PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING

Abstract : The growing interest in machine learning among business leaders and decision makers demands that researchers explore its use within business organisations. One of the major issues facing business leaders within companies is the loss of talented employees. 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 first experiment involved training the original class-imbalanced dataset with the following machine learning models: support victor machine (SVM) with several kernel functions, random forest and Knearest neighbour (KNN). The second experiment focused on using adaptive synthetic (ADASYN) approach to overcome class imbalance, then retraining on the new dataset using the abovementioned machine learning models. The third experiment involved using manual undersampling of the data to balance between classes. As a result, training an ADASYN balanced dataset with KNN (K = 3) achieved the highest performance, with 0.93 F1-score. Finally, by using feature selection and random forest, F1-score of 0.909 was achieved using 12 features out of a total of 29 features.

Introduction:

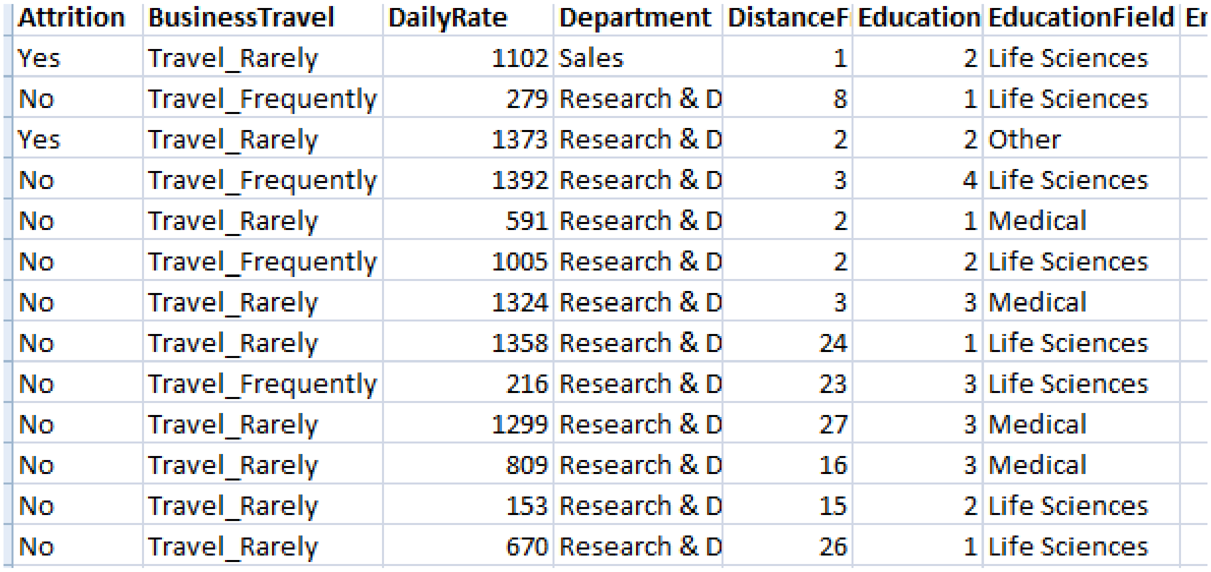
Employee attrition can be defined as the loss of employees due to any of the following reasons: personal reasons, low job satisfaction, low salary, and a bad business environment.

Predicting employees attrition at a company will help management act faster by enhancing their internal policies and strategies. Where talented employees with a risk of leaving can be offered several propositions, such as a salary increase or proper training, to reduce their likelihood of leaving. Using machine learning models can help companies predict employees attrition. Using the historical data kept in human resources (HR) departments, analysts can build and train a machine learning model that can predict the employees who are leaving the company. Such models are trained to examine the correlation between the features of both active and terminated employees.

Dataset :

Dataset used for this paper is IBM HR Analytics Employee Attrition and Performance dataset. It is a fractional dataset created by IBM data scientists which contains the comparable data of employee performance and attrition. The data set contains 35 features like employee age, job role, marital status etc. One of the columns is job satisfaction which is numeric value of rating ranges between 1 to 4. We are going to predict this feature with our proposed algorithms.

The sample of the dataset is shown bellow.



