# Abstract

The purpose of this research project is to assist human resource managers and project managers in improving the retention rate of key employees in the organization, thereby reducing company staff costs. The study was conducted in three phases. Efforts are being made for the first time to analyze the employee retention system present in the form of researches and they are studied for their achievements and shortfalls to understand the system and build a better one. Second, decide which approach should be applied, finalizing the machine learning and deep learning platform in order to build a prediction system. Various analysis were performed when selecting a machine learning algorithm for predictive model training. The deep learning algorithm is used to create a predictive model where it will give you the most accurate results. After conducting several analysis on how to select an important employee and the hypothetical use of the method, a definite model is created with the help of conditional advertising, in which the employees are important. Finally, the HR analytical program is developed using python and H2O framework which identifies factors influencing employees intention to resign from the company. Also, there has been research on these topics among HR professionals. The special staff source file containing informationon employees has been compiled and they may be on verge of resignation. Using the python and its libraries, a classifier was developed that shows all the factors that contribute to the recruitment of employees so that HR managers and projectors can use them to perform critical work. Used in the development and evaluation of the IBM employee dataset (IBM, 2018). HR administrators can use this application to facilitate staff retention decisions.We used Deep Learning (Neural Networks), Distributed Random Forest (DRF), Generalized Additive Models (GAM), Gradient Boosting Machine (GBM), RuleFit, Stocked Ensembles and XGBoost. In a hybrid model, it is a combination of two models stacked on top of each other, making a very efficient model. This helps us to check if the stacked results are better than each model number. The current H2O.ai framework is implemented using the Python language. The performance of these algorithms is summarized and compared with the most efficient model performance. The model is compared with best model evaluation matrices such as accuracy, F1 score, processing time and AUC.

**Keywords:** Employee Attrition, Retention, Deep learning approach, Neural network, Distributed Random Forest (DRF), Generalized Additive Models (GAM), Gradient Boosting Machine (GBM), Rule Fit, Stacked Ensembles, XGBoost, H2O.ai framework, Accuracy, F1-Score, Training time and AUC.

Table of Content

[Abstract 1](#_Toc80521824)

[Table of Figures 4](#_Toc80521825)

[1. Introduction 6](#_Toc80521826)

[1.1. Purpose 8](#_Toc80521827)

[1.2. Research Questions 9](#_Toc80521828)

[1.3. Related Works 9](#_Toc80521829)

[1.4. Thesis Layout 10](#_Toc80521830)

[2. Literature Review 11](#_Toc80521831)

[2.1. Detection methods 11](#_Toc80521832)

[2.2. Machine learning approaches 13](#_Toc80521833)

[2.2.1. Supervised Machine Learning Approach 17](#_Toc80521834)

[2.2.2. Un-supervised Machine Learning Approach 19](#_Toc80521835)

[2.2.3. Reinforcement Machine Learning Approach 20](#_Toc80521836)

[2.3. Rule based approaches 20](#_Toc80521837)

[2.4. Deep learning approaches 21](#_Toc80521838)

[2.4.1. Employee Turnover 22](#_Toc80521839)

[2.4.2. Maintenance Management 23](#_Toc80521840)

[2.4.3. Factors in which individuals are engaged in their work 24](#_Toc80521841)

[3. Methodology 25](#_Toc80521842)

[3.1. Dataset Description 25](#_Toc80521843)

[3.2. Data Preparation 26](#_Toc80521844)

[3.3. Class Balancing 29](#_Toc80521845)

[3.4. Feature Extraction and Selection 31](#_Toc80521846)

[3.5. Model Training 32](#_Toc80521847)

[3.6. Model Evaluation 36](#_Toc80521854)

[4. Experiments and Results 37](#_Toc80521862)

[4.1. Experimental Setting 37](#_Toc80521863)

[4.1.1. H2O Sparkling Water Setup 37](#_Toc80521864)

[4.1.2. Data Preparation 38](#_Toc80521865)

[4.2. Models 49](#_Toc80521866)

[4.2.1. Deep Learning 49](#_Toc80521867)

[4.2.2. Distributed Random Forest 50](#_Toc80521868)

[4.2.3. Gradient Boosting Machine 51](#_Toc80521869)

[4.2.4. RuleFit 52](#_Toc80521870)

[4.2.5. XGBoost 53](#_Toc80521871)

[4.2.6. Stacked Ensemble 53](#_Toc80521872)

[4.3. Evaluation 54](#_Toc80521873)

4.3.1 [Comparison of Precision, Recall and F1-score 54](#_Toc80521874)

[4.3.2 Accuracy Comparison 56](#_Toc80521875)

[4.3.3 F1-Score Comparison 57](#_Toc80521876)

[4.3.4 AUC Comparison 58](#_Toc80521877)

[4.3.5 Training Time Comparison 59](#_Toc80521878)

[5. Conclusion 60](#_Toc80521879)

[5.1. Contributions 60](#_Toc80521880)

[5.2. Future Work 60](#_Toc80521881)

[References 61](#_Toc80521882)

[Appendix 65](#_Toc80521883)

# Table of Figures

[Figure 1: Human Resource Management layout 6](#_Toc80031703)

[Figure 2: Employee Attrition Analysis 7](#_Toc80031704)

[Figure 3: AutoML framework in H2O environment for model training and model deployment 9](#_Toc80031705)

[Figure 4: Data analysis in H2O environment 11](#_Toc80031706)

[Figure 5: Strategic analysis for employee attrition prediction on organisation 13](#_Toc80031707)

[Figure 6: Employee Attrition dataflow analysis 13](#_Toc80031708)

[Figure 7: Employee selection process and data analysis 14](#_Toc80031709)

[Figure 8:Responsibilities of Human Resource Management 15](#_Toc80031710)

[Figure 9: Supervised learning approach 17](#_Toc80031711)

[Figure 10: Un-supervised learning approach 18](#_Toc80031712)

[Figure 11: Reinforcement/ Deep learning approach 18](#_Toc80031713)

[Figure 12: Human Resource planning to avoid employee attrition 21](#_Toc80031714)

Figure 13: Process Flow Diagram of employee attrition detection system 24

Figure 14: The classes in data and Under and Over Sampling 29

Figure 15: The H2O environment setup 37

Figure 16: Data information 38

Figure 17: Data missing value check 39

Figure 18: Data information 40

Figure 19: Class balancing 41

Figure 20: Data classes after Random Oversampling 41

Figure 21: Correlation Heatmap 42

Figure 22: Data information 43

Figure 23: Chi square test to find best feature 43

Figure 24: Chi square test scores 45

Figure 25: Chi-square test results 46

Figure 26: Data information 47

Figure 27: Data H2O streaming 47

Figure 28: H2O data 48

Figure 29: Deep learning model bulid 48

Figure 30: confusion matrix of Deep learning 49

Figure 31: Distributed Random Forest model built 49

Figure 32: confusion matrix of DRF 50

Figure 33: GBM model bulid 50

Figure 34: confusion matrix of GBM 50

Figure 35: Rulefit model rules 51

Figure 36: confusion matrix of Rulefit 52

Figure 37: xgboost model built 52

Figure 38: confusion matrix of xgboost 52

Figure 39: stacked ensemble model 53

Figure 40: confusion matrix of stacked ensemble 53

Figure 41: Model result 54

Figure 42: Model result in barplot for comparitive analysis 54

Figure 43: Model result in line plot for comparative analysis 55

Figure 44: Accuracy result in bar plot for comparative analysis 55

Figure 45: Accuracy result in line plot for comparative analysis 56

Figure 46: F1- Score result in bar plot for comparative analysis 56

Figure 47: F1- Score result in line plot for comparative analysis 57

Figure 48: AUC - Score result in bar plot for comparative analysis 57

Figure 49: AUC - Score result in line plot for comparative analysis 58

Figure 50: Time taken result in bar plot for comparative analysis 58

Figure 51: Time taken result in line plot for comparative analysis 59

# Introduction

In today’s competitive economy with unparalleled technological advancement, the discovery, study and analysis of data brings new information known as the “information economy”. Information technology is not just a source of data, it is, above all, a positive aspect of data analysis that makes it possible to process data collections and extract information from them. Data has become a strategic asset for many companies in many sectors, including those involved in business processes. All types of companies benefit from adopting new technologies (Fallucchi, 2020) and offer a number of advantages in terms of data collection, management and analysis capability and competitive advantages. In fact, analyzing large amounts of data can lead to better decision-making processes, the implementation of pre-established business goals, and better business competitiveness (Tomasen, 2016).

Organizations that adopt the implications of artificial intelligence for corporate decision-making activities have a number of areas (Mohbey, 2020). In recent years, there has been a greater focus on employee (HR) because employee quality and capabilities are a growth factor and have a real competitive advantage in companies (Ribeiro, Singh &Gestrin, 2016). In fact, after further adoption in the retail and marketing sectors, artificial intelligence has now begun to guide company decisions regarding its employees, with the goal of supporting human resource management decisions in objective data analysis rather than subordinate expectations.



Figure 1: Human Resource Management layout

Often, companies try to maximize profits. In companies where employees do routine tasks, they may be inclined towards hiring, part-time worker (such as gig economy) as they have less contractual obligations. However, in companies where employees perform specialized tasks, professional expertise and continuity in employee work is required.

The importance of persistent skills, knowledge and learning ability has proven to be important for businesses. The use of artificial intelligence in the field of human resources allows companies to convert information into action using imaginary models: such models allow employees to use data collected by the company over the years, thereby identifying important issues. Reduced and all HR activity improved (Tomasen, 2016; Kelher, Nami & D'RC, 2015).



Figure 2: Employee Attrition Analysis

Companies invest a lot of time and resources to recruit and train employees according to their strategic needs. Therefore, employees (more or less) represent real investment of companies. When an employee leaves a company, the company loses not only a significant number of employees, but also resources, especially the money and effort of human resources employees, who are responsible for recruiting and selecting those employees and assisting them in their duties. As a result, the company must continue to invest in recruiting, training and developing new employees to fill vacancies. Training a new employee is a lengthy and costly process and controlling and reducing the layoff rate is highly beneficial to the company: Retirement is defined as a retired or retired employee in a company. Furthermore, satisfied, highly motivated and loyal employees are the core of the organization and contribute to the productivity of the organization. In the literature, some authors suggest that employees should be happier and more motivated to be more creative, productive, and better, which ultimately produces and maintains strong and effective performance (Jordan & Michel, 2015; Ayodel, 2010; Tomsen, 2016). Job satisfaction is reflected in financial texts as a good estimate of profit. Job satisfaction statistics are powerful estimates of separation and resignation, which also control wages, hours and the average population and employment diversity(Marjanovic, Sess-Kekmanovic&Widgen, 2018).

Many researchers have shown the usefulness of human resource management (HRM) in working conditions, production and management, and in identifying relationships with production. In fact, the results show that the impact of HRM on output has a positive impact on business capital growth and resilience. Much of the research focuses on customer analysis and behavior, and does not focus on the major assets of the company that its employees represent. Several studies have analyzed employee attractiveness. Current research(Tomasen, 2016) has shown that the number of employees and job-related characteristics are very important factors in attracting employees, such as salary and working hours.

## Purpose

Maintaining employee attractiveness in order to get a cheap and healthy price in a company is very important to keep the company running smoothly and therefore competitive. (Shaw, 2010; Ajit and Punnos, 2016; Koyas, 2006). The pressure on HR departments to provide organizational value led to the introduction of data-based decision making and electronic learning (Tomasen, 2016). Today, high-level recruitment is considered an issue for companies and has increased the responsibility on HR departments to maintain employee recruitment levels at the highest level (Park & ​​Shaw, 2013; Gleebeek& Box, 2017; USAHO, 2010).

The aim of this study is to examine how machine learning models can be used in large models (HR) in assessing employee attractiveness and decision making in an organization. Should be publicly examined. That is, the goal of this study is to evaluate the feasibility of estimating employee turnover based on employee data sets and to make physicians an effective working tool for estimating employee turnover.

## Research Questions

The following research questions were framed in such a way that all the design and implementation will be able to justify the employee attrition analysis using machine learning techniques and models.

1. Is it possible for the companies to predict employee’s attrition using the historical data kept in human resources department?
2. Which model from Deep Learning (Neural Networks), Distributed Random Forest (DRF), Boosting Machine (GBM), Rule Fit, Stacked Ensembles, and XGBoost performs better for employee attrition prediction?
3. Which machine learning model from Deep Learning (Neural Networks), Distributed Random Forest (DRF), Boosting Machine (GBM), Rule Fit, Stacked Ensembles, and XGBoost have faster learning /training when applied to employee attrition dataset?

## Related Works

Research conducted by Fallucci analyzes how objective factors affect employee attractiveness, identifies key factors influencing an employee’s decision to leave a company, and assesses whether a particular employee will leave the company. After training, the employee attractiveness assessment model is tested on real data provided by IBM Analytics, which has 35 features and approximately 1500 models. Considered classification algorithms: Native base classifier for Gaussian naive bayes, multivariate Bernoulli models, Logistic Regression Classifier, K-Near Companion (K-NN), Decision Tree Classifier, Random Forest Classifier (S) ). The test results are presented according to the classical matrix of the Gaussian navie bayes classifier algorithm. It reveals an excellent memory rate (0.54) because it measures the variability in detecting all positive conditions and provides an unrealistic value of 4.5% of the total sequences (Faluchi, 2020).

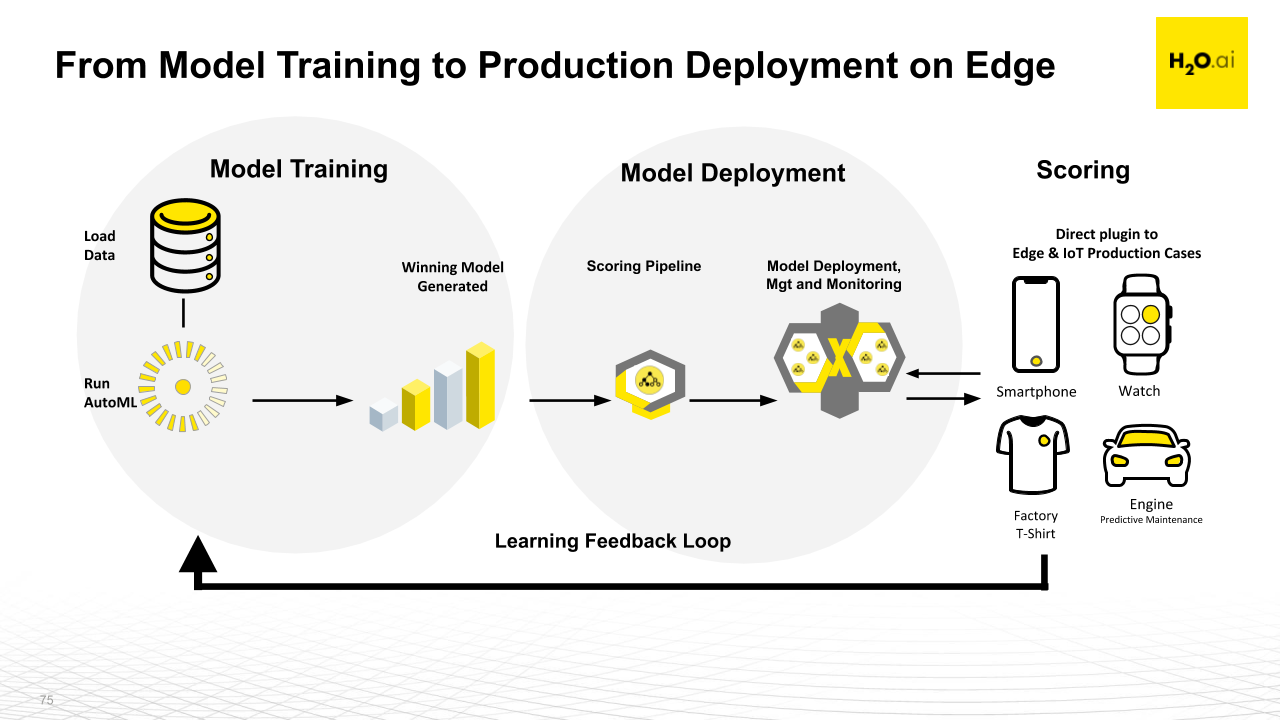


Figure 3: AutoML framework in H2O environment for model training and model deployment

Research by Matt Dancho uses the H2O approach to the AutoML framework to attract employees. H2O AutoML can be used to perform machine learning tasks, including automatic lifting and adjustment of multiple models at user-defined time intervals. The model is 88% accurate but the attrition = NO in the selection is 79% accurate, which is not a good model (business-science.io).

## Thesis Layout

This research is organized as follows: A brief background that introduces some of the areas where specialized practice is used today, and the importance of task management for organizations benefit. Based on the introduction and background, objective and research questions are presented. Second, Literature reviews on machine learning, human resource management, and machine learning are introduced into human resource management theory. Third, an explanation of how this paper is presented. Fourth, all academic influences are on appendix. Fifth, analysis and discussion of the material is provided. Finally, a conclusion was reached along with future research.

# Literature Review

Author KK Mohe Naive Bayes conducted a comparative analysis between different machine learning methods such as SVM, Decision Tree, Informal Forestry and Dimentionality Reduction. The result presented helped to identify the behaviour of future employees of company. Test results show that systematic practice can reach 86% higher accuracy than other machine learning methods(Mohbe, 2020).

Another research paper developed by Adarsh ​​Patel is the creation of a workforce assessment system based on the staff database from the Kaggle website. For assessment purposes, they used four machine learning algorithms such as KNN (K-Near Neighbourhood), SVM (Support Vector Machine), Decision Tree, and Random Forest. Random Forest Tree performed better with 86% accuracy(UPL, 2020).

The model performance could be more refined and tuning the model as per the data would help to get better results in the researches mentioned above. Also, All the papers use traditional classification models for the prediction.

## Detection methods

Today, companies are making a lot of efforts in the Department of Human Resources (HR), where one of the most important functions for management is to reduce attrition by attracting workers (Coys, 2006; Ajit&Punnos. 2016). This is because of the consequences that come with attracting employees. Cost reduction for new jobs and money spent in the form of training to replace experienced employees who have left companies (Alduwaz& Rajput, 2018;Sexton et al. 2005). Furthermore, as work progresses, clear and concise information is lost and important social relationships are severed (Ashworth, 2006; Droze&Hubbler, 2003; Levy, 2011).

HR employees often struggle to define the structure of their organizations and their job is to make more important and better decisions. Today, there is an expansion of HR departments trying to make decisions based on data. Data-based decisions lead to better organizational performance which means that if HR departments make decisions on data, it benefits the organization.

(Tomasen, 2016). Field data analysis refers to the decision-making process based on data analysis. Businesses use data analytics to understand and capture data attributes from data used to make decisions (Kelleher, Nami & D'RC, 2015). Data analysis in the field of human resources is growing and has the potential to increase the importance of HR segments in organizations (King, 2016).

Machine learning is a lesson in how computers learn from experience (Jordan & Michel, 2015; Ayodel, 2010; Tomsen, 2016). i.e. teaching a computer to assess example results. The data used by machine learning algorithms play a big role in its success. As the amount of data available increases, so does the ability to solve large and complex problems (Domingos, 2012). Electronic learning methods are widely used today and can be found in places like automotive, stock markets or healthcare (Domingos, 2012). Although the use of ML models is limited to human desires (Ribeiro, Singh &Gestrin, 2016) a variety of machine learning strategies are available: monitoring, and advanced learning (Ayodel, 2010). It is the same but the way and the requirements are different.

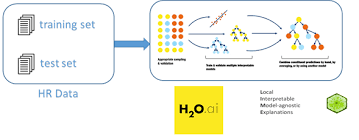


Figure 4: Data analysis in H2O environment

Machine learning and prediction models have been shown to accomplish tasks that humans cannot. However, how well a specific learning program results. If customers cannot trust them. Note Ribeiro, Singh, and Gestrin (2016) "If users do not trust the model or do not make predictions, they will not use it." And there are two definitions of trust: (1) "reliance on expectation", which means that the user's expectations must be sufficiently reliable to apply the information obtained from that estimate, and (2) "trusted models", i.e. trust the customer. The model should behave as expected. Both are related, but the first represents an assessment and the second represents model trust.

Before using it in real world situations. Both depend on how well the user understands the functionality of the model, so that the intermediate step between input and output does not look like a blurry magic box (Ribeiro, Singh &Gestrin, 2016). Relying on machine learning models becomes even more important when file results are used in decision making. And understanding what the model actually does plays a big role. Take for example in a field such as health care where ML is used for a variety of activities such as medical care. In order to process the ML diagnostic result, one must rely on the model used as the results are devastating. Therefore, the use of learning tools depends on people who rely on different models and whether we can use machine learning in areas such as health care or personal vehicles, we must understand the patterns and know where they go wrong (Heaven, 2020).

## Machine learning approaches

Machine learning (ML) is a standard method for assessment purposes in data analysis (Domingos, 2012; Kelleher, Nami & D'RC, 2015). Machine learning is a lesson in how computers learn from experience (Jordan & Michel, 2015; Ayodel, 2010; Tomsen, 2016). The idea is that the algorithm learns from the data set and improves after new data is revealed. Most likely, ML can be used in HR departments to assess employee recruitment. The problem with this is that decisions made in the human resources departments can have serious consequences for employees, their families and the organization (O'Neill, 2016; Marjanovic, ICECs-Kekmanovic&Widgen, 2018; Devbold et al. 2019).Therefore, this paper examines whether it is possible to use learning tools to assess employee attractiveness. This is done by analyzing two different strategies to evaluate employee recruitment and examine algorithms and life-changing decisions on democracy in revealing algorithmic decisions.

Employee income plays a major role in the success of a company and there are many reasons for this (Ajit and Punnos, 2016). As mentioned in the introduction, there are many impacts on the entry of employees, which can have negative consequences for the organization. Some but not all, first time workers benefits are expensive. People who give up voluntarily must change. The hiring process takes time, which is considered to be more productive jobs. Usually, it takes time for the new employee to reach at full speed again.



Figure 5: Strategic analysis for employee attrition prediction on organisation

Investing in new employee training is not uncommon (Coys, 2006; Aldous & Rajput, 2018; Sexton et al. 2005). Second, production may be affected. Experienced employees have the knowledge and experience to contribute to production. Inexperienced workers are less likely to be produced as skilled workers and more experienced workers take longer to reach productive levels (Elduce& Rajput, 2018; Sexton et al. 2005). Third, higher wage levels are likely to adversely affect the morale of the rest of the employees. When employees leave, one reason for this is, some employees have to perform tasks that someone else has to do. Work in progress. If the turnover rate is high, it means it will increase the burden of impairing motivation and behavior (Ashworth, 2006; Droze&Hubbler, 2003; Levy, 2011). Fourth, employee benefits affect profits. The above results are impressiveOrganizational performance also affects the ability to work at the desired level. Cost, loss of information and low productivity are both affected by profits.

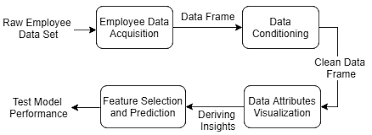


Figure 6: Employee Attrition dataflow analysis

The results on hiring are reasons why it is important for companies to use effective and prudent strategies to minimize employee benefits. Assessment capacityStaff can add tools to the Access Toolbox when developing such strategies. These strategies increase the likelihood of terminating employee benefits and thereby reduce the risk of all the side effects listed above (Ajit&Punnos, 2016; Sexton et al. 2005).

Today, companies are trying to estimate workers' profits and use that knowledge to reduce labor costs. However, most expectations for staff today focus on overall company performance, not on an individual level. Assessing employee benefits at a general level in an organization can help identify benefits as an issue and provide life-saving strategies and analyze why people are leaving. However, estimating employee growth at each level will give companies better opportunities to work harder and take the necessary steps to reduce employee attractiveness and thereby positively impact employee turnover. It has a positive impact on the organization's performance to retain technology, maintain and improve productivity, maintain social stability and work ethic, and prevent disruption in projects. Happens. (Ajit&Punnos, 2016; Tomasen, 2016; Marjanovic, Sess-Kekmanovic&Widgen, 2018).

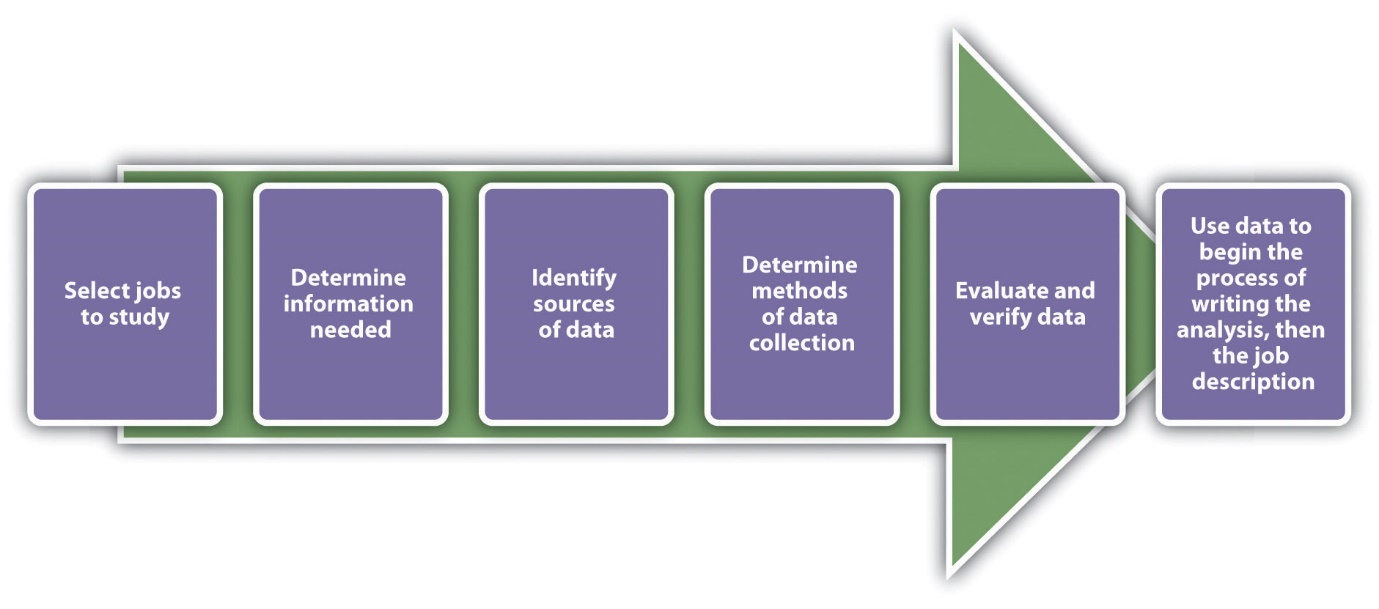


Figure 7: Employee selection process and data analysis

It took a long time to learn machine learning to enter the field of human resources and was not used until a few years ago (Fagela, 2019). According to LinkedIn (2018) only 22% of companies have launched analytics in their HR segments and do not know how to use these analytics effectively. There is great access to big data in areas like marketing, for example in product sales in the country. It contains data on how many products have been sold, when they were sold and how many times you received the product. In those cases, ML is a very effective tool that you can use to analyze large data sets. (Kapelli, Tambe&Yakubovich, 2018) The use of machine learning in human resources poses some problems and challenges. First, the data stored on employees is very small compared to the data sets used in data science and the data is generally not stored (Capelli, Tambe et al., 2018). Even a large company with 10000 employees is nowhere near getting HR information in the form of product sales data. Second, human resource management decisions have far-reaching consequences (Sutherland &Wock, 2011;Kapelli, Tambe&Yakubovich, 2018). For example, the decision of who was fired and who is doing the job. Third, individual performance is often difficult to measure because complex characters often rely on other roles - collaboration - and individual performance in team performance is difficult to pin (Capelli, Tambe&Yakubovich, 2018; Murray &Enerson, 2007). Fourth, deciding whether or not to hire an individual not only on the basis of technical ability but also on the basis of social and psychological relationships among employees.



Figure 8:Responsibilities of Human Resource Management

(Klugger et al. 2002; Capelli, Tambe&Yakubovich, 2018). Finally, machine learning algorithms require training and the algorithm is very clear about it. For example, this can be a problem when hiring new employees. If machine learning algorithms train current employees and the majority of employees are said to be white, the algorithm will be more suitable for white people when looking for new candidates. (Capelli, Tambe&Yakubovich, 2018) This exact example happened on Amazon‌ in 2018 and they removed gender directly from their model as a process, but the model still learns to serve a specific group of people.

However, machine learning has enormous potential in HR. This is because performance and human resource outcomes affect the performance of the organization in many ways. These include demonstrations such as hiring new employees, training new and old employees, identifying good and bad performance, deciding who to promote, retaining employees and employee benefits(Kapelli, Tambe&Yakubovich, 2018).

### Supervised Machine Learning Approach

A wide variety of machine learning strategies are available: supervised, un-supervised, and reinforced learning (Iodel, 2010; Librecht& Noble, 2015).

Supervised machine learning algorithms are provided with training examples, including input data and output data (Iodel, 2010; Molnar, 2020). The algorithm is trained by linking the link between the input data and the output data provided. In training, the algorithm contains a coherent output data that allows the algorithm to identify patterns between the input data and the output data so that the algorithm can find out which input data is associated with the corresponding output data (iodale, 2010). The goal is divided into two parts. First the algorithm learns to associate the input data with the correct output data (such as categories). Second, a new set can be assigned to the algorithm if the input data and accurate results are estimated. That is, the “correct answer” is given to learn the mapping between the two algorithms in the learning stages. At the application stage new input data is given to the algorithm and based on experience, it evaluates the results. The level of accuracy of learning machine learning algorithms is determined by using test data set inputs and results, providing input to a trained algorithm and verifying the relevant results in the test data set. For example, if we have 100 inputs and 100 outputs, we can say that the algorithmic accuracy is 95% if 95 of those inputs are separated by the correct product. The accuracy of the algorithm can be increased (or decreased) by changing the amount of training.



Figure 9: Supervised learning approach

Algorithmic training for the study of machine monitoring is as common as possible in the current problem (Domingos, 2012). The algorithm is used in the real-world problem, which is not likely to be trained in exactly the same model. Therefore, the algorithm must be designed in a certain way so that it can solve new problems. For this reason, when testing an algorithm, it should be tested with real-world data, not data that can be used to train the algorithm for training (Domingos, 2012).

There are two types of monitoring learning algorithms: decentralized and isolated (Domingos, 2012). Withdrawal is used when predicting the persistent outcome of a problem, which is a particular type of value (Ayodley, 2010). Separation is used when there is a problem finding a different classification of data (Cotyantis, 2007; Iodel, 2010). For example, we will look at the rain forecast for the festival day. The segmentation algorithm is used estimate whether it will rain - yes or no - and the regression algorithm will be used to estimate how much it will rain, i.e. the amount of rain.

### Un-supervised Machine Learning Approach

Unlike supervised learning, non-supervised learning algorithms do not look for results. That is, the algorithm reads only samples from the data sample input without given output (Gentleman et al. 2008). Instead of explaining how inputs are integrated, work here (Ayodel, 2010). Integration is the process of managing data into different groups are somehow united (Gentleman et al. 2008). As an example, we can see the division of fruits into groups. Think basket with apples, pears and oranges. The task of the algorithm is to divide the fruit into different groups with the same fruit. Unsupervised learning algorithm can find similarities between all apples without knowing the results and sort them into a group individually, as well as find similarities between all the oranges and sort them into a group.

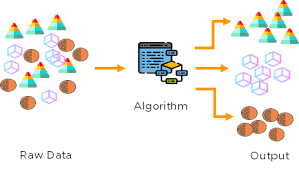


Figure 10: Un-supervised learning approach

### Reinforcement Machine Learning Approach

Reinforced learning is a mixture of supervised and supervised learning. With this method the algorithm is trained with a given installation without proper extraction. Yet the solution specified by the algorithm comes with a positive or negative feedback (Iodel, 2010). That is, if the suggested algorithmic solution is correct, a positive answer is also obtained if the suggested solution gives a false negative response.

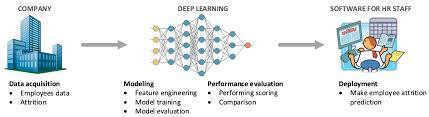


Figure 11: Reinforcement/ Deep learning approach

## Rule based approaches

Models used Traditional speculative models and results are often inconsistent with formal retrospective models (Yilnaz and Kinar, 2011; Paliwal and Kumar, 2009). Over the past 20 years, traditional models have been challenged by machine learning models intended to assess multiple areas (Domingos, 2012; Paliwal and Kumar, 2009). Machine Learning Models (Lundberg &Su-In, 2017; Paliwal& Kumar, 2009). And then it could be human resources by assessing workers ’interests. There have been only a few studies on estimating employee benefits or employee placements at each level. That means trying to estimate when the employee is leaving. Sexton et al. (2005) conducted a study in which they attempted to assess employee attractiveness towards people who use learning tools. The results were somewhat successful, but not Sexton et al. (2005)This suggests that further research and numerous similar studies need to be conducted before reaching any conclusions on how to assess employee benefits at each level.

A similar study was done by Ajith and Punnus (2016). Their results have also shown promise in assessing employee benefits, but as Sexton et al (2005) suggest, further research and evaluation in many organizations should be done before concluding a favorable model that can be used to assess employee benefits. As well as Sexton et al. (2005) and Ajith and Punnos (2016) identify several potential companies benefits if the model is successful in assessing employee benefits. It has been a practice of the last decade for leaders to be interested in using learning tools in their organizations to make decisions (Alduwaz and Rajput, 2018). One reason for this is that researchers have found that organization performance is positively impacted when making decisions based on data, and machine learning is widely used to design decision-making models that support learning from large data sets. Predictive data are used in testing (Tomasen, 2016; Kelher, Nami & D'RC, 2015). Machine learning is used to assess employee attractiveness at the individual level, rather than the ability to develop end-to-end strategies that can minimize employee benefits and thereby have a positive impact on the organization. It also increases the number of HR segments, which increases their value to companies.

Since the use of learning tools to assess employee attractiveness affects people and it can be seen as sensitive information, it is important that people be prepared to use such models. Ribeiro, Singh and Gestrin (2016) say that people need to rely on machine learning models without any side effects. It is therefore also necessary to examine the willingness of individuals to be exposed to such models.

## Deep learning approaches

The use of algorithms to make decisions in the personal and private life of the business is becoming more common (Marjanovic, Sess-Kekmanovic&Widgen, 2018). These decisions play a big role and often change lives. In HR it can be in the recruitment process to evaluate job applicants, decide who should or should not continue this process and make career decisions (Marjanovic, Sess-Kekmanovic&Widgen, 2018; Devbold et al. 2019). Although algorithmic decisions are beneficial in most cases, allowing algorithms to make decisions for us can have dire consequences (Marjanovic, ICECz-Kekmanovic&Widgen, 2018; Devbold et al. 2019).

Taking the recruitment process as an example, the algorithm can examine thousands of job applications with just one click, decide which students are happy, and one person has to manage all the activities manually. It is clear that using algorithms saves a lot of time compared to the time taken. In the same scenario, consider the case of a prominent baptized candidate who was automatically removed due to the algorithm. Algorithmic decision-making transformation services that change or alter people's lives are dangerous and even harmful to individuals and their families and organizations (O'Neill, 2016; Marjanovic, Sess-Kekmanovic and Widgen2018). Flexible services are usually services such as health care, school and geriatric care. However, with the definition of an exchange service as a service that changes people's lives, HR decisions are included.

### Employee Turnover

The goal of all companies is to gain competitive advantage through their competition. Another important area to consider to gain competitive advantage is human resource management (HRM). Employee happiness and satisfaction levels have been shown to affect their participation and how well they do and the performance of the organization (Anita, 2014). Effective HRM has the potential to have a positive impact on employees, improving their performance and thereby increasing the competitive advantage of the organization (Coys, 2006).

Satisfying the majority of employees in HRM. Studies by Ryan, Schmidt and Johnson (1996) have shown that employee satisfaction is related to employee benefits. High satisfaction often drives employee income. Numerous artistic studies have shown that employees benefit themselves and often contribute to the smooth running of the organization.

Influencing competitive advantage (Choice, 2006; Davis, 2013; Morrow & McElroy, 2007; Des & Shaw, 2001). Some of the ways in which an organization can be affected are: impact, (2) increased workload on other employees, (3) corporate behavior, (4) lower prices i.e. lower employment as a result of training activities and the recruitment process (Coys, 2006; James & Matthews, 2012). Reduces costs. Employee benefits have clear and unambiguous implications. It is important to look at both of them to gain a deeper understanding of the actual results and to learn how employee benefits affect competitive advantage.

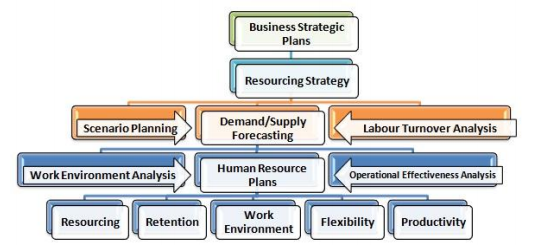


Figure 12: Human Resource planning to avoid employee attrition

Cogmar et al. (2006) argue that a particular employee benefit is beneficial to the organization, but in many cases this benefit is too costly and can disrupt workflow. An example of how much effort it takes to turn our eyes to the US fast food industry (2006).

The first step in reducing employee attrition is to understand why this is happening and this is very important. Because, knowing the cause of the problem gives companies the opportunity to take action.

### Maintenance Management

Retention management is part of HRM and studies how companies can work to retain their employees for longer. In order for you to become a proactive maintenance manager, companies must approve and evaluate the status quo (James & Matthews, 2012). Knowledge of how the position of profits is not adequate. Studies have shown that in order for companies to function, they must also have a strategic plan for practical actions to retain their employees and that companies with existing retention strategies are more likely to retain their employees. Raising and retaining employees creates the attraction of attracting new employees (James & Matthews, 2012). Furthermore, James & Matthews (2012) argues that the company strives to retain its best employees and that failure to do so causes them to lose their best employees. That is why end management has shifted to an important part of HRM and the results have a huge impact on competitive advantage.

### Factors in which individuals are engaged in their work

Voluntary change can be achieved in one person, but it can be difficult and necessary (Mitchell et al. 2003). Many researchers have observed this over the years why workers left their companies and went to others (Mitchell et al. 2003; Davis, 2013; Ladelsky and Catano, 2013)? The simple answer is very simple. If an employee is dissatisfied with his or her job and lacks some commitment to the company, that employee is often fired (Mitchell et al. 2003).

There are many factors that influence people’s desire to be in their organizations both externally and internally. External factors may be job opportunities or people moving to another area (Mitchell et al. 2003; House‌nect, Roda& Howard, 2009). Internal factors embedded within job satisfaction and organizational commitment (Mitchell et al. 2003).

Job satisfaction in some subdivisions results in a variety of factors that contribute to the choice of why employees should be in their organizations. Factors affecting job satisfaction include the need for employees to connect with other employees and activities within the organization, employees to feel that they are their own organization, employees to feel that their companies support them in their work and employees to understand their workload (Michel et al. 2003; House‌nect, Roda& Howard , 2009; Cho, Johansson &Guchait, 2008).

# Methodology

The methodology of our proposed model is divided into the multiple stages. In this work, our main objective is to utilise the H2O framework for the prediction of employee attrition. The overall framework of the proposed architecture is shown in Figure 13.

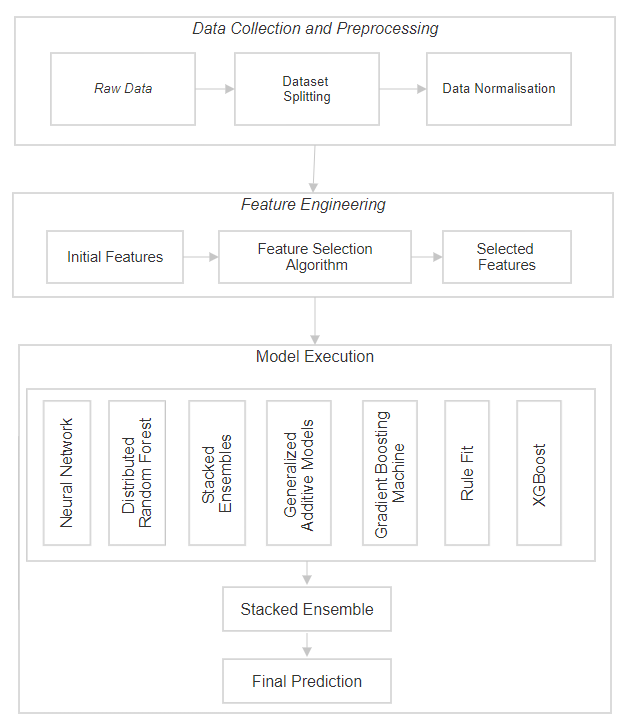


Figure 13: Process Flow Diagram of employee attrition detection system

## Dataset Description

The design starts by importing the employee data downloaded from kaggle.com into the python notebook environment in python machine. Then the task of data pre-processing will be performed. This step it is checked for its structure, null/missing values, class balancing, and outliers. All these issues will be treated, and encoding the data for the classification models. The first task in the data pre-processing step is to remove the missing/null values. This can be done in two ways by either removing the missing values or by adding a default value.

## Data Preparation

Exploratory data analysis is the key to make high-performance data models, hence that is the appropriate process to win the battle in data Maintenance. Moreover, exploratory data analysis and data pre-processing are the major two terms that are involved in any kind of task which is overlapped. During that period, the process is used at interchangeable aspects.

***Exploratory Data Analysis***

It is important to analyse some basic questions before focus on the key concept of the exploratory data analysis process.

* How can you select the appropriate algorithm for a set of data?
* How will you ensure that you have the ability to apply the particular algorithm which is related to the machine learning process to a data set?
* How can you define and modify different feature variables that could effectively be used for the modelling of data?

From that aspect, with the help of the EDA process, it will be easy to answer all the questions properly.

EDA is the potential process that helps to visualize, summarize,and in deapth analyses of vital traits of a data set. Considering the impact of business, domain knowledge can deliver great help in recognizing the data and their extracted result at the time of using the EDA process effectively. Exploratory data analysing is highly valuable in the section of data science. That allows getting close look at the particular period. Hence, the effective results can be obtained from an effective model for future aspects which will be consist of validation, accuracy, interpretation and appropriate application in regards to the context.

For meeting the particular level of certainty, here some tasks can be done by the exploratory data analysis.

* It needs to identify the way of a collection of raw data for the process of work
* It is important to generate the data in a valid approach for bringing in a positive outcome
* It is important to be more flexible with different kinds of data with different characteristics
* It is vital to learn about the different features of the data set and identify their interrelated relationship in a proper way
* Extracting valid information regarding the business operation from stakeholders is equally needed with a proper data set
* For reducing the anomalies, human error and missing values with outliers, it is important to check and extract valid data set properly
* It is highly necessary to identify the effective pattern in the data set which is hidden. That will help to make effective comprehension of the problem of business operation adequately

After completing the exploratory data analysis, data scientists have a set of features at there disposal. That can be used for the appropriate data modelling section.

Generally, data scientists need to go back to the data collection and data cleaning section by the use of data science methods properly. Hence, data scientists will be able to transform the entire data set in regards to the business outcomes which is expected.

**Pre-processing steps of data**

In the data science section, data processing is highly effective before start the phase of the modelling process. The steps mainly consist with,

1. *Maintaining highly effective features with correlation:*

While two features are linked in a correlated way and introduce equal information, then clustering and correlation can be plotted in the proper way. Moreover, correlation among two features creates a 99 % higher amount, and then one information can be removed in the proper and safe way based on the general rule. For making a proper correlation, the threshold percentage can help to decide based on the issues of business operation.

1. *Erase duplicate point of data:*

While the training data set will be high in size, then identical data points can be repeated several times approach. In that aspect, to remove the bias ness of data at the time of modelling, it is required to erase harmonious data points properly.

1. *Handling data imbalance*

In the case of the imbalanced data set, data scientists can be able to create an oversample of the class with low data points. In that aspect, they can create duplicate points of data with the proper way to extract the positive results. A data scientist can be able to erase the similar data points which are under-sample of the class with a high number of data points.

1. *Controlling data features with low variances*

A data scientist can be able to erase the entire variance which has low values. In the target of variables, data features can be constant in the data set in the proper manner and that failed to explain with influence the variation.

1. *Monitoring missing values:*

For handling the missing values in the data set, there have various techniques while importing of data set and libraries. Considering the low percentage of missing values, data scientists can drop the rows containing missing values in terms of low feature values. In that aspect, while the missing values of a high percentage are more than 40 to 50 per cent, then, it can be dropped properly.

In the data analysing process, the data set can be complete rarely. Even, data can be missed for various reasons such as data not captured properly, collected without validation, and much more reasons. From that aspect, data scientists can be continuing analysing the data after selecting the missing values. The process is mainly called imputation. Impute of missing values can be done by the mean and median form with numerical features and mode for features at categorical way.

1. *Scaling of features:*

Several kinds of data feature mainly depend on the scaling in a data set. Scaling is a highly effective method that helps to be deployed to maintain standard the range of data features with independent variables effectively. Even, some data features can dominate the rest by bringing all features on the same scale effectively.

1. *Encoding of categorical features:*

Considering the data set, some data are in the qualitative form that can be categorized as data in text forms. As maximum models of data are dependent on the equation of mathematical approach, calculation and incorporate numerical data set for input, hence, features can be converted into numerical data set for the categorical approach. In that aspect, the data analyses can be able to use the ratio of supervised. Even, they can use a variable binary representation and label encoding process while several categorical features are not present.

1. *Test and train data set:*

It is important to analyze and check the entire distribution of trains and test data sets are similar to each other. Then, it will be possible to conduct effective data analyzation otherwise; it will not make any kind of positive outcome.

1. *Reduction of dimensionality:*

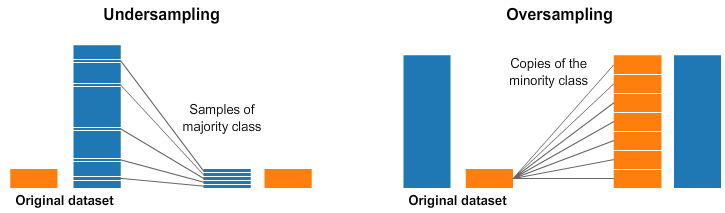
It is highly necessary to conduct preprocessing steps before dealing with the big data sets as the data set have thousand or more features. In that aspect, in this process, data scientists can be able to apply principle component technique analysis properly. With the application of the technique, the linear combination of a set of actual features can be transformed properly into a new set of features. That can be done by minimizing the high range of features area while maximum information and retained as much as possible.

In a general aspect, only 20% of the data set is mainly incorporate to conduct the test set and the 80% data set is mainly used in the training set process. Thus the reason, they can train a machine learning process on the training set. Even, the test will be conducted on the training set to identify the way of prediction of data sets properly. Moreover, it needs to shuffle the set of data in the proper way. That help the model to learns properly at single interation about the points of data with high variations.

## Class Balancing

Class Measurement is a major task for segmentation models. This is because if the correct balance is not present then the model is discriminated against and the accuracy of the model is affected. Choosing the right class balance can sometimes be difficult. Making a simple different frequency may not always work very well. Reduction of data can help, but even that will reduce all the examples of each class equally. Therefore, another way to measure the data by doing it directly, by Figure 14 below is an illustration of under and over sampling of the data.

Class balancing can be understood by considering the example of binary classification. The binary classifier that means that you are trying to classify items into two bins. The training data will contain a roughly equal number of rows for Class A and Class B but this happens only in an ideal world. Most real-world data sets contain imbalanced data which means that the number of rows of data belonging to Class A will be very different from the number of rows of data that belongs to Class B, for example, there are hundred thousand rows of data that belong to Class A and just ten thousand rows of data that belong to Class B. This is a highly imbalanced data set and if tried to make a machine learning model with this imbalanced data it's going to give very wrong results so make the machine to learn equally from both these data both these classes of data. Well one way to do this is to understand from the more dominant class so instead of using all hundred thousand rows of data from Class A randomly pick a subset of data and use that for building the model. The other approach is actually to oversample the less dominant class. These means have to create extra rows of data for the Class B or the less dominant class you can do this by simply copy paste the data bunch of times so that you are rebalancing the data or can use a more sophisticated algorithm likes mode that creates more rows of synthetic data and there is a third approach that is actually a mix of these two approaches and the hybrid approach actually uses a little bit of under sampling from the more dominant class and a little bit of over sampling from the less dominant class. To mix the two approaches have a algorithm called imbalance method.

  
*Figure 14: The classes in data and Under and Over Sampling*

In both the left and right sides of the image above, the blue class has more samples than the orange class. In this regard, have two pre-processing methods that can help train the machine learning models.

Undoing means that will only select specific data for a larger category, using only more examples than that sub-category. This selection should be made to maintain a possible distribution of the category. It was easy! simply enter the database by taking small samples.

Guessing means will build copies of the sub-category so that have the same number of multiple class examples. Copies will be made to distribute class properly. simply enter the data. Sorting can be another great way to measure class if you find that class imbalance tools are difficult to use properly.

## Feature Extraction and Selection

Then the task performed next is feature selection. This is an important step as the accuracy of the model depends a lot on the features that it is built on. The overfitting/ underfitting of the models, which means in overfitting running the machine learning with too many features and some of them are insignificant. This is treated by checking the features for their level of significance and the features that are best for training the model will be the only features kept in the data.

The classifications models work best on categorical data in form of numeric labels rather than string labels. Then encoding on the data is done, to convert such features so the model efficiency is improved. Encoding the data also helps in the fast processing of the models, when the data is huge model training part can run up to days. After all these steps the data will streamed using pysparkling and will be converted into the H2O data stream from pandas data frame. This step will be done later as the functions for data pre-processing and feature selection are not available in H2O framework.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algortihm1: **Feature Selection** | | | | | | |
|  | **n** = sample\_count | | | | | |
|  | **Selected\_features** = (Φ) | | | | | |
|  | **For** p= Ito desired\_feature\_count | | | | | |
|  |  | **For** c = 1 to feature\_count | | | | |
|  |  |  | Data \_all = data(Selectedfeatures + feature(c)); | | | |
|  |  |  | **For** i = 1 to class\_number | | | |
|  |  |  |  | **Train\_data\_class**(i) = partition(rand(Dataall(class == i), 0.75) | | |
|  |  |  |  | **Test\_data\_class**(i)= partition(rand(Data\_all(class = i)),others); | | |
|  |  |  |  | [**data**(i)) = chisquare(train\_data\_dass(i)); | | |
|  |  |  | **End for i** | | | |
|  |  |  | **Test\_data\_all** = merge(test\_data\_class); | | | |
|  |  |  | **For** i size(Test\_data\_all) | | | |
|  |  |  |  | **For** k = 1: class\_count | | |
|  |  |  |  |  | **For** j count(data(k)) | |
|  |  |  |  |  |  | **distance**((, j) = oklid\_distance(data k)-test\_data (i); |
|  |  |  |  |  | **End for j** | |
|  |  |  |  |  | **min\_dist**(k) = min(distance(k)); | |
|  |  |  |  | **End for k** | | |
|  |  |  |  | (**result\_of\_test\_data**(i), class of test data(i)] = min(min\_dist); | | |
|  |  |  | **End for**; | | | |
|  |  |  | **Performance\_criteria**(feature(c)) = sum(class\_of\_test\_data(i) == class\_of\_test\_data\_expected), | | | |
|  |  | **End for c** | | | | |
|  |  | **Best\_feature**(c)= arg max(performance\_criteria(feature(c)) | | | | |
|  |  | **Selected\_features** = Selected\_features + best\_feature(c) | | | | |
|  | **End for p** | | | | | |

## Model Training

Then the models Deep Learning (Neural Networks), Distributed Random Forest (DRF), Generalized Additive Models (GAM), Gradient Boosting Machine (GBM), Rule Fit, Stacked Ensembles, and XGBoost will be implemented. For the stacked ensemble model which is a combination of two models one stacked over the other will be created with two best performing model. This will also help us evaluate if the result of stack is better than the individual model score. Which are present the H2O.ai framework will be implemented using python language. The use of these models adds to the novelty of the research under study.

# Deep Learning (Neural Networks)

H2O’s Deep Learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation. The network can contain a large number of hidden layers consisting of neurons with tanh, relu, activation functions. Advanced features such as adaptive learning rate, rate annealing, momentum training, dropout, L1 or L2 regularization, checkpointing, and grid search enable high predictive accuracy. Each compute node trains a copy of the global model parameters on its local data with multi-threading (asynchronously) and contributes periodically to the global model via model averaging across the network.

A feedforward artificial neural network (ANN) model, also known as deep neural network (DNN) or multi-layer perceptron (MLP), is the most common type of Deep Neural Network and the only type that is supported natively in H2O-3. Several other types of DNNs are popular as well, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). MLPs work well on transactional (tabular) data; however if you have image data, then CNNs are a great choice. If you have sequential data (e.g. text, audio, time-series), then RNNs are a good choice.

# Distributed Random Forest (DRF)

Distributed Random Forest (DRF) is a powerful classification and regression tool. When given a set of data, DRF generates a forest of classification or regression trees, rather than a single classification or regression tree. Each of these trees is a weak learner built on a subset of rows and columns. More trees will reduce the variance. Both classification and regression take the average prediction over all of their trees to make a final prediction, whether predicting for a class or numeric value. (Note: For a categorical response column, DRF maps factors (e.g. ‘dog’, ‘cat’, ‘mouse) in lexicographic order to a name lookup array with integer indices (e.g. ‘cat -> 0, ‘dog’ -> 1, ‘mouse’ -> 2.)

DRF no longer has a special-cased histogram for classification (class DBinomHistogram has been superseded by DRealHistogram) since it was not applicable to cases with observation weights or for cross-validation.

# Generalized Boosting Model (GBM)

Gradient Boosting Machine (for Regression and Classification) is a forward learning ensemble method. The guiding heuristic is that good predictive results can be obtained through increasingly refined approximations. H2O’s GBM sequentially builds regression trees on all the features of the dataset in a fully distributed way - each tree is built in parallel.

The current version of GBM is fundamentally the same as in previous versions of H2O (same algorithmic steps, same histogramming techniques), with the exception of the following changes:

* Improved ability to train on categorical variables (using the nbins\_cats parameter)
* Minor changes in histogramming logic for some corner cases

There was some code cleanup and refactoring to support the following features:

* Per-row observation weights
* Per-row offsets
* N-fold cross-validation
* Support for more distribution functions (such as Gamma, Poisson, and Tweedie)

# RuleFit

H2O’s Rulefit algorithm combines tree ensembles and linear models to take advantage of both methods: the accuracy of a tree ensemble and the interpretability of a linear model.

The general algorithm fits a tree ensemble to the data, builds a rule ensemble by traversing each tree, evaluates the rules on the data to build a rule feature set, and fits a sparse linear model (LASSO) to the rule feature set joined with the original feature set.

# Stacked Ensembles

Ensemble machine learning methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Many of the popular modern machine learning algorithms are actually ensembles. For example, [Random Forest](http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html) and [Gradient Boosting Machine (GBM)](http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html) are both ensemble learners. Both bagging (e.g. Random Forest) and boosting (e.g. GBM) are methods for ensembling that take a collection of weak learners (e.g. decision tree) and form a single, strong learner.

H2O’s Stacked Ensemble method is a supervised ensemble machine learning algorithm that finds the optimal combination of a collection of prediction algorithms using a process called stacking. Like all supervised models in H2O, Stacked Ensemeble supports regression, binary classification, and multiclass classification.

|  |  |  |
| --- | --- | --- |
| Algortihm2: **Stacked Ensemble** | | |
|  | **Input**: Training data | |
|  | **Output**: An ensemble classifier H | |
|  | **Step 1**: Learn first-level classifiers | |
|  | **for** t ← 1 to T | |
|  |  | Learn a base classifier ht based on Dk |
|  | **end for** | |
|  | **Step 2**: Construct new data sets from D | |
|  | **for** I ← 1 to m do | |
|  |  | Construct a new data set that contains where |
|  | **end for** | |
|  | **Step 3**: Learn a second-level classifier to: Learn a new classifier h' based on the newly constructed data set | |
|  |  | return H(x)= h' (h1(x), h2(x)), , hT (x)) |

# XGBoost

XGBoost is a supervised learning algorithm that implements a process called boosting to yield accurate models. Boosting refers to the ensemble learning technique of building many models sequentially, with each new model attempting to correct for the deficiencies in the previous model. In tree boosting, each new model that is added to the ensemble is a decision tree. XGBoost provides parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. For many problems, XGBoost is one of the best gradient boosting machine (GBM) frameworks today.

The H2O XGBoost implementation is based on two separated modules. The first module, h2o-genmodel-ext-xgboost, extends module [h2o-genmodel](https://github.com/h2oai/h2o-3/tree/master/h2o-genmodel) and registers an XGBoost-specific MOJO. The module also contains all necessary XGBoost binary libraries. The module can contain multiple libraries for each platform to support different configurations (e.g., with/without GPU/OMP). H2O always tries to load the most powerful one (currently a library with GPU and OMP support). If it fails, then the loader tries the next one in a loader chain. For each platform, H2O provide an XGBoost library with minimal configuration (supports only single CPU) that serves as fallback in case all other libraries could not be loaded.

## Model Evaluation

The performance of these algorithms will be captured and then compared to declare the best performing model. The models will be compared on basis evaluation metrics like, accuracy, F1- score, time of processing and AUC.

# AUC

AUC stands for Area under the ROC Curve. That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example

# F1-Score

F1-score is the weighted average of precision and recall. F1-score takes both the false positive and false negative values into consideration. F1-score is used to prevent the model from false alarms. The highest value of F1-score can be 1. It can be formulated as :

### *F1-Score = 2 × ( Precision × Recall ) ⁄ ( Precision + Recall )*

# Accuracy

Accuracy is the measure of correctly predicted observation to total observations. It is very helpful when the false positive and false negative values are almost the same. The accuracy can be formulated as :

### *Accuracy = (TP + TN) / (TP +FP + FN + TN)*

### Where TP represents the True positives, FP represents False positives, TN represents True negative and FN represents the false negative values.

# Training Time

Every algorithm in machine learning takes different amounts of time to train the model with respect to the dataset. In our experiment we evaluate and compare machine learning algorithms based on distributed and standalone platform with respect to training time as well.

# Experiments and Results

## Experimental Setting

The experimental setting section describes the details of the experiments carried out in this research. This section comprises of the H2O environment setup, data preparation whereby the exploratory data analysis, data preprocessing, class balancing, label encoding and so on.

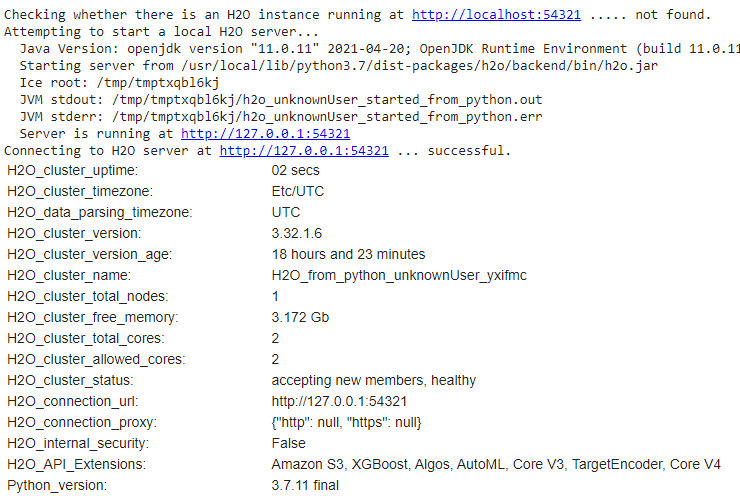
## H2O Sparkling Water Setup

Sparkling Water framework gives the user the capability of the Spark being fast and scalable into the machine learning algorithms. Sparkling Water gives the users power to drive the computation from the languages like Scala, R and Python. It utilizes the H2O Flow UI to generate an ideal machine learning platform for the developers.

Sparkling Water contains the same features and functionality as H2O but provides a way to use H2O with Spark, a large-scale cluster framework.

Sparkling Water is ideal for H2O users who need to manage large clusters for their data processing needs and want to transfer data from Spark to H2O (or vice versa).

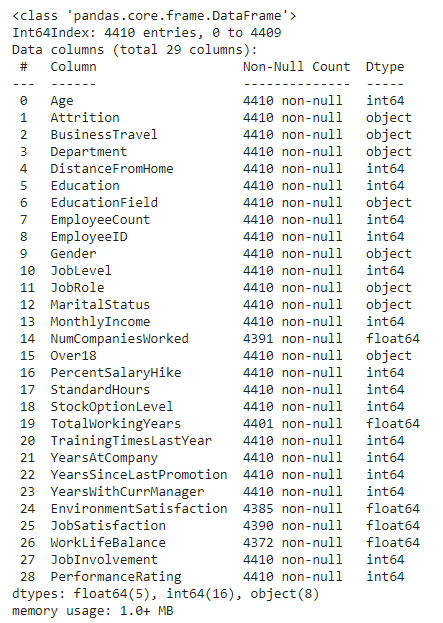
There is also a Python interface available to enable access to Sparkling Water directly from PySpark.



*Figure 15: The H2O environment setup*

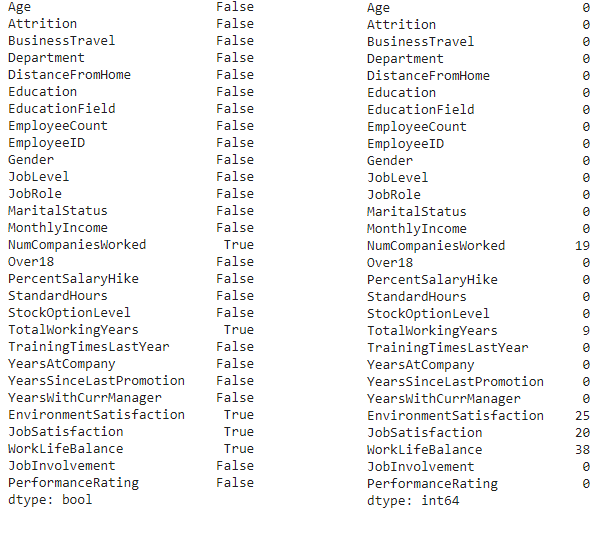
## Data Preparation

The dataset that is in this research belongs to research done by IBM. It has been divided into 5 files. The research is based on 3 files general data, manager survey and employee survey. The other two files are having the data regarding the in and out the time of the employee. The data is the merged version of the three files and has the shape 4410x 29 where the number of rows is 4410 and 29 columns with the data of the employees. This data includes features as 'Age', ' Attrition ', ' BusinessTravel ', ' DistanceFromHome ', 'JobLevel', 'MonthlyIncome', and more as shown in the Figure 16 below.



*Figure 16: Data information*

Then the process to check the data for the missing values is done. In this step, the result is generated in two steps. The first part is when the data column is checked if the missing value is present in that column. The result is shown below. The value False in the column represents there are no null values in the data. The value True in the column represents there are null values in the data. The columns NumCompaniesWorked, TotalWorkingYears, EnvironmentSatisfaction, JobSatisfaction and WorkLifeBalance has the result True. The second part of figure 17 below shows the count of the missing values present in the data. The column NumCompaniesWorked has 19 missing values, TotalWorkingYears has 9, EnvironmentSatisfaction has 25, JobSatisfaction has 20 and WorkLifeBalance has 30 missing values in the data.

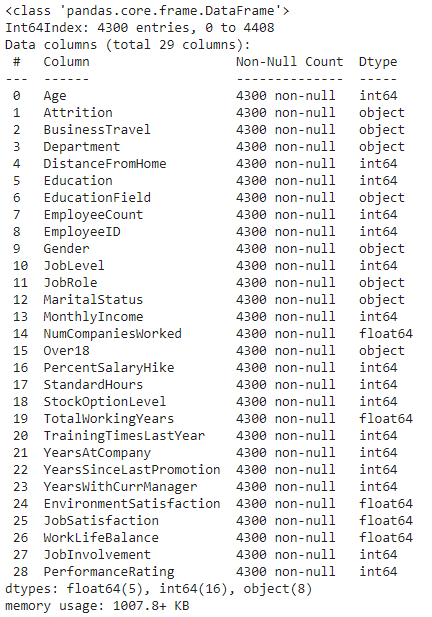


*Figure 17: Data missing value check*

The missing values in the data can be treated in two ways:

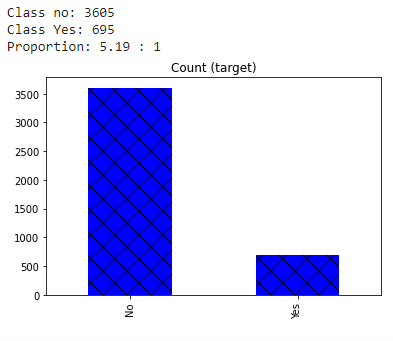
* Replacing the null with a default value
* Removing the data row with missing column

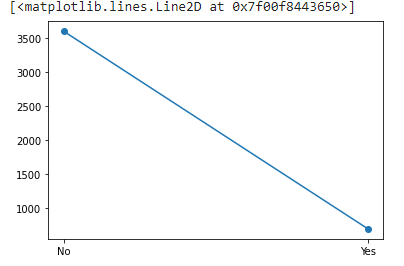
The default value approach can lead to a scenario where the data might lose its credibility. This is because the columns like EnvironmentSatisfaction, JobSatisfaction and WorkLifeBalance can not be filled with any default value as it entirely depends on the employees perspective. Thus, the rows the missing data are dropped from the data, the data information after dropping the rows is shown in figure 18 below.



*Figure 18: Data information*

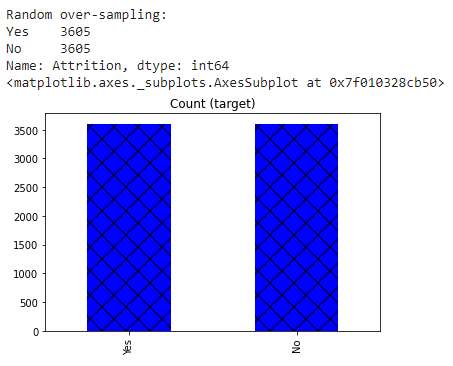
The data is then checked for the balance of the classes. The data under study has two classes: Yes and No. The Yes class signify that the employee with attrite and has 695 rows in data. Class No means that the employee will not attrite. The No class has more values 3605 than the Yes class. This shows the balance of the classes is off, as shown in the figure 19 below.



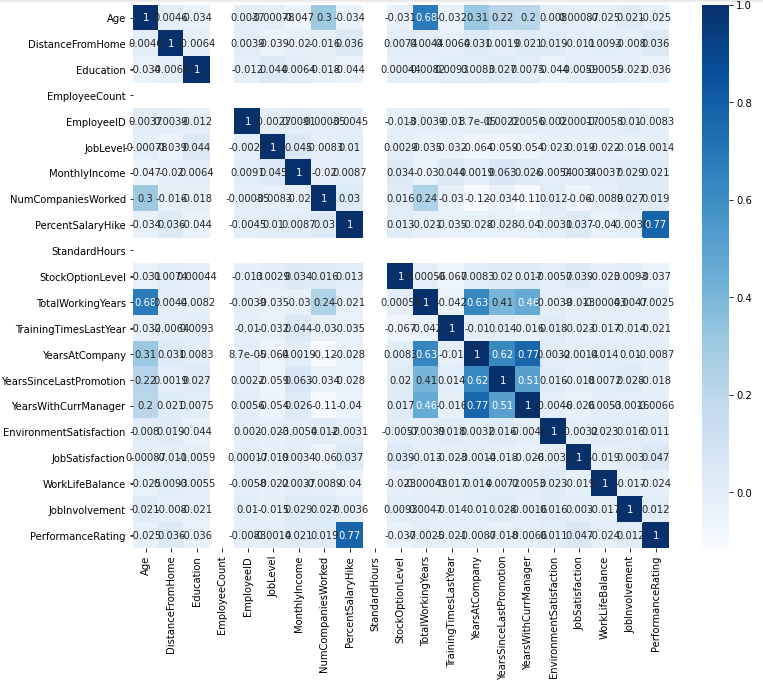


*Figure 19: Class balancing*

The class balancing can be fixed in two ways, either undersample the data or oversample the data. If in the research, undersampling is used, then the classes count will be twice(number of classes) of 695. If in the research, oversampling is used, then the classes count will be twice(number of classes) of 3605. Thus oversampling is applied. The classes after random oversampling are shown in the figure 20 below.

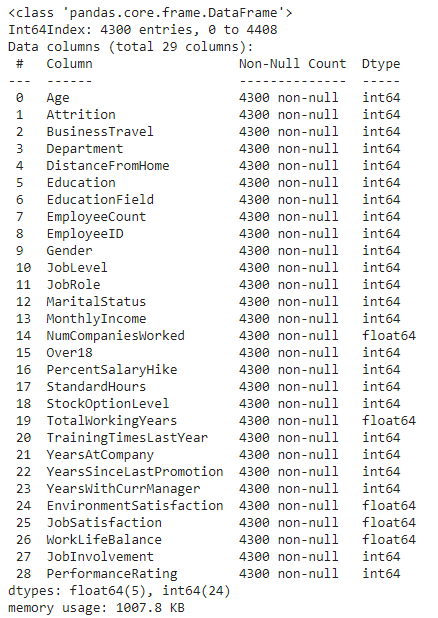


*Figure 20: Data classes after Random Oversampling*



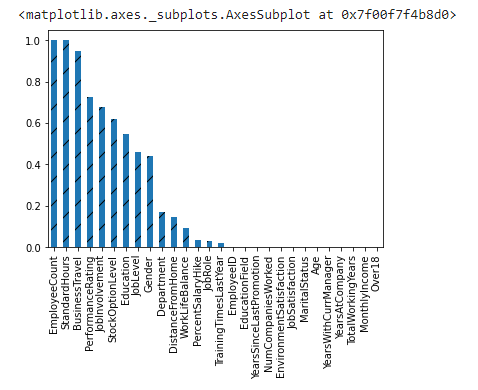
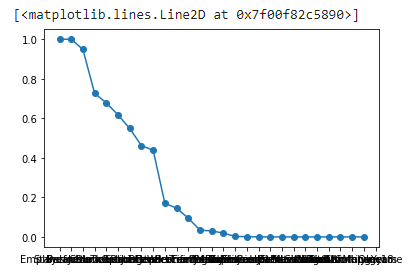
*Figure 21: Correlation Heatmap*

The correlation heatmap of the data is plotted to check for correlation. The correlation heatmap is shown is figure 21 above. In the plot, the darker shades represent the higly correlated features. This lays the ground for feature selection where by the step of removing the non-significant i.e. highly correlated features is done.



*Figure 22: Data information*

Then label encoding on the data is performed. This is done to encode the data of object type which implies the data is in the form of string or character. This can be understood by the attritition column which has only two values Yes and No, so the data will . The encoding helps in faster execustion of the data. The Figure 22 above shows the type of all the columns is now of type integer and float.



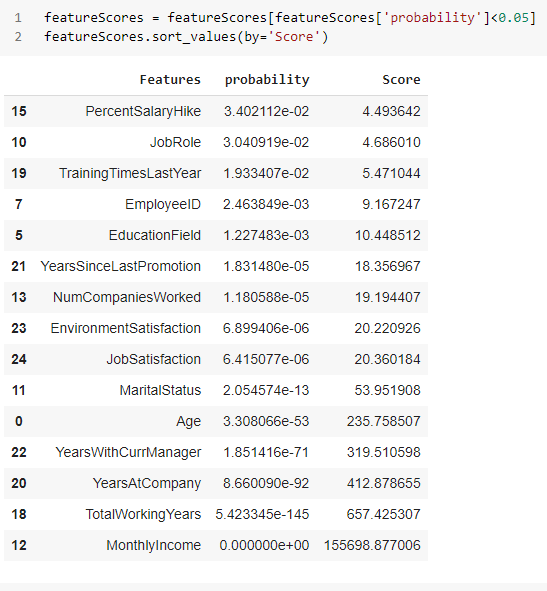
*Figure 23: Chi-square test to find the best feature*

Then chi-square test on the data is performed. This is done to find the significant features. The plots above in figure 23 shows the probability bars or points. The probability here signifies that these columns are more probable of beings highly correlated and thus are less significant and should be dropped from the data. The figure 24 below shows the probability scores and the chi-score for all the features.



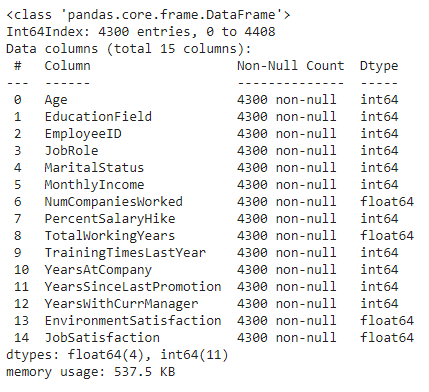
*Figure 24: Chi-square test scores*

Figure 25 below shows the probability scores and the chi-score for all the features which has a probability score of less than 0.05. This enables getting the features list of the features which are significant and should be kept in the data.



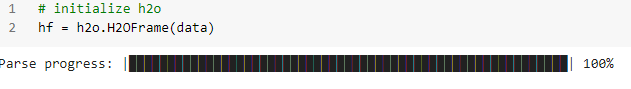
*Figure 25: Chi-square test results*

Figure 26 below shows the data information after removing the insignificant columns.



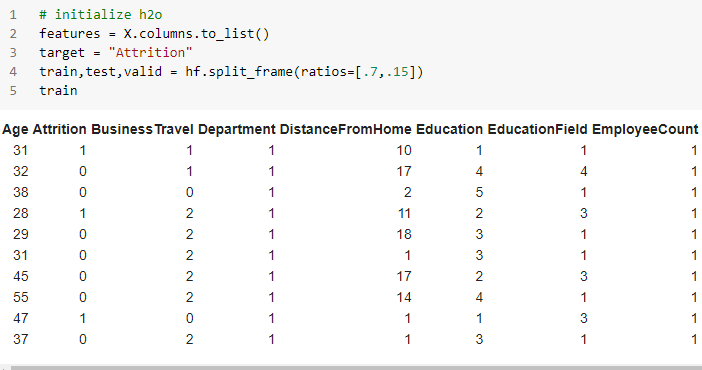
*Figure 26: Data information*

The data is now ready for the model building. The data is now loaded into the H2O Frame. This is done as H2O won’t be able to recognise all the pandas Dataframe. The data is loaded later in the H2O framework as there are certain functions for data preprocessing in the H2O framework that is missing. This is shown in Figure 27 below.



*Figure 27: Data H2O streaming*

Figure 28 below shows the feature vector, target vector set for the data. The train, test and valid data where the training part of the data constitute 70% of the data and the test and valid data both had remaining 30% data split into 15% ratio each.



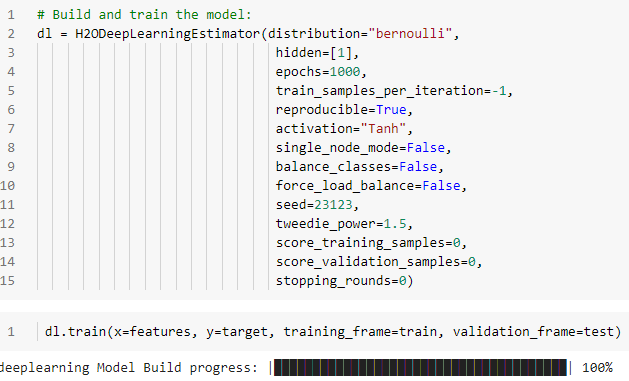
*Figure 28: H2O data*

## Models

## Deep Learning

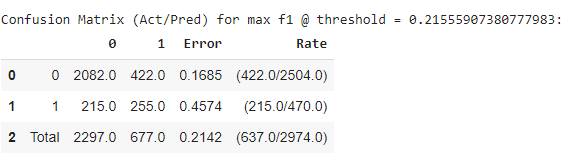
The deep learning model in H2O framework built with Bernoulli distribution, 100 epochs.

The process is shown in the figure 29 below.



*Figure 29: Deep learning model bulid*

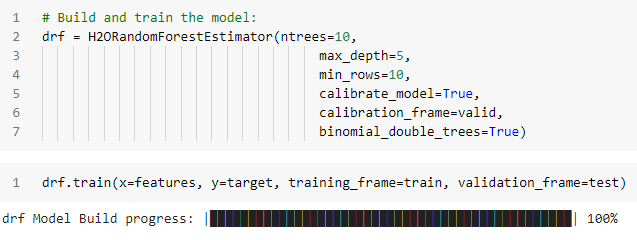
The confusion matrix for the model output is shown in the figure below. The confusion matrix shows the actual v/s predicted value. The true positive of the model is 2082, the false positive of the model is 422, the true negative is 255 and the false negative is 215. This shows that model is predicting the true positive class more accurately.



*Figure 30: confusion matrix of Deep learning*

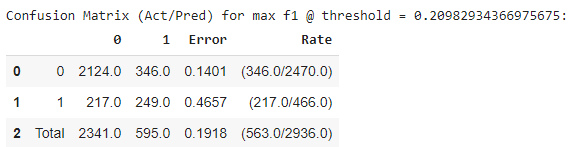
## Distributed Random Forest

The distributed random forest trees model in H2O framework built with 10 trees, max depth of 5, minimum rows 10, calibration true with valid frame and binomial double trees true. The process is shown in figure 31 below.



*Figure 31: Distributed Random Forest model built*

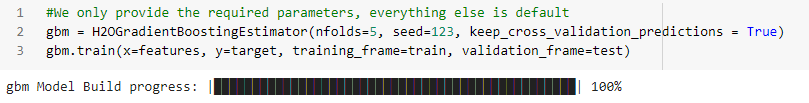
The confusion matrix for the model output is shown in the figure below. The confusion matrix shows the actual v/s predicted value. The true positive of the model is 2124, the false positive of the model is 246, the true negative is 217 and the false negative is 249. This shows that model is predicting the true positive class more accurately.



*Figure 32: confusion matrix of DRF*

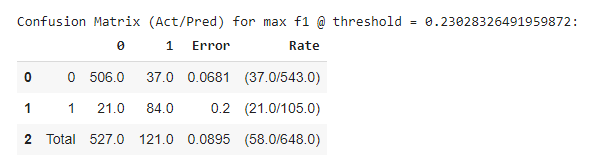
## Gradient Boosting Machine

The gradient boosting machine model in H2O framework built with 5 number of folds, and keept the cross validation true. The process is shown in the figure 29 below.



*Figure 33: GBM model bulid*

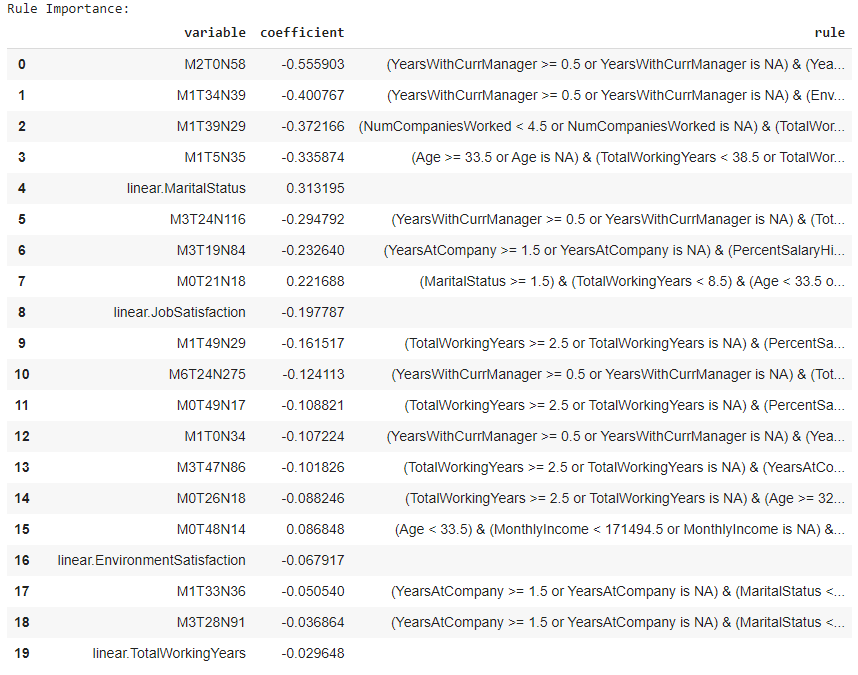
The confusion matrix for the model output is shown in the figure below. The confusion matrix shows the actual v/s predicted value. The true positive of the model is 506, the false positive of the model is 37, the true negative is 21 and the false negative is 2. This shows that model is predicting true positive class and true negative accurately.



*Figure 34: confusion matrix of GBM*

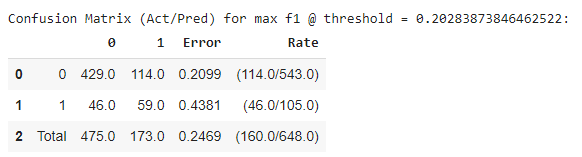
## RuleFit

The rulefit model in H2O framework built with the set of rules. These rules are shown in the figure 35 below.



*Figure 35: Rulefit model rules*

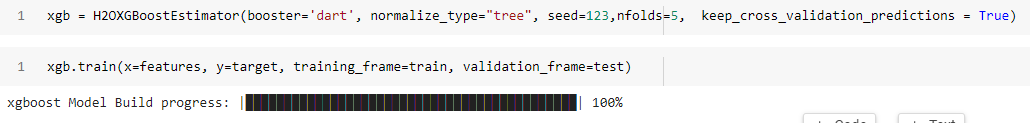
The confusion matrix for the model output is shown in the figure below. The confusion matrix shows the actual v/s predicted value. The true positive of the model is 429, the false positive of the model is 114, the true negative is 59 and the false negative is 46. This shows that model is predicting true positive class more. Accurately.



*Figure 36: confusion matrix of Rulefit*

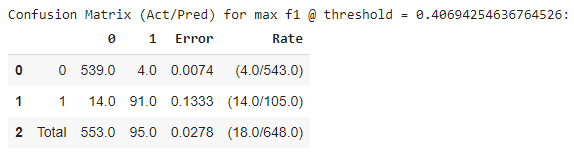
## XGBoost

The xgboost model in H2O framework built with dart booster, tree as normalization type, n folds 5 and kept cross-validation true. The process is shown in figure 37 below.



*Figure 37: xgboost model built*

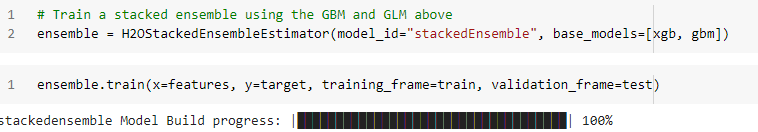
The confusion matrix for the model output is shown in the figure below. The confusion matrix shows the actual v/s predicted value. The true positive of the model is 539, the false positive of the model is 4, the true negative is 91 and the false negative is 14. This shows that model is predicting both true positive class and rue negative class more accurately.



*Figure 38: confusion matrix of xgboost*

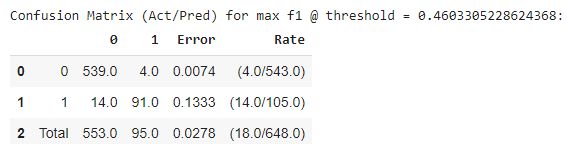
## Stacked Ensemble

The stacked ensemble model in H2O framework built with xgboost classifier and gradient boosted machine classifier. These models are chosen based on their performance in classifying both true positive and true negative efficiently. The process is shown in the figure 39 below.



*Figure 39: stacked ensemble model*

The confusion matrix for the model output is shown in the figure below. The confusion matrix shows the actual v/s predicted value. The true positive of the model is 539, the false positive of the model is 4, the true negative is 91 and the false negative is 14. This shows that model is predicting true positive and true negative classes more accurately.



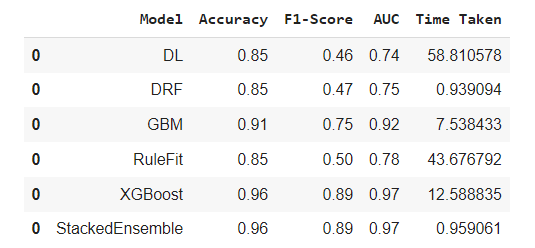
*Figure 40: confusion matrix of stacked ensemble*

## Evaluation

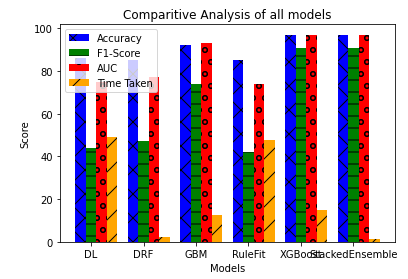
### 

### 4.3.1 Comparison of Precision, Recall and F1-score

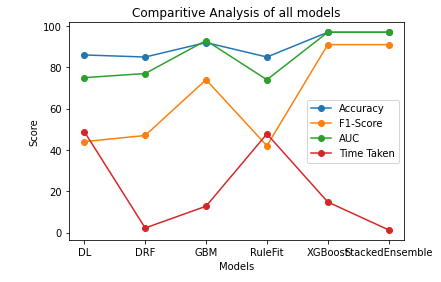
From the table below in figure 41, the details of all the model's performance can be seen. The best performing model in all metrics is Stacked Ensemble. The model scored 96% accuracy, 89% f1-score, 97 AUC score and 0.959061 ms as the time taken by the model. The model Deep learning has an accuracy of 0.85, F1-score of 0.46, AUC of 0.74 and it took 58.81ms for execution. The model Distributed Random Forest has an accuracy of 0.85, F1-score of 0.47, AUC of 0.75 and it took 0.93 ms for execution. The model Gradient Boosting Machine has an accuracy of 0.91, F1-score of 0.75, AUC of 0.92 and it took 7.53 for execution. The model Rule Fit has an accuracy of 0.85, F1-score of 0.50, AUC of 0.78 and it took 43.67ms for execution. The model XGBoost has an accuracy of 0.96, F1-score of 0.89, AUC of 0.97 and it took 12.58ms for execution. The model stacked Ensemble has an accuracy of 0.96, F1-score of 0.89, AUC of 0.97 and it took 0.95ms for execution. These results are shown in a bar plot in the figure 42 below and in line plot in figure 43 below.



*Figure 41: Model result*



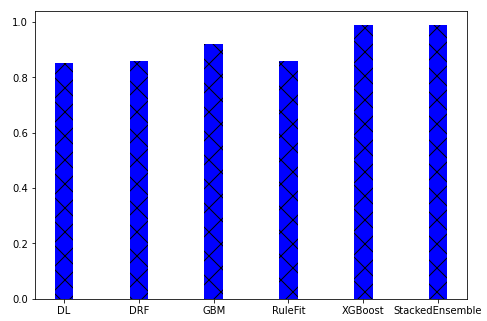
*Figure 42: Model result in barplot for comparitive analysis*



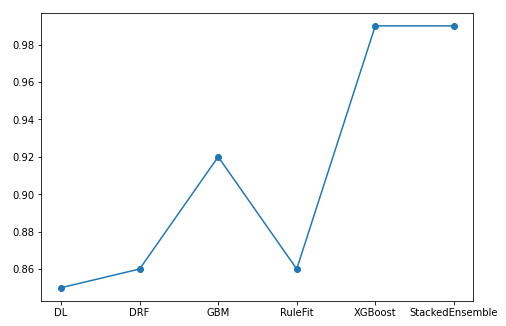
*Figure 43: Model result in line plot for comparative analysis*

### Accuracy Comparison

The figure below, shows that XGBoost and stacked ensemble are highest performing model. The second best model is Gradient Boosted machine. The model Deep Learning, DRF,and RuleFit came third as they all have similar performance bar.



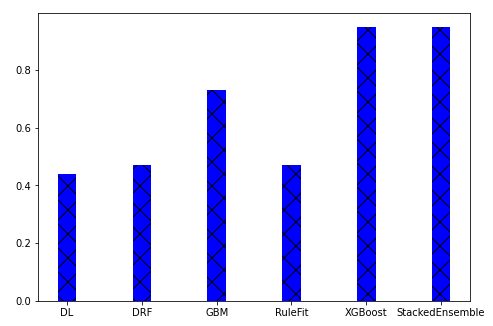
*Figure 44: Accuracy result in bar plot for comparative analysis*



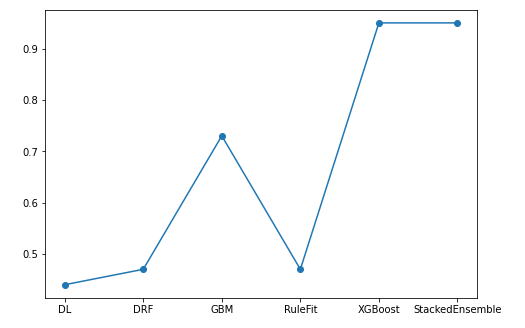
*Figure 45: Accuracy result in line plot for comparative analysis*

### F1-Score Comparison

The figure below, shows that XGBoost and stacked ensemble are highest performing model. The second best model is Gradient Boosted machine. The model DRF,and RuleFit came third as they all have similar performance bar. Deep Learning has the lowest F1-Score.



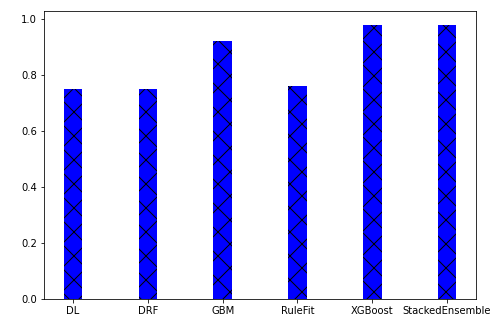
*Figure 46: F1- Score result in bar plot for comparative analysis*



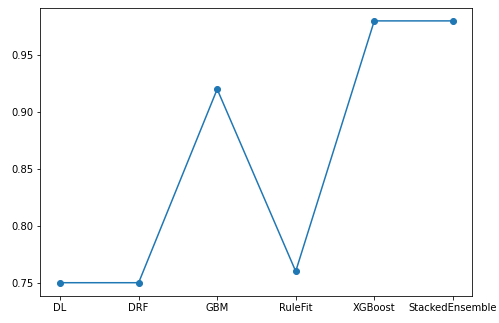
*Figure 47: F1- Score result in line plot for comparative analysis*

### AUC Comparison

The figure below, shows that XGBoost and stacked ensemble are highest performing model. The second best model is Gradient Boosted machine. Rulefit model stand third with his performance. The models Deep Learning, and DRF came third as they all have similar performance bar.



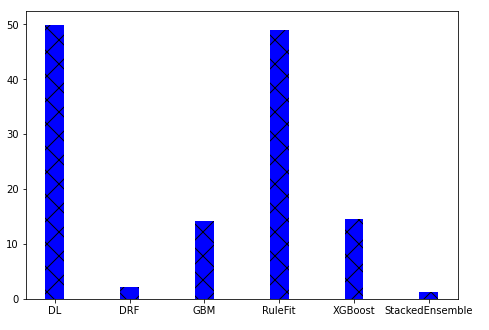
*Figure 48: AUC - Score result in bar plot for comparative analysis*



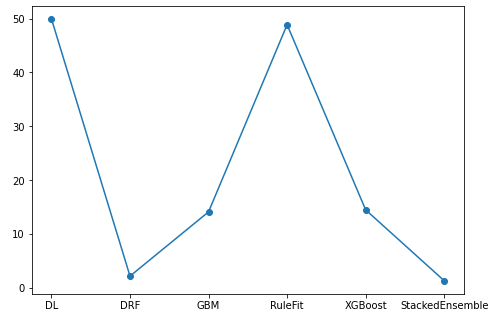
*Figure 49: AUC - Score result in line plot for comparative analysis*

### 4.3.5 Training Time Comparison

The figure below, shows that RuleFit and Deep Learning are highest time complexity. The second place in time complex model are GBM and XGBoost. The DRF model and Stcked Ensemble comes next.



*Figure 50: Time taken result in bar plot for comparative analysis*



*Figure 51: Time taken result in line plot for comparative analysis*

# Conclusion

## Contributions

The experiments can be concluded with the result that the Stacked Ensemble model works best concerning all the metrics that are taken. The model has an accuracy score of 99%, F1-score of 95% and AUC of 98%. The balance between the accuracy score and F1-Score is also good, there is not much difference between these. This concludes that the model is free from issues of class balancing and overfitting and Underfitting of the data. The AUC is also higher the better, so the AUC for the model is also good. The XGBoost has a similar performance. This is due to the reason that Stacked Ensemble has XGBoost and Gradient Machine models stacked. Though the XGBoost model takes a lot more time exactly 14.486 than Stacked Ensemble. The time taken by Stacked Ensemble is 1.319 which is the lowest score.

The research concluded that the companies can predict employee’s attrition using the historical data kept in the human resources department. The data can be run through the model for training and prediction purposes. The list of employees that may attrite can be shown as the result.

The research also finds to find the best performing deep learning model from Deep Learning (Neural Networks), Distributed Random Forest (DRF), Boosting Machine (GBM), Rule Fit, Stacked Ensembles, and XGBoost performs better for employee attrition prediction in H2O environment. From the experiments conducted it can be concluded that the Stacked Ensemble model works better for the employee attrition prediction. The model has the best scores in the evaluation metrics and also has the best execution time, the fastest and more efficient model. This is the novelty of the research that using the H2O framework is the best performing model and fastest model for the employee attrition prediction.

## Future Work

# The work that can be done as an extension to this research is to implement it using the Hadoop architecture. This sparkling water framework that is followed in the project is based on the spark that is a part of Hadoop architecture. The further use of tools and techniques from the Hadoop architecture can give good results.

# It is also recommended to work with advanced and more complex neural networks, as the neural networks are working efficiently in a majority of sectors. Neural networks are a part of artificial intelligence, which has revolutionalised the world with its technology. Artificial intelligence works like a human brain considering all the scenarios and thinking over again and again. This helps the model learn and it learns with every good and bad prediction made. This could help in building a good employee attrition prediction system.

# References

1. business-science.io (2017). *HR Analytics: Using Machine Learning to Predict Employee Turnover*. [online] Business Science. Available at: <https://www.business-science.io/business/2017/09/18/hr_employee_attrition.html>.
2. Fallucchi, F., Coladangelo, M., Giuliano, R. and William De Luca, E. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, 9(4), p.86.
3. Mohbey, K.K. (2020). Employee’s Attrition Prediction Using Machine Learning Approaches. *Machine Learning and Deep Learning in Real-Time Applications*, pp.121–128.
4. ‌Bhatnagar, S., Upadhyay, S., Gupta, S.P., Bhole, G.V. and Singh, P. (2017). Image Textural Style Transfer using Neural Network. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(5), pp.163–165.
5. Punnoose, R. and Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. International Journal of Advanced Research in Artificial Intelligence, 5(9).
6. Al-Radaideh, Q. A., & Al Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employees performance. International Journal of Advanced Computer Science and Applications, 3(2).
7. Alao, D., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics and Allied Research Journal, 4.
8. Alduayj, S. S., & Rajpoot, K. (2018). Predicting Employee Attrition using Machine Learning. 2018 International Conference on Innovations in Information Technology (IIT), 93–98. 10.1109/INNOVATIONS.2018.8605976.
9. Cortes, C., &Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. doi:10.1007/ BF00994018.
10. Ho, T. K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278–282.
11. Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm. International Journal on Computer Science and Engineering, 2(8), 2526–2534.
12. Kaur, S., & Vijay, M. R. (2016). Job satisfaction-A major factor behind attrition of retention in retail industry. Imperial Journal of Interdisciplinary Research, 2(8), 993–996.
13. McKinley Stacker, I. V. (2015). IBM waston analytics. Sample data: HR employee attrition and performance [Data file]. McKinley Stacker.
14. Moncarz, E., Zhao, J., & Kay, C. (2009). An exploratory study of US lodging properties’ organizational practices on employee turnover and retention. International Journal of Contemporary Hospitality Management, 21(4), 437–458. doi:10.1108/09596110910955695.
15. Morgan, J. N., &Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association, 58(302), 415–434. doi:10.1080/01621459.1963.10500855.
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & ... . (2011). Scikitlearn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
17. Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Academic Press.
18. Rane, A., & Kumar, A. (2018, July). Sentiment Classification System of Twitter Data for {US} Airline Service Analysis. In 2018 {IEEE} 42nd Annual Computer Software and Applications Conference ({COMPSAC}). IEEE. 10.1109/compsac.2018.00114.
19. Raschka, S. (2015). Python machine learning. Packt Publishing Ltd.
20. Saradhi, V. V., &Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999–2006. doi: 10.1016/j.eswa.2010.07.134.
21. Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., & Smith, A. M. (2005). Employee turnover: A neural network solution. Computers & Operations Research, 32(10), 2635–2651. doi: 10.1016/j.cor.2004.06.022.
22. Sikaroudi, E., Mohammad, A., Ghousi, R., &Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). Journal of Industrial and Systems Engineering, 8(4), 106–121.
23. Tzeng, H.-M., Hsieh, J.-G., & Lin, Y.-L. (2004). Predicting Nurses’ Intention to Quit with a Support Vector Machine: A New Approach to Set up an Early Warning Mechanism in Human Resource Management. CIN: Computers, Informatics, Nursing, 22(4), 232–242. PMID:15494654.
24. Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. Proceedings of SAI Intelligent Systems Conference, 737–758.
25. Pujahari, A. and Sisodia, D.S. (2019). Modeling Side Information in Preference Relation based Restricted Boltzmann Machine for recommender systems. *Information Sciences*, 490, pp.126–145.
26. Exit interview -A Competency based Practice for Employee Retention. (2019). *International Journal of Innovative Technology and Exploring Engineering*, 8(12), pp.2707–2711.
27. ‌Mahesh, K.S. (2017). Employee Attrition and Retention Strategies in Amara Raja Batteries Limited: An Empirical Investigation. *International Journal of Scientific Research and Management*.
28. Customer Attrition and Retention. (2020). International Journal of Innovative Technology and Exploring Engineering, 9(7), pp.925–927.
29. Psychological Contract and Attrition in Information Technology Sector. (2019). International Journal of Innovative Technology and Exploring Engineering, 8(12), pp.3951–3956.
30. Importance of Social Responsibility: An Employee Standpoint. (2019). *International Journal of Innovative Technology and Exploring Engineering*, 8(11S), pp.274–279.
31. ‌Sivakumar, B. (2021). DIGITAL INDIA - A KEY TO SMART India. *PARIPEX INDIAN JOURNAL OF RESEARCH*, pp.15–16.
32. Mahesh, K.S. (2017). Employee Attrition and Retention Strategies in Amara Raja Batteries Limited: An Empirical Investigation. *International Journal of Scientific Research and Management*.
33. ‌Attrition and Retention Practices in Multi-National Companies (MNCs) in India. (2016). *International Journal of Science and Research (IJSR)*, 5(4), pp.1821–1824.
34. Kamath, Dr.R.S., Jamsandekar, Dr.S.S. and Naik, Dr.P.G. (2019). Machine Learning Approach for Employee Attrition Analysis. *International Journal of Trend in Scientific Research and Development*, Special Issue (Special Issue-FIIIIPM2019), pp.62–67.
35. ‌Murugan, Dr.J.S.V. and Murugan, S.B. (2011). Employee Attrition And Retention In Private Insurance Sector A HRM Challenge. *Indian Journal of Applied Research*, 1(5), pp.115–117.
36. ‌Srivastava, P.R. and Eachempati, P. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction. *Journal of Global Information Management*, 29(6), pp.1–29.
37. Sharma, B. and Garg, N.K. (2015). Impact of Self-Efficacy on Employee Attrition: A Study Through Structural Equation Modeling. *SSRN Electronic Journal*.
38. Mann, Y.N. and Mann, K.N. (2016). Employee Attrition in Educational Institutions in India Causes and Remedies. *SSRN Electronic Journal*.
39. Sharma, A., Sharma, K.R. and Borah, S.B. (2011). Do you Love Me? Attrition During Probation Period: Employee vs Employer. *Prabandhan: Indian Journal of Management*, 4(7), p.8.
40. ‌Sharma, A., Sharma, K.R. and Borah, S.B. (2011). Do you Love Me? Attrition During Probation Period: Employee vs Employer. *Prabandhan: Indian Journal of Management*, 4(7), p.8.
41. ‌Geetha, Y., Sravanthi, Mrs. and Rangaiah, Dr.V. (2019). Employee Satisfaction At Gold Stone Technologies Limited. *Think India*, 22(2), pp.71–76.
42. Umamaheswari, S. and Krishnan, J. (2014). Impact of Work Environment and Organisation Commitment with Employee Retention in Ceramic Industries in India. *Siddhant- A Journal of Decision Making*, 14(4), p.337.

‌

‌

‌

‌

‌

‌

# Appendix