PREDICTION OF INCREASED ATTRITION RATE OF EMPLOYEE USING PEOPLE ANALYTICS

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Computational Engineering

2022

Muhammad Bilal Hassan

Examiners: Associate Professor Lassi Roininen

Associate Professor Matylda Jablonska-Sabuka

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Computational Engineering

Muhammad Bilal Hassan

Prediction of Increased Attrition Rate of Employee Using People Analytics

Master's Thesis

2022

48 pages, 22 figures, 1 table

Examiners: Associate Professor Lassi Roininen and Associate Professor Matylda Jablonska-Sabuka

Keywords: People Management, People Analytics, Employee Turnover, Employee Satisfaction, Attrition Rate, Predictive Analytics, Business Performance, Employee Retention

The major aim of this thesis is to explore the use of people analytics in people management. This study is a quantitative research which has been conducted using a certain employee data from an online website Kaggle. The machine learning algorithm used for carrying out the analysis is decision tree algorithm. The aim is to determine attrition rate of employees in an organization with the help of people analytics. Therefore, machine learning algorithms have been used to analyze the employee data and find out how the attrition rate can be reduced using data visualizations.

The findings of this research reveal that there are several reasons behind increased attrition rate of employees, which involves different attributes such as satisfaction level of employees, number of projects done by an employee, average monthly salary, time spent at the company, workplace accidents, promotions, and sales. From the findings of this research, it is clear that people analytics improves the attrition rate by identifying the different attributes on which the organization has to work, therefore it is recommended for the companies to use people analytics for managing increased attrition rate of employees as well as train them and perform other HR functions with the help of predictive analytics.

ACKNOWLEDGEMENTS

First of all, I would like to pay warm regards to LUT University for providing me with this opportunity to do my Master's degree. I have improved my research skills, gained more knowledge about my domain and improved my critical thinking ability to advance my career and grow as a professional. The time that I have spent at LUT, I have made new friends and gained more knowledge and understanding, made valuable memories and experiences that I will remember forever. LUT University has treated me really well, therefore I will cherish the good times till the end of my life.

I would further like to thank my supervisor Lassi Roininen and Matylda Jablonska-Sabuka for their continuous efforts and guidance throughout this project and providing valuable feedback for this work of writing which has led me to achieve the objectives of this project successfully. Lastly, I would like to thank my parents, friends and late brother who have always believed in me and supported me to achieve my goals in life and helped me to become more career oriented.

28 April 2022

Table of contents

1 Introduction

For understanding the value of people in an organization, the major terminologies used are human resource or intellectual capital. Every company is considering these aspects. The return on investment of an organization as well as its profitability are dependent on the quality of workforce that it has. People management is a series of practices which comprise of end-to-end processes of talent optimization, talent acquisition, as well as talent retention while offering a continual support for firms and a direction for employees of the firm. People management, as a chief sub-set of human resource management (Aslam, et al., 2014).

Total quality people is a form which makes up the right foundation for an organization. This basically consists of people with integrity, character, positive attitude, and good values. People in an organization are considered to be the core asset of that organization but are not always given the right devotion that they deserve or need from their subordinates or line managers. According to Aslam, et al. (2014), the approach of a manager must set out examples for all the subordinates. The superior-subordinate relationships have to be considered as power recipient and power wilder relationships. Every organization is considered to be more successful, if it follows a specified plan for managing people, which consists of a defined number of policies related to people management. From the start of 1990 there has been an increasing interest of organizations in managing people and analyzing its impacts on the employee as well as the organization (Becker and Gerhart, 2010).

The concept of people management involves the role of line managers for shaping the perceptions of employees with the help of enactment of HR policies and their actions of leadership. Currently, senior managers in organizations expect their managers to make the best use of available resources, which leads to faster delivery while using fewer people and lower budgets. Sometimes, this produces tensions at workplace, which has to be managed in most suitable manner. Another important measure which plays a crucial role in an organization is social psychology. Companies are a social place where people work together with each other including their managers, clients and colleagues. For capturing the best out

of their employees, the company managers have to look for innovative ways of leading and motivating their team members. The attitudes of employees toward their colleagues and managers are in relation to social impact. Organizations have to have a productive and motivated workforce for carrying out their roles and responsibilities resourcefully and hence ensure that a level of profitability is maintained by the organization. This is the major reason behind the popularity of people management to be one of the most influential soft leadership skill (Peeters, et al., 2020).

People management is a key sub-set of HR management which covers all the features of how people behave, work, grow and engage in a work environment. There are different systems which are employed for management of people by different organizations (Boakye, 2020). The human resource management systems affect the overall mechanisms of the organization and therefore have to be followed for active puzzle pieces without losing the idea of the bigger picture.

