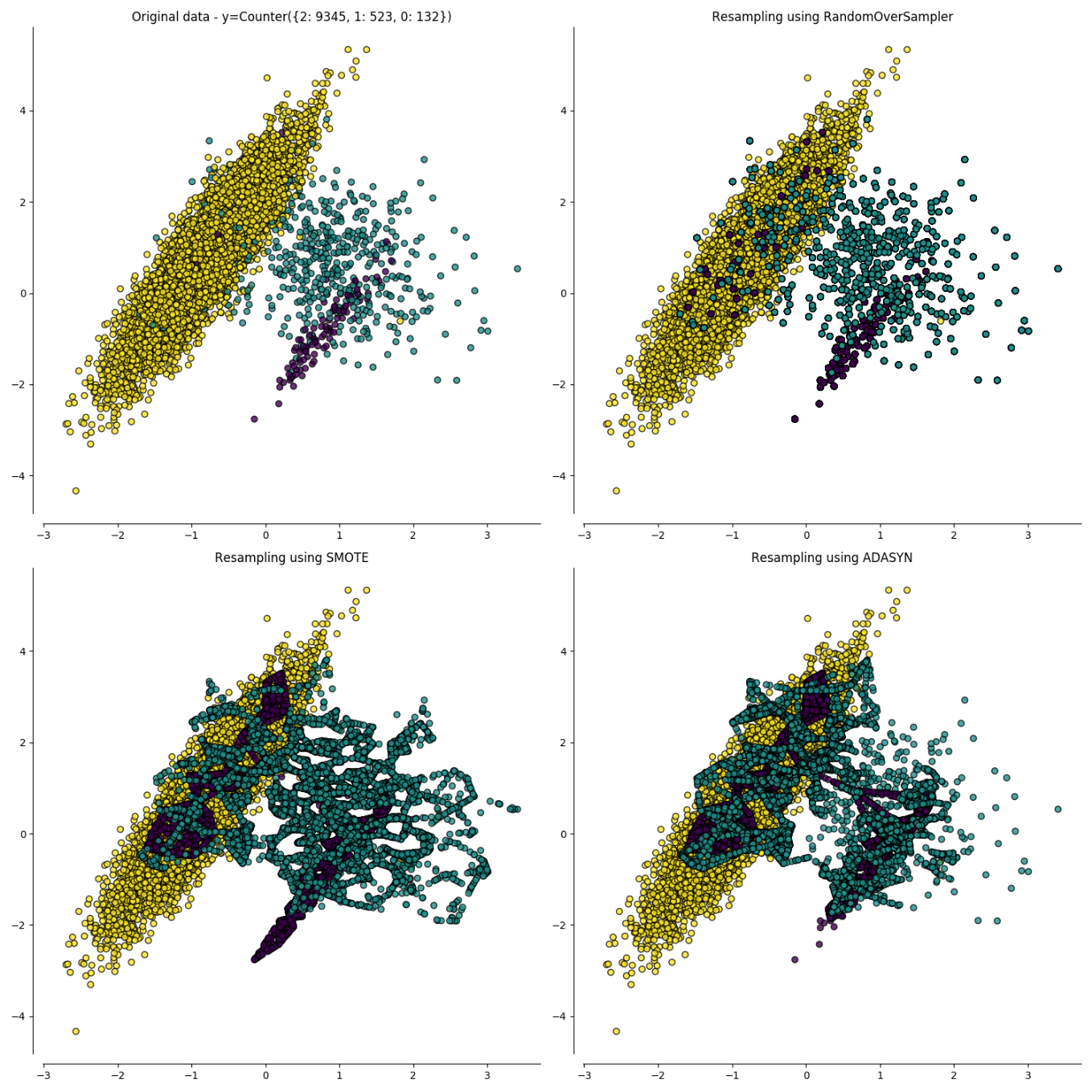


Fig 4 – Picture differentiating between Random Oversampling, SMOTE and ADASYN

**Support Vector Machine (SVM):**

Support Vector Machine (SVM) is one of the popular 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