Attrition	BusinessTravel	DailyRate	Department	DistanceF	Education	EducationField	Er
Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	
No	Travel_Frequently	279	Research & D	8	1	Life Sciences	
Yes	Travel_Rarely	1373	Research & D	2	2	Other	
No	Travel_Frequently	1392	Research & D	3	4	Life Sciences	
No	Travel_Rarely	591	Research & D	2	1	Medical	
No	Travel_Frequently	1005	Research & D	2	2	Life Sciences	
No	Travel_Rarely	1324	Research & D	3	3	Medical	
No	Travel_Rarely	1358	Research & D	24	1	Life Sciences	
No	Travel_Frequently	216	Research & D	23	3	Life Sciences	
No	Travel_Rarely	1299	Research & D	27	3	Medical	
No	Travel_Rarely	809	Research & D	16	3	Medical	
No	Travel_Rarely	153	Research & D	15	2	Life Sciences	
No	Travel_Rarely	670	Research & D	26	1	Life Sciences	

Algorithms :

In this paper 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 comparision.

Modules:

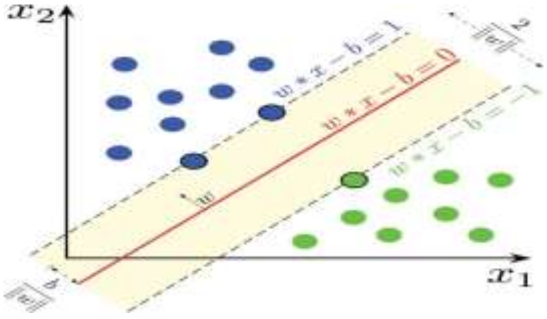
Support Vector Machine (SVM) :

Support Vector Machine (SVM) is one of the standard machine learning algorithms used in pattern recognition, spam filtering and anomaly network intrusion detection. SVM can learn the pattern in gives accurate classification by using class labels. The accurate classification is achieved by training machine to classify unknown samples with the training dataset model. SVM has the capability to find the global optimal solution by performing the linear separation finding an optimal hyperplane that separates two classes. The closest data to the hyperplane are support vectors and by getting the features the predicted class is declared.

Given the training dataset of n points of the form

$(\overrightarrow{x1}, y1), \ldots, (\overrightarrow{xn}, yn)$
Where “yi” are either 1 or -1 each which indicates the class to the point $\vec{xi}$ belongs.

The hyper lane can be written as the set of points xi satisfying w.x-b=0

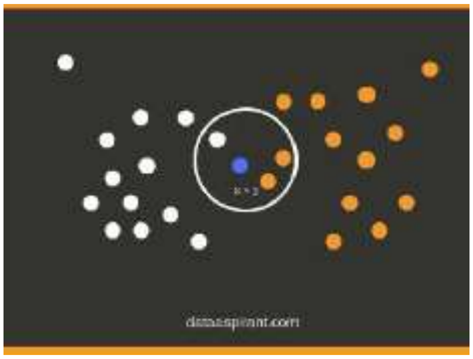


KNN (K Nearest Neighbours) Classifier :

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).
We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.



Naive Bayes classifier :

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$\underset{\text{Posterior Probability}}{P(c|x)} = \frac{\overset{\text{Likelihood}}{P(x|c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

Proposed Methodology:

Data Preprocessing :

The dataset is highly imbalanced and it is converted to balanced data by using upsampling 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.

FRONT END

DJANGO

Django is a free and open source web application framework written in Python. A framework is nothing more than a collection of modules that make development easier. They are grouped together, and allow you to create applications or websites from an existing source, instead of from scratch.

This is how websites - even simple ones designed by a single person - can still include advanced functionality like authentication support, management and admin panels, contact forms, comment boxes, file upload support, and more. In other words, if you were creating a website from scratch you would need to develop these components yourself. By using a framework instead, these components are already built, you just need to configure them properly to match your site.

BACKEND:

ANACONDA

It is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

Anaconda distribution comes with more than 1,500 packages as well as the Conda package and virtual environment manager. It also includes a GUI, **Anaconda Navigator**, as a graphical alternative to the Command Line Interface (CLI).