There are several features of people management that have been shown in Figure 1. The figure 1 shows that the tasks and sub-aspects which support the chief pillars of managing people comprise of the recruitment, compensation, rewards, branding of employer, performance management, safety, organization development, wellness, training, engagement and motivation of employees, administration and communication. These features of people management, altogether interlace the cultural fabric within a company and lead to providing experience which helps in attracting and retaining the right talent (Dutta and Chaudhry, 2021).

In line with (Kaufman, 2014), people management and approaches related to analytics have been working together over the past few years. People management concept started to be observed in research from the early 1900s and the first ever publication which highlighted the topic was in a book “How to measure people management”, which was published in 1984 (Thakre, 2021).

Figure 1. People management feastures in an organization.

Considering that data of people has been regulated progressively by the law and number of ethical questions concerning the procedure of people analytics is increasing, this thesis focuses on finding if an employee will leave a particular organization or not. Also, this thesis shall focus on the achievement of core features that may be included in the team of people analytics for attaining legitimacy and compliance considering the external and internal stakeholders. This thesis also takes into account the features that are required for people analytics team to be successful (Thakre, 2021).

According to (Boakye, 2020), people analytics is a slogan with an increased hype, there is a limited amount of academic research. The implementation of people analytics in business landscape in the face of research has been frequently linking people analytics with progressive business outcomes (Marler and Boudreau, 2016). In Finland, there are a few master's theses that have been done on people analytics which focus on qualitative research. This thesis entails analysis of using people analytics in general and using predictive analysis to attain valuable insights. In contrast to other studies, this research has been conducted using

relevant data related to employee turnover in general firms from a verified online platform Kaggle (Sagar, 2020).

According to Jones et al. (2008), people analytics in Finland has not yet developed fully. It has also been reflected through the responses of 70% of respondents who have been using descriptive analytics for answering a simple question related to "what has happened or might happen" (Jones, et al., 2008). A few organizations are using advanced people analytics, which involves the application of machine learning algorithms and statistical methods on datasets, which syndicate the business and people data to support decision making. A common example of such statistical analysis is predictive analysis.

People analytics is defined as a method for collecting and analyzing data of talent for enhancing business outcomes and perilous talent. The leaders in people analytics are responsible for allowing the HR leaders to produce data-driven insights for making informed decision related to talent, enhance the processes of workforce and encourage positive experience among employees (Vermeeren, et al., 2014). The major role of people analytics in an organization is to gather and analyze employee data for gaining the insights into the core motivations of employees. This data can then facilitate in making better business decisions and have a greater impact from every hire, promotion or assignment. In this regard, there are many factors that count along with the data related to employee attribute, relational analytics also helps in the transmission of insights about the relationship among people (Peeters, et al., 2020).

People analytics, also known as HR analytics, talent analytics, or workforce analytics, comprises of the collection of analysis and reporting of the data of people. It also allows the company to understand and measure the influence of a variety of metrics involved in managing people on the performance of business operations and how it affects the decision making criteria of managers on the basis of data. To explain further, people analytics is a data-focused approach toward the management of human resource (Tursunbayeva, et al., 2018).When the companies evaluate the relationship between individuals and groups in which they work, the business outcomes and employee experiences can be improved by

leadership. People analytics gives a macro-view of the influence of employee drive and the conditions under which they can take full advantage of their contributions. The use of people analytics helps in identifying the key trends, patterns and relationships which allow the leaders to make well-informed decisions, identify key problems and act effectively for enhancing their business operations. The value of people analytics is appreciated even outside the silos within people management. People analytics that has been entrenched in methods of leadership and software applications aid in the application and collection of more valued data. (Shrivastava et al., 2018)

1.1 Objectives and limitations

The objective of this thesis is to apply people analytics by using a machine-learning algorithm to predict whether a person will leave an organization or not. This study focuses on exploring the implications of people analytics in reducing employee turnover in an organization. The topic of this research has been selected due to an identified research gap from previous studies. People analytics is a growing concern of many organizations all over the world. Although, there are some studies on relevant topic, these do not take into account the use of people analytics, specifically by organizations in Finland (Jones, et al., 2008). We show the relevant suggestions to organizations to incorporate people analytics in overcoming employee turnover. Research questions for supporting the research topic have been formulated below.

- How attrition rate of employee can be managed in an organization using people analytics?
- What types of people analytics are used by companies to manage employee turnover?

The above questions have been proposed for conducting this research. The first question entails the management of employee attrition rate using people analytics by companies. For attaining the objective of this thesis, it is important to briefly discuss the concerning factor, metrics, predictors as well as different strategies which can affect the employees to leave an

organization. The second research question is focused on finding the types of people analytics that are used by companies to manage attrition rate of employees. This may also involve the strategies and techniques in business analytics which can be incorporated in company's functions and help retain employees.