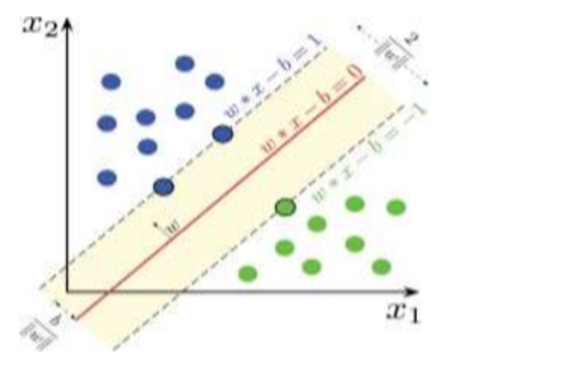


Fig 5- SVM model

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

K-Nearest Neighbors is one of the most fundamental but essential machine learning classification algorithms. It is part of the supervised learning domain and is used intensively in pattern recognition, intrusion detection and data collection. It is very widely available in real-life scenarios as it is non-parametric, which means that it makes no underlying assumptions about data distribution (unlike other algorithms, such as GMM, which 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 simple algorithm which stores all existing cases and classifies new cases on the basis of an equivalence (e.g. function of distance). As early as in the early 1970s, KNN was employed as a non-parametric technique in statistical estimation and pattern recognition.

F=Fig

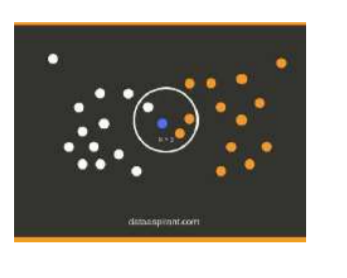


Fig 6 - KNN model

**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 's theorem determines the probability of a happening because of the possibility of another occurrence. Bayes’ theorem is stated mathematically as the following equation.

**P(c/x)=(P(x/c)P(c) )/P(x)** where

P(c/x)= Posterior Probability

P(x/c)= Likelihood

P(c)= Class Prior Probability

P(x)= Predictor Prior Probability

**Random Forest Classifier:**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds that is **“A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.”** The low correlation between models is the key.The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don’t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

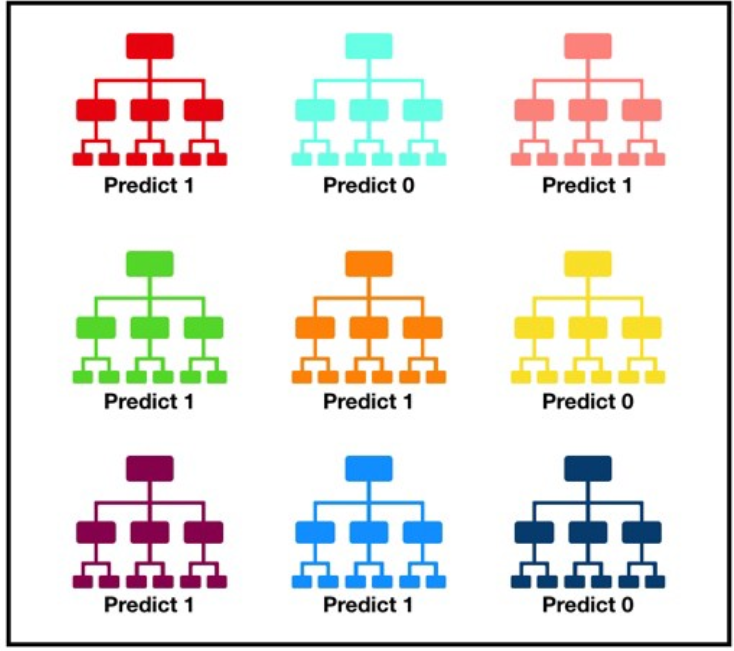


Fig 7 – Random Forest Trees

**Artificial Neural Networks (ANN):**

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.**“Artificial Neural Networks or ANN is an information processing paradigm that is inspired by the way the biological nervous system such as brain process information. It is composed of large number of highly interconnected processing elements (neurons) working in unison to solve a specific problem.”.**

Neural networks require a trainer in order to describe what should have been produced as a response to the input. Based on the difference between the actual value and the predicted value, an error value also called **Cost Function** is computed and sent back through the system. “Cost Function: One half of the squared difference between actual and output value”.