Anaconda Navigator is a desktop Graphical User Interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator

- Jupyter Notebook
- QtConsole
- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

VISUAL STUDIO

Microsoft **.NET** is a set of Microsoft software technologies for rapidly building and integrating XML Web services, Microsoft Windows-based applications, and Web solutions. The .NET Framework is a language-neutral platform for writing programs that can easily and securely interoperate. There's no language barrier with .NET: there are numerous languages available to the developer including Managed C++, C#, Visual Basic and Java Script. The .NET framework provides the foundation for components to interact seamlessly, whether locally or remotely on different platforms. It standardizes common data types and communications protocols so that components created in different languages can easily interoperate.

".NET" is also the collective name given to various software components built upon the .NET platform. These will be both products (Visual Studio.NET and Windows.NET Server, for instance) and services (like Passport, .NET My Services, and so on).

Microsoft **VISUAL STUDIO** is an Integrated Development Environment (IDE) from Microsoft. It is used to develop computer programs, as well as websites, web apps, web services and mobile apps

LITERATURE SURVEY:

1. TITLE: HISTOPATHOLOGICAL IMAGE ANALYSIS: A REVIEW

AUTHORS: S. Kaur and R. Vijay

DESCRIPTION:

The Employee turnover has always been a crucial matter of concern for organizations. In today's era of globalization there are ample opportunities for talented people in this world, therefore, employees are inclined to move from one organization to another. Due to this Corporates are facing the problem of attrition in the world of economic revival. A large degree of employee turnover is highly deleterious to both the organization as well as the employees. How to reduce employees attrition is a decisive challenge for the HR managers. Lucrative incentives and motivational theories have become useless and are considered as old practices of the human resource management. This article presents a holistic view of attrition and retention of employees in this competitive scenario with reference to Retail Industry. Along with other factors, Job Satisfaction has been considered as the major source of attrition and retention. The relevant literature review has been done for compiling this research paper in order to find out the various factors responsible for the attrition of employees in the retail sector. The research is based on the relevant literature review and also from the secondary data available on the internet.

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

DESCRIPTION:

Most compensation managers implicitly assume (or perhaps hope) that high pay levels will maintain and enhance future performance. To date, this assumption has been largely untested. Given the importance of pay level and the large expense that pay represents to most organizations, understanding how and why pay level influences the behaviour of employees in organizations is an important question. The purpose of this study is to examine the motivational effects of pay level on employee performance. To examine these issues, we collected field study data from a variety of sources, at three different times, and assessed the effects of employee pay level on subsequent self-esteem and performance. Specifically, we hypothesized that the effects of pay level on performance would be mediated by pay level effects on organization-based self-esteem. We base this hypothesis on the premise that level of pay within an organization communicates a sense of how much the organization values an employee and thus affects employee organization-based self-esteem which, in turn, enhances job performance. After controlling for organization tenure, and previous pay change, results supported a mediated model that suggests that pay level affects employee self-esteem, which in turn, affects employee performance.

3. TITLE: AN EXPLORATORY STUDY OF US LODGING PROPERTIES' ORGANIZATIONAL PRACTICES ON EMPLOYEE TURNOVER AND RETENTION

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

DESCRIPTION:

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. Design/methodology/approach--Using the Directory of Hotel & Lodging Companies, a convenient sample group of 24 management companies are selected. A self-administered mail survey instrument is developed to measure and test organizational initiatives and practices on employee turnover and retention. Using SPSS 16.0, two statistical tests are employed to test study hypotheses. Correlation analysis is used to identify the relationships between predictor and response variables. Likewise, regression analysis is used to examine the relationships between predictor and response variables hypothesizing that the effectiveness of practicing the human resource management organizational initiatives on management and non-management retention and turnover will differ. Findings--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. Moreover, Organizational Mission, Goals and Direction, and Employee. Recognition, Rewards and Compensation were found to positively reduce non-management employee turnover. Research limitations/implications--Owing to the study methodology and the relatively low response rate, generalization of the study findings is limited. Future replication studies are recommended. Practical implications--The 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 positive impact on workforce productivity. Originality/value--This study makes a major contribution to the relative influence of the practice of eight study-defined organizational initiatives on turnover in lodging businesses. [ABSTRACT FROM AUTHOR]

4. TITLE: Employee churn prediction

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

DESCRIPTION:

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 study, we applied well-known classification methods including, Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes methods on the HR data. Then, we analyze the results by calculating the accuracy, precision, recall, and F-measure values of the results. Moreover, we implement a feature selection method on the data and analyze the results with previous ones. The results will lead companies to predict their employees' churn status and consequently help them to reduce their human resource costs.