1.2 Structure of thesis

This thesis comprises of six Chapters. Chapter 1 involves a general introduction about topic and research background. Chapter 2 includes the problem statement and information about data. Chapter 3 involves the description of decision tree models and analysis of synthetic data. Chapter 4 involves the discussion on key findings of research and data analysis. Chapter 5 involves a brief exploration of research thesis and concluding the research topic with the help of data exploration. Lastly, it involves proposing useful recommendations in Chapter 6.

2 Problem statement

This research is being conducted to cover the identified research gap from previous studies in people analytics. This research focuses on analyzing certain employee data to gather results and recommend useful measures for managing the attrition rate. Machine learning algorithms will be applied for model building and data analysis. Based on the results different options will be explored through which the attrition rate of employees can be managed in the organization using people analytics.

2.1 Data

The dataset utilized in this research consists of almost 15000 observations. The dataset in question is secondary and obtained directly from Kaggle which has a huge repository of the community published datasets and models to promote data science activities and competitions (Sagar, 2020). There are nine main attributes in the dataset which will be discussed in detail in the data exploration section.

2.2 Data exploration

Data exploration is the initial step in data analysis which comprises of the use of data visualizations and other statistical techniques by data analysts for describing the characteristics of datasets such as the quantity, size and accuracy for understanding the nature of data. The techniques of data exploration involve both automated and manual data analysis with the use of data exploration software solutions that identify and explore relationships between different variables visually. This involves the structure of dataset, outliers, value distribution for revealing research patterns and points of interests which enables data analysts for gaining greater comprehension of raw data. The identification of data from Python has been shown in Figure 2.

Data exploration plot between average monthly hours spent by employees and the number of employees shows that between the ranges of 800 to 1000 employees, an average of 150 hours per month has been spent while, 600 to 800 employees show a range of 250 hours per month. Whereas less than 200 employees show an average of 100 hours per month. The data related to salary of employees show that more than 7000 employees have low salary, more than 6000 employees have been receiving medium salary and only 1000 employees have been receiving a higher salary. In terms of last evaluation, around 1000 employees have had their last evaluation 5 months ago, around 800 employees have had their last evaluation about a year ago and less than 200 employees have had their last evaluation less than 4 months ago.