For each layer of the network, the cost function is analyzed and used to adjust the threshold and weights for the next input. Our aim is to minimize the cost function. The lower the cost function, the closer the actual value to the predicted value. In this way, the error keeps becoming marginally lesser in each run as the network learns how to analyze values.We feed the resulting data back through the entire neural network.

As long as there exists a disparity between the actual value and the predicted value, we need to adjust those weights. Once we tweak them a little and run the neural network again, A new Cost function will be produced, hopefully, smaller than the last. We need to repeat this process until we scrub the cost function down to as small as possible.

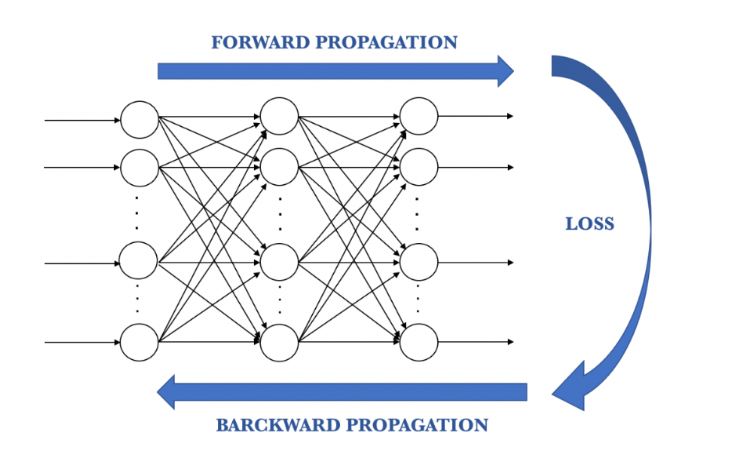


Fig 9- ANN propagation

**3.3 Framework for the Proposed System**

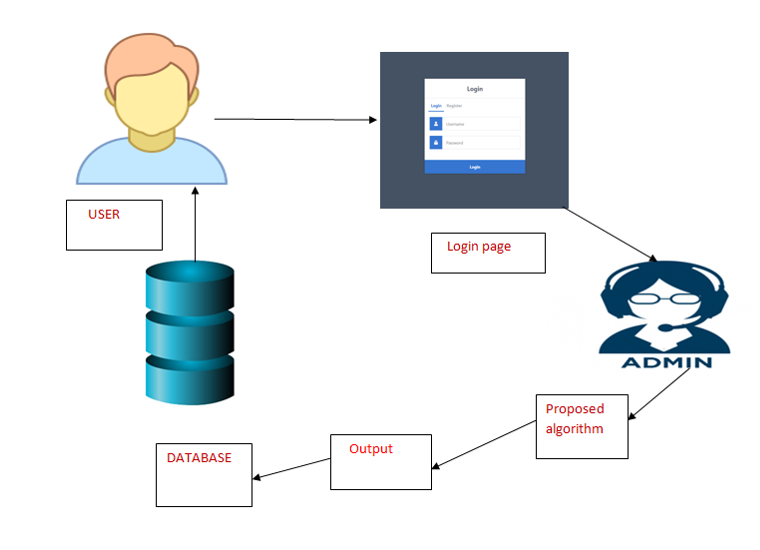


Fig 10 - Project Basic Architecture