5. TITLE: ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS

AUTHORS : D. A. B. A. Alao

DESCRIPTION:

Employee turnover is a serious concern in knowledge based organizations. When employees leave an organization, they carry with them invaluable tacit knowledge which is often the source of competitive advantage for the business. In order for an organization to continually have a higher competitive advantage over its competition, it should make it a duty to minimize employee attrition. This study identifies employee related attributes that contribute to the prediction of employees’ attrition n organizations. Three hundred and nine (309) complete records of employees of one of the Higher Institutions in Nigeria who worked in and left the institution between 1978 and 2006 were used for the study. The demographic and job related records of the employee were the main data which were used to classify the employee into some predefined attrition classes. Waikato Environment for Knowledge Analysis (WEKA) and See5 for Windows were used to generate decision tree models and rule-sets. The results of the decision tree models and rule-sets generated were then used for developing a predictive model that was used to predict new cases of employee attrition. A framework for a software tool that can implement the rules generated in this study was also proposed. Keywords: Employee Attrition, Decision Tree Analysis, Data Mining.

6. TITLE: Machine learned job recommendation

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

DESCRIPTION:

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. Moving to a new job, however, is not an easy decision, which may depend on many factors, such as salary, job description, and geographical location. These patterns may involve features extracted from the business profiles of employees, the profiles of institutions, and the job transitions themselves. In this paper, the authors address the problem of recommending suitable jobs to people who are seeking a new job and have formulated it as a supervised machine learning problem. Their technique exploits all past job transitions as well as the data associated with employees and institutions to predict an employee's next job transition. 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.

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

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

DESCRIPTION:

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. However, it is limited to an assumption that the local space between any two positive instances is positive or belongs to the minority class, which may not always be true in the case when the training data is not linearly separable. However, mapping the training data into a more linearly separable space, where the SMOTE algorithm can be conducted, can fix this problem. 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.

Experimental results have demonstrated that this approach has better performance than traditional SMOTE.

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

DESCRIPTION:

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 the learning process of the ML algorithms, if the ratio of minority classes and majority classes is significantly different, ML tends to be dominated by the majority classes and the features of the minority classes are recognize slightly. As a result, the classification accuracy of the minority classes may be low when compared to the classification accuracy of the majority classes. The features in the minority classes are normally difficult to be fully recognized. 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.

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

AUTHORS : He-Yong Wang

DESCRIPTION:

Imbalanced data learning is problematic as traditional machine learning approaches fail to provide satisfactory results due to skewed class distribution. There are two solutions to this problem: increasing the number of minority class examples, called over-sampling, or decreasing the number of majority class examples, called under-sampling. A new approach to construct the classifiers from imbalanced datasets is proposed in this paper by combining SMOTE and Biased-SVM approaches. Often real-world data sets are predominately composed of normal examples with only a small percentage of abnormal examples. The cost of misclassifying an abnormal example into a normal example is often much higher than that of the reverse error. Experimental results confirms that the proposed combination approach of SMOTE and biased-SVM can achieve better classifier performance.

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

DESCRIPTION:

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. In this study, the authors propose a set of new algorithms to enhance the effectiveness of classification for 5-year survivability of breast cancer patients from a massive data set with imbalanced property. Results from this study show that the hybrid algorithm of SMOTE + PSO + C5 is the best one for 5-year survivability of breast cancer patient classification among all algorithm combinations. They conclude that, implementing SMOTE in appropriate searching algorithms such as PSO and classifiers such as C5 can significantly improve the effectiveness of classification for massive imbalanced data sets.