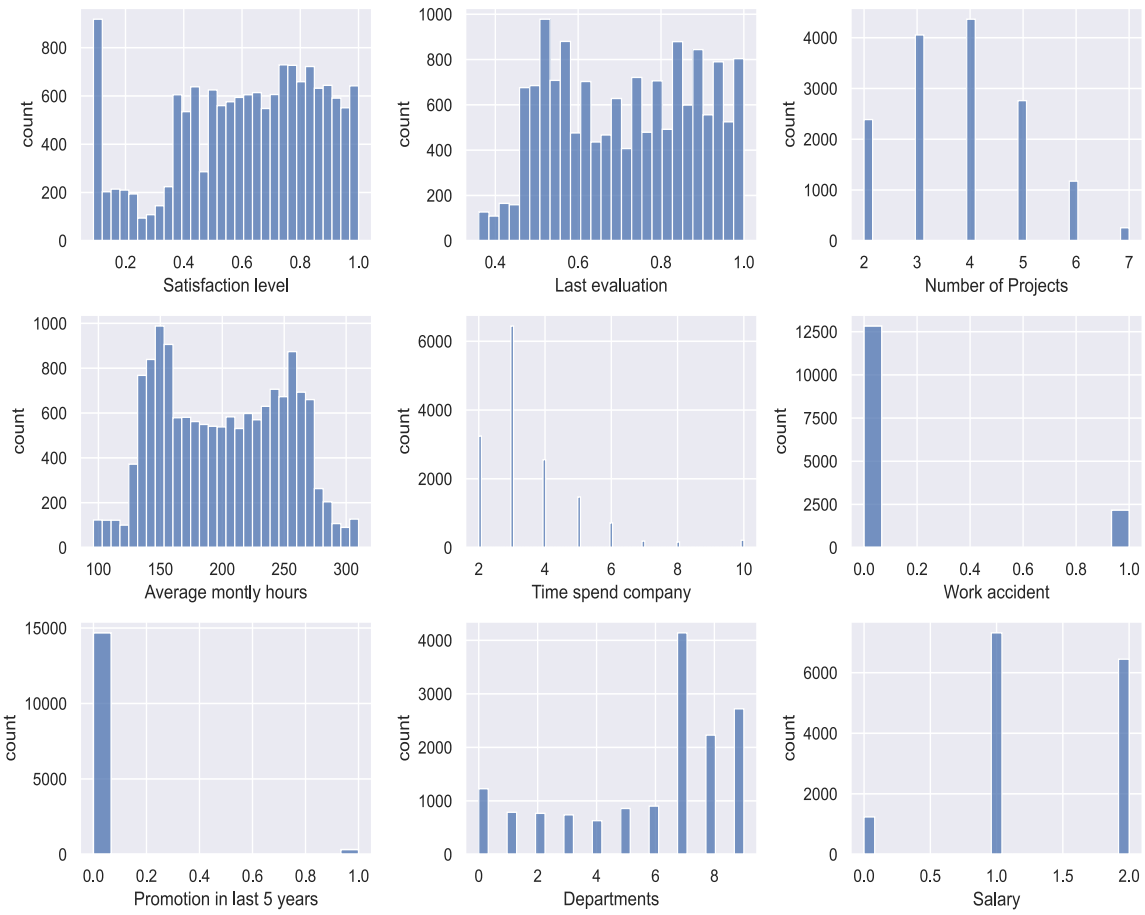

Figure 2. Data exploration – graphs generated from the analysis of original employee data from Kaggle data.

The exploration of data related to department shows that 4000 employees are from sales department, around 500 employees are from accounting and HR department, around 2500 employees are from technical and support department, less than 500 employees are from management, 1000 from IT. Whereas, less than 1000 employees are from product, marketing and R&D. It can be observed from total number of projects taken by employees that around 4000 and more employees have completed 3 to 4 projects, more than 2500 employees have completed 5 projects, more than 2000 employees have completed 2 projects, around 1000 employees have completed 6 projects and less than 500 employees have completed 7 projects.

The data related to promotion of employees shows that more than 14000 of employees have been promoted less than a year ago whereas, less than 500 employees were promoted more than 4 years ago. The satisfaction level of employees shows that more than 800 employees have a satisfaction level less than 0.1, around 600 employees have a satisfaction level of 0.4 and around 700 employees have a satisfaction level more than 0.8. The total time spent by employees at the company shows that more than 6000 employees have spent 3 years at the company, more than 3000 employees have only spent 2 years at the company and less than 500 employees have spent 10 years at the company. In terms of data related to work accidents at the company, it has been depicted that more than 12000 employees have had no accidents at the workplace. Whereas only 2000 employees have been reported to have faced one accident so far.

3 Decision tree model

This chapter includes the decision tree models that have been used in this project. This will discuss the decision tree algorithms used, analysis of decision tree algorithm on synthetic data and sensitivity analysis on synthetic data.

3.1 Use of decision tree

A decision tree algorithm can be used to solve two types of predictive problems. The first one is the categorical classification and the second is regression. These problems are differentiated on the basis of the target variables (i.e. categorical variable, continuous variable) for the subject problem. The algorithm used for this project is the decision tree algorithm (Lamrini , 2020).

The decision tree algorithm basically comes from a family of administered learning algorithms. An algorithm is used to create decision trees, which stratifies data into smaller groups depending on the class label (Kingsford and Salzberg, 2008). Based on a measure of impurity, the decision tree algorithm decides the split at each node. The number of inaccurate classifications at each node is measured by impurity. The impurity may be measured in a variety of ways, including Gini and entropy (Song and Lu, 2015). All characteristics are evaluated at each node to see whether they may be utilised as the splitting variable. Every value inside each feature's range is examined for splits.

Decision tree is a form of a tree visualization in which each internal node postulates a test on an attribute of data in question and each edge between a child and a parent signifies a decision or a consequence on the basis of that particular test. To predict a label class in a decision tree for record, the starting point is the root of tree. The values of root attribute are compared with the attribute record. For comparing them, branch is followed which is

compared to that value and then comes the next node (Lamrini, 2020). After the algorithm is complete, the resulting structure will have many nodes, each with corresponding distributions that can be analyzed to determine the importance of features in predicting the target class. A decision tree algorithm has been used as it helps in creating a training model which can be used for predicting the value and class of defined variables with the help of learning different rules for decision making from historical data. (Kingsford and Salzberg, 2008).

3.1.1 Entropy

Entropy is a method for identifying randomness in the information that is being processed. The higher the value of entropy, the harder it gets in concluding information. The entropy of a random variable, or more particularly its probability distribution, measures how much information it contains. The entropy of a skewed distribution is low, but the entropy of a distribution with equal probability is higher (Mitchell, 1997). For instance, flipping a coin provides random information.

$$E(S) = \sum_{i=1} - p_i \, log_2 \, p_i. \quad (1)$$

Here S is known as current state and p_i is known as probability or a percentage of class (i) in a node of (S) state. Entropy for several variables is represented below;

$$E(T, X) = \sum_{c \in X} P(c)E(c). \quad (2)$$

Here T and X are selected attributes.