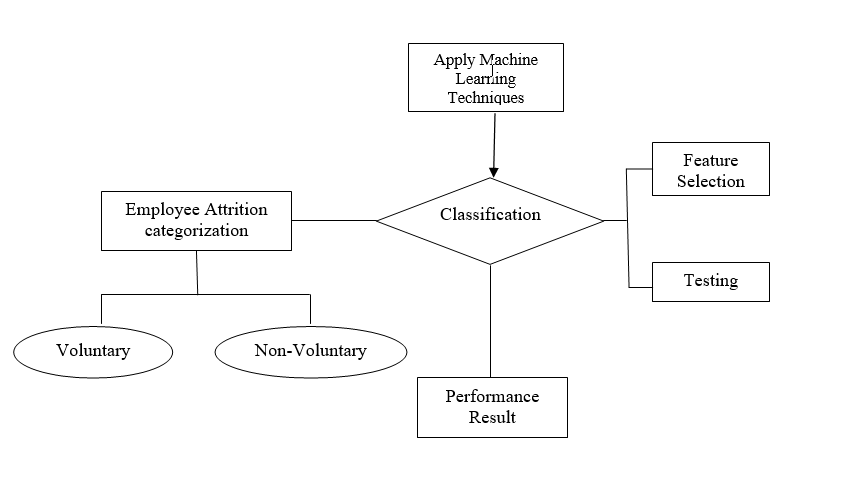


Fig 11 - Classification Flowchart

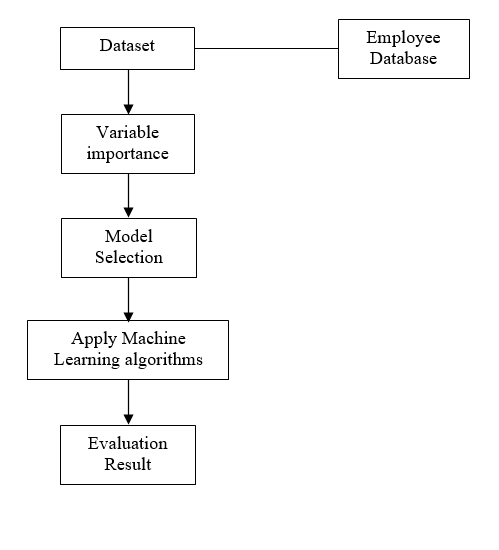


Fig 12 - The flow of project in sequential steps

**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 attempted to anticipate employee attrition by utilizing the first imbalanced dataset. In the subsequent trial, they have introduced the adaptive synthetic sampling approach to solve the class imbalance problem. This methodology included oversampling the minority class which was for this situation the "yes" class. The third experiment involved random under sampling of the data where they have selected an equal subset of each class arbitrary. Besides, each test included preparing and approving a lot of classifiers to anticipate the attrition.

1. **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

1. **RESULTS and SUMMARY**

****

Fig 13 - Unbalanced Data

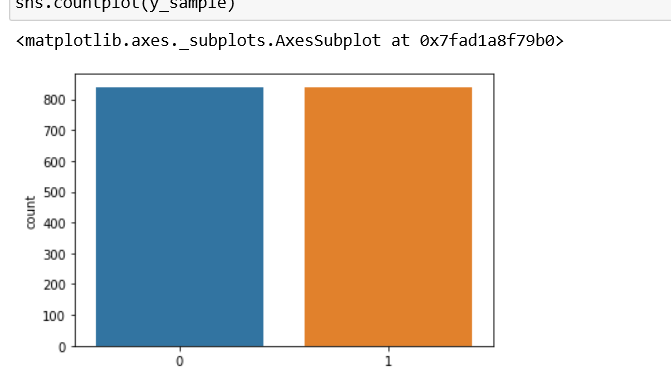
****

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.