11. TITLE: Predicting Employee Attrition using Machine Learning

AUTHORS: Sarah S. Alduayj, Kashif Rajpoot

DESCRIPTION:

The growing interest in machine learning among business leaders and decision makers demands that researchers explore its use within business organisations. One of the major issues facing business leaders within companies is the loss of talented employees. This research studies employee attrition using machine learning models. Using a synthetic data created by IBM Watson, three main experiments were conducted to predict employee attrition. The first experiment involved training the original class-imbalanced dataset with the following machine learning models: support vector machine (SVM) with several kernel functions, random forest and K-nearest neighbour (KNN). The second experiment focused on using adaptive synthetic (ADASYN) approach to overcome class imbalance, then retraining on the new dataset using the abovementioned machine learning models. The third experiment involved using manual under sampling of the data to balance between classes. As a result, training an ADASYN-balanced dataset with KNN (K = 3) achieved the highest performance, with 0.93 F1-score. Finally, by using feature selection and random forest, F1-score of 0.909 was achieved using 12 features out of a total of 29 features.

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

AUTHORS: Akhil Gokuldas Warrier, Rajiv Prasad

DESCRIPTION:

The IT (information technology) industry has been a prominent contributor to the Indian economy over the last two decades. As a result it has been able to absorb a significant part of the talent coming out of the country's educational institutions. But in recent times, many new avenues are opening up for the best talent. As a result the employee attrition rate is quite high in the IT sector these days. 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. Herzberg's theory of workplace motivation has been one of the most validated theories of motivation in the western world. But, it has been found to operate with some differences in different countries, especially Asia, because of cultural differences. For example, see Sithiphand, 1978; Hauff, 2014. Attempts have been made to understand the workplace motivation of employees in the Indian IT sector. But there have been hardly any effort to understand the way the factors pointed out by Herzberg affect job satisfaction of employees in the Indian IT industry. In this study we examined the role of these two sets of factors in the Indian IT industry. Data was collected from 153 IT industry employees. It was found that contrary to what is predicted by the theory, the hygiene factors play a stronger role in predicting job satisfaction of the Indian IT sector employees. The implications of this finding are discussed in the paper.

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

AUTHORS: Tince Etlin Tallo, Aina Musdholifah

DESCRIPTION:

An imbalanced dataset is a condition that has a minority class which is a class has far fewer instance distributions than other classes. The imbalanced condition can affect the performance of standard classifier algorithms that lead to the biased of the results classification or tend to become a majority class. The SMOTE method overcomes the imbalanced masses by creating synthetic instances of minority classes. However, the implementation of SMOTE resulted in overgeneralization because generated instances have the same amount regardless of the distribution of instances. As a result, the boundaries between classes are unclear. The SMOTE-Simple Genetic Algorithm (SMOTE-SGA) method is used to determine the sampling rate of each instance in order to obtain unequal amounts of synthetic instances. The tests were performed using some imbalanced datasets by comparing the classification results measured using G-means and F-Measure. The results of the application of genetic algorithm at SMOTE can improve the classification result by obtaining better G-means and F-measure value.

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

AUTHORS: Tince Etlin Tallo, Aina Musdholifah

DESCRIPTION:

Data classification is highly significant in data mining which leads to a number of studies in machine learning with preprocessing and algorithmic technique. Class imbalance is a problem in data classification wherein a class of data will outnumber another data class. Sentiment Analysis is an evaluation of written and spoken language which determines a person's expressions, sentiments, emotions and attitudes and is commonly used as dataset in machine learning. This study is a comparative analysis 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 for classification of data given the same Sentiment Analysis datasets gathered by students of University of San Carlos. Weka, a Graphic User Interface (GUI) with a collection of machine learning algorithms for data mining, is use to preprocess and classify the datasets. The results had shown that 10 Folds validation provides better findings compared to 70:30 split in testing SVM and NBM with SMOTE. However, it also depends on how the datasets is preprocessed especially when it contains noisy data.

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

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

DESCRIPTION:

Human capital is of a high concern for companies' management where their most interest is in hiring the highly qualified personnel which are expected to perform highly as well. Recently, there has been a growing interest in the data mining area, where the objective is the discovery of knowledge that is correct and of high benefit for users. In this paper, data mining techniques were utilized to build a classification model to predict the performance of employees. To build the classification model the CRISP-DM data mining methodology was adopted. Decision tree was the main data mining tool used to build the classification model, where several classification rules were generated. To validate the generated model, several experiments were conducted using real data collected from several companies. The model is intended to be used for predicting new applicants' performance**.**