3.1.2 Information gain

According to Mitchell (1999), information gain is a statistical property which depicts how well the training is separated from a given attribute according to their target classification. The construction of a decision tree includes the identification of an attribute which returns

highest information gain and smallest entropy. Information gain is known as a reduction in entropy. Calculating the difference between entropy before split and average entropy after splitting datasets on the basis of given values of attributes. The mathematical representation of information gain is as follows.

$$\text{Information Gain } (T, X) = \text{Entropy } (T) - \text{Entropy } (T, X) \quad (3)$$

3.1.3 Mathematical calculations of decision tree model

In these mathematical calculations, we have taken the sample data and shown the calculation and working of our decision tree model. We have shown here how our splitting criteria are decided and how on a single node we decide our splitting criteria. The same algorithm is run in every node and every level and splitting criteria are decided under this algorithm until we get a completely pure split. This means that on one side we get completely positive values and on the other side we get all negative values. The splitting criteria can also be stopped when we have set a maximum depth of a decision tree and if the depth is reached then the splitting criteria are matched and then it will stop otherwise it will keep running. By using this algorithm we have made our decision tree.

Table 1. Sample data.

Sr. No.	Satisfaction level	Time spent at company	Number projects	Last evaluation	Left
1	1	1	1	2	1
2	1	1	1	1	1
3	2	1	1	2	0
4	3	2	1	2	0
5	3	3	2	2	0

6	3	3	2	1	1
7	2	3	2	1	0
8	1	2	1	2	1
9	1	3	2	2	0
10	3	2	2	2	0
11	1	2	2	1	0
12	2	2	1	1	0
13	2	1	2	2	0
14	3	3	1	1	1

3.1.4 Attribute: satisfaction level

The first attribute is satisfaction level having values 1, 2 and 3.

Values (Satisfaction level) = 1, 2, 3

The Entropy (set) of satisfaction level is S = [9+, 5-]

$$\text{Entropy}\ (S) = -\frac{9}{14}\log_2\ \frac{9}{14} - \frac{5}{14}\log_2\ \frac{5}{14} = 0.94$$

S_1 ← [4+, 0-]

$$\text{Entropy}\ (S_{v1}) = -\frac{4}{4}\log_2\ \frac{4}{4} - \frac{0}{4}\log_2\ \frac{0}{4} = 0$$

S_2 ← [3+, 2-]

$$\text{Entropy}\ (S_{v2}) = -\frac{3}{5}\log_2\ \frac{3}{5} - \frac{2}{5}\log_2\ \frac{2}{5} = 0.971$$

S_3 ← [2+, 3-]

$$\text{Entropy } (S_{v3}) = -\frac{9}{14}\log_2 \frac{9}{14} - \frac{5}{14}\log_2 \frac{5}{14} = 0.94$$

$$\text{Gain } (S, \text{ Satisfaction Level }) = \text{Entropy } (S) - \sum_{v \in \{v_1, v_2, v_3\}} \frac{|S_v|}{|S|} \text{Entropy } (S_v)$$

$$= E\ (S) - \frac{5}{14} E\ (S_{\text{v1}}) - \frac{4}{14} E\ (S_{\text{v2}}) - \frac{5}{14} E\ (S_{\text{v3}})$$

$$\text{Gain } (S, \text{ Satisfaction Level }) = 0.94 - \frac{5}{14} 0.971 - \frac{4}{14} 0 - \frac{5}{14} 0.971 = 0.2464$$

From the calculation, information gain has been determined which is 0.2464. Similarly, by calculating the attributes of time spend in a company, number of projects completed and last evaluation we will get the following values of information gain.The values of all these attributes are listed below.

Gain $(S, \text{satisfaction level}) = 0.2464$

Gain $(S, \text{number of projects}) = 0.1516$

Gain $(S, \text{time spend } in \text{ company}) = 0.0289$

Gain $(S, \text{last evaluation}) = 0.0478$

Information gain for last evaluation is 0.0478. This shows the quality of split which is lowest in this case. We select the attribute for splitting whose information gain is more compared to other attributes. These values of information gain show that the highest value is of satisfaction level, showing a better quality of split compared to other attributes. The second highest value is of number of projects completed. Then comes time spent at company and lastly, the last evaluation.

3.2 Decision tree algorithm

A machine learning algorithm has been chosen for predicting the model. For this purpose, a decision tree classifier has been used and entropy has been used to draw results. Test and train datasets have been considered to split the attributes of decision tree. The purpose of train datasets is to train the model whereas, test datasets have been used to illustrate the accuracy of the model. For this model, test size has been set at 20 percent. The decision tree has been trained on training set and was then scored on test set. Here we make 2 decision tree models one with Synthetic data and one with adding Gaussian noise in the synthetic data to check the sensitivity analysis of our model.

To better explain the model and compare the data, we have generated synthetic data. A simple decision tree algorithm has been generated from synthetic data. The decision tree algorithm of synthetic data includes three columns. Two columns are of features and one is of the target column. The decision tree algorithm of synthetic data has been illustrated in Figure 3. The accuracy of the decision tree model is 52%. The decision has been made on the basis of entropy.

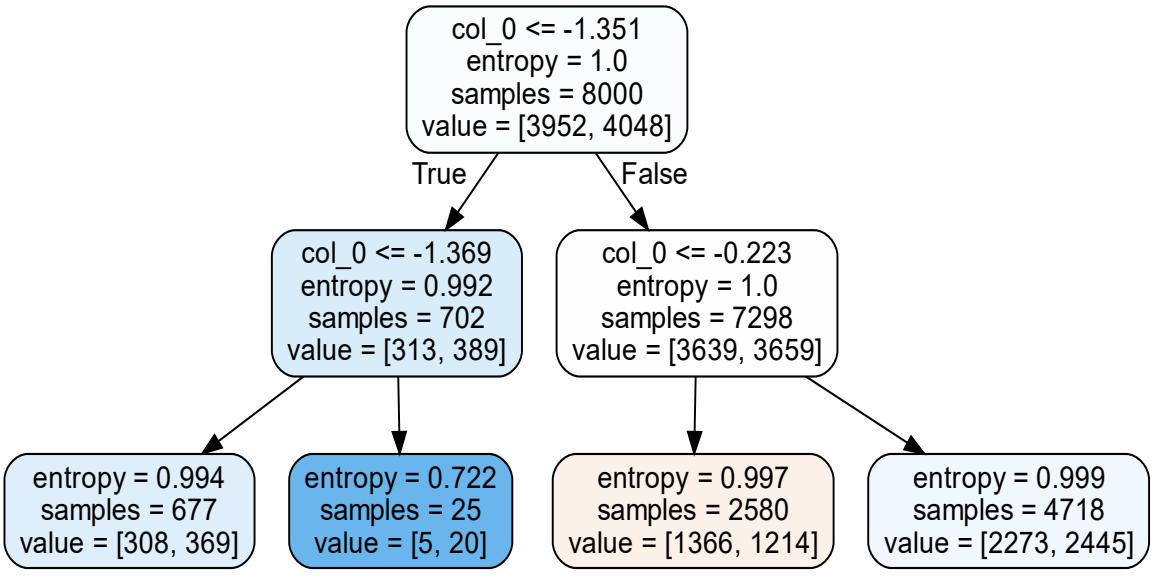

Figure 3. Synthetic data decision tree model.

The confusion matrix of synthetic data has been displayed in Figure 4. In many situations, models are built using a descriptor or fingerprint representation of the components, and the model's ability to distinguish between the two classes is assessed using the results of a prediction on a separate dataset. Because the predicted instances were pre-labeled with their known class and also contain a label coming from prediction, this produces four sorts of outcomes for two-class issues. True positives (TP), false positives (FP) (also known as false discoveries or type I errors), true negatives (TN), and false negatives (FN) (also known as missed discoveries or type II mistakes) are the four sorts of results. The confusion matrix is a term used to describe the collection of all four outcome categories (Brown, 2018). Figure 4 shows the findings generated from synthetic data before adding noise. Correct predictions of synthetic data show the following; (1) true negatives of the model are 16.95% and (2) true positives are 35.4%. Incorrect predictions include the following; (1) the value of false positives is 32.45% and (2) false negatives has been observed to be 15.20% which have been falsely predicted in this synthetic data model.

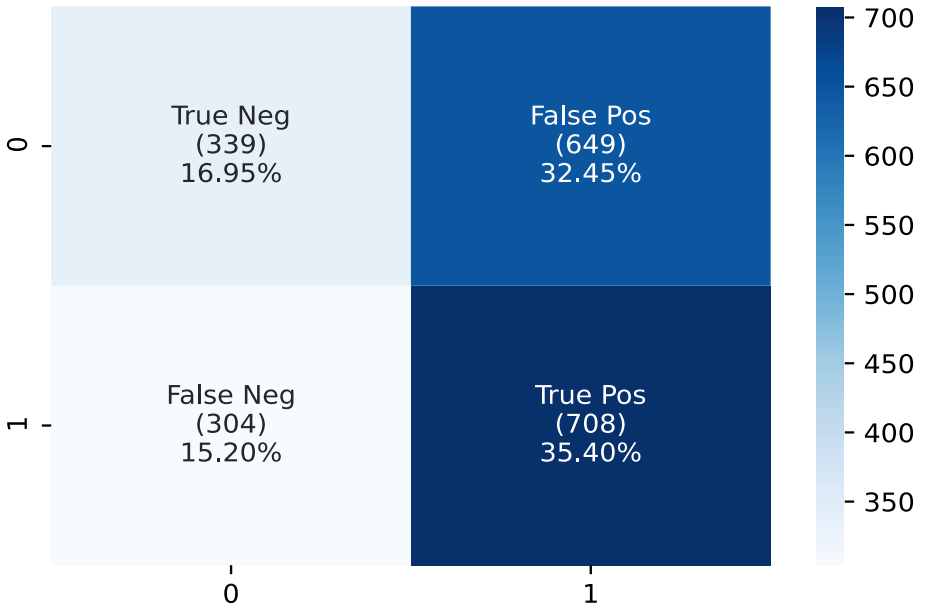

Figure 4. Results of confusion matrix.

3.3 Sensitivity analysis on synthetic model

Sensitivity analysis is always an important part of decision making, and it frequently focuses on probabilities in decision trees when the user has little knowledge about the real values of probabilities in the stochastic model under consideration (Kamiński, 2018). Sensitivity analysis of synthetic data has been performed to evaluate the performance of the model by adding Gaussian noise to the synthetic data.

3.3.1 Decision tree algorithm after Gaussian noise in the synthetic model

The decision tree algorithm after adding Gaussian noise in synthetic data has been illustrated in Figure 5. Figure 5 shows the decision tree algorithm after adding noise, showing how the model has slightly changed from previous results. The accuracy of model has slightly reduced from 52% to 50% after adding noise. This shows that the previous model (i.e. before adding noise) was better than this model.

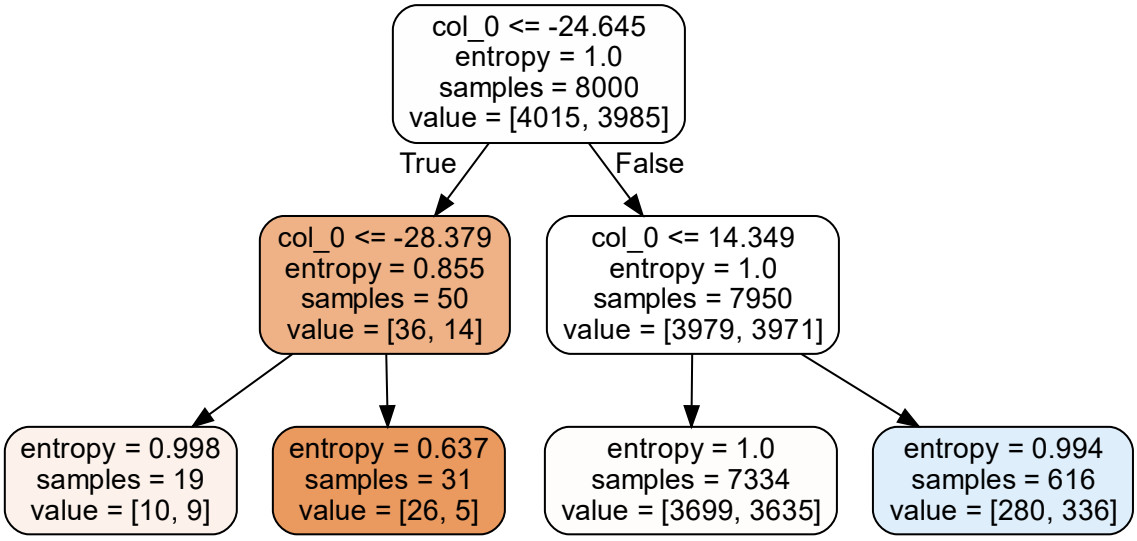

Figure 5. Decision tree of perturbed synthetic data.

3.3.2 Confusion matrix after adding Gaussian noise in the synthetic model

The confusion matrix after adding Gaussian noise in synthetic data shows the following results as demonstrated in Figure 6.

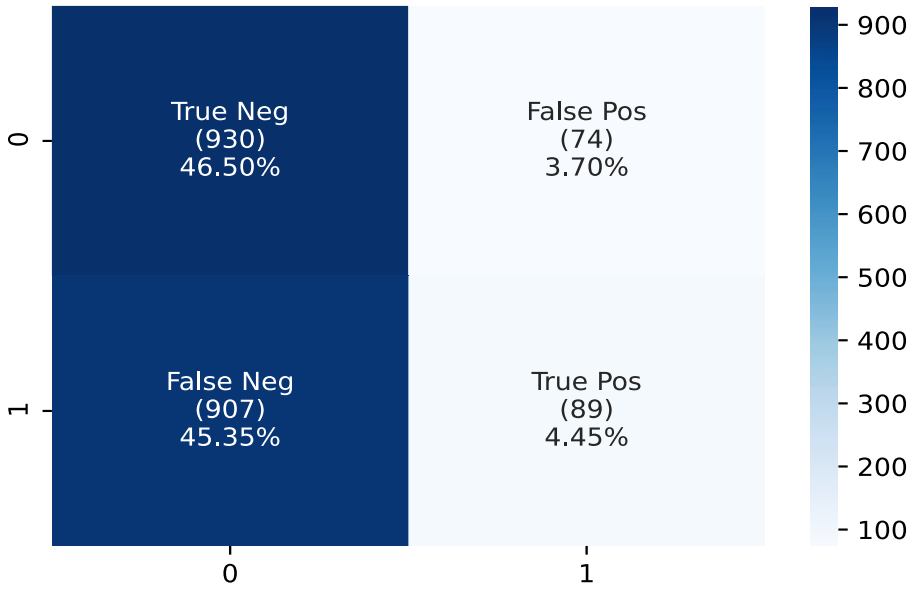

Figure 6. Confusion matrix of perturbed synthetic data.

Correct Predictions have slightly changed due to Gaussian noise. (1) true negatives of the model after adding Gaussian noise are 46.5% and (2) true positives are about 4.45%. Changes in incorrect predictions include the following; (1) the value of false positives is 3.70% which have been falsely predicted and (2) the value of false negatives has been observed to be 45.35%.

3.4 ROC curve of synthetic model and noisy synthetic model

Receiver operating characteristic (ROC) curve is a visual representation depicting the performance of an ordering model at all ordering thresholds. This curve intrigues two factors: true positive rate and false positive rate. For a set of thresholds, the ROC curve is a plot of the true positive rate (TPR, sensitivity) vs the false positive rate (FPR, 1 - specificity).

The threshold for predicting the default categorization is 0.50 by default, although any threshold can be used (Fawcett, 2006). The ROC curve of synthetic and noisy synthetic data has been displayed in Figure 7. Figure 7 shows the ROC curve, which shows a comparison of synthetic and noisy synthetic data between true positive rate (on y-axis) and false positive rate (on x-axis). The accuracy of ROC curve of synthetic data is 45%, while that of noisy synthetic data is 43% which is slightly different from the synthetic one and because of adding slightly noise the curves have slight differences as well.

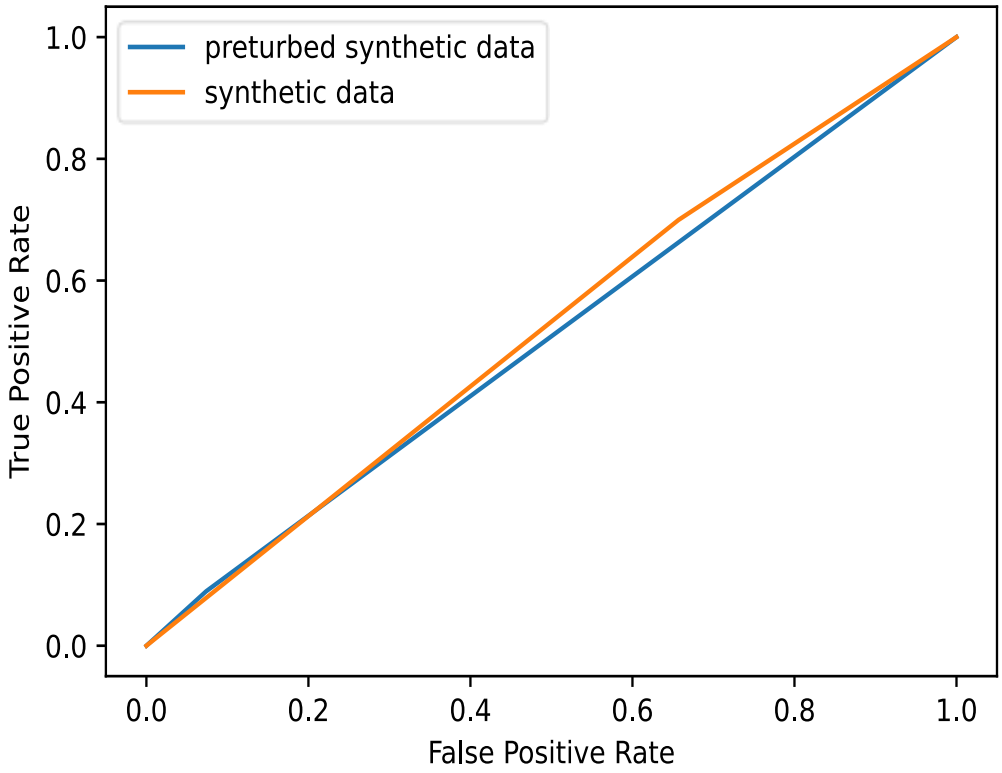

Figure 7. ROC curves w.r.t decision tree and decision tree with Gaussian noise.

The actual accuracy of our model is 52% without noise and 50% after noise. In conclusion, the sensitivity