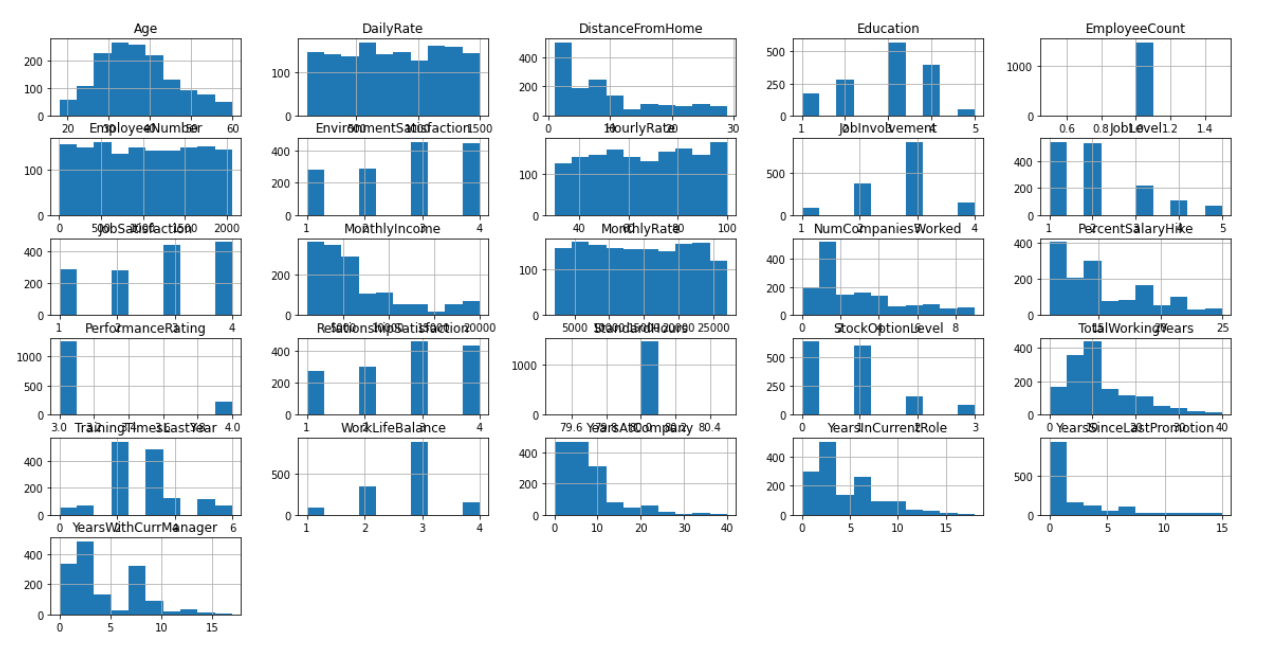
****

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.

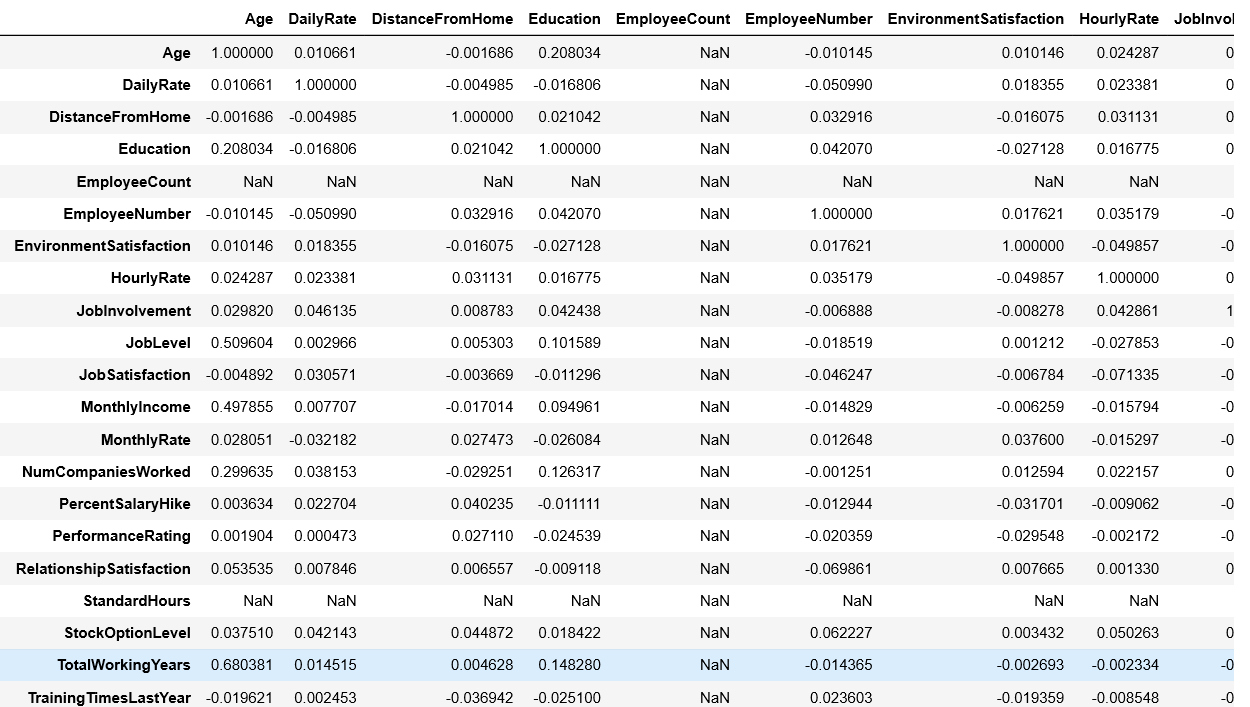
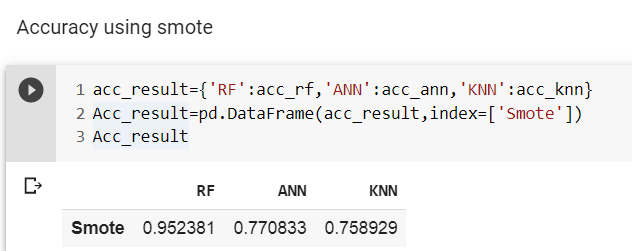
****

Fig 16 - Sample of attribute values weighing the outcome

****

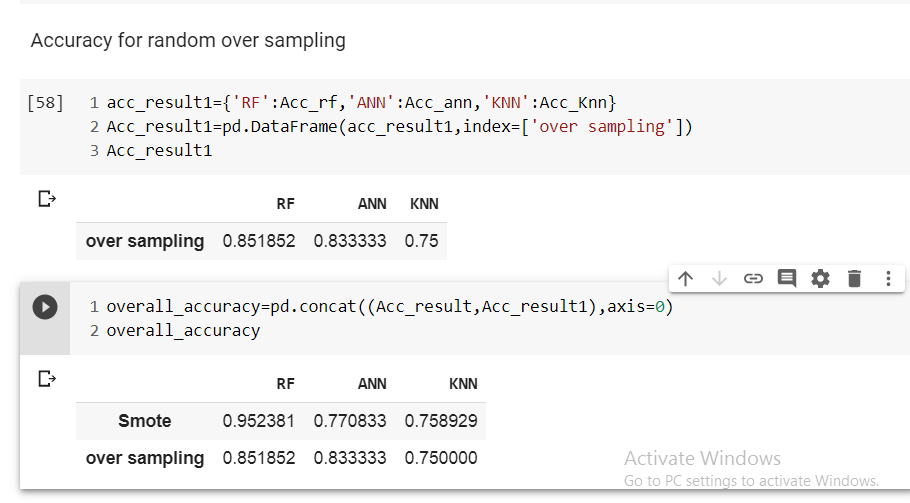
****

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 oversampling technique we had the following accuracies:

Random Forest: 0.851852

ANN: 0.833333

KNN: 0.75

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. Random oversampling combined with KNN has the least accuracy.
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.

**5) 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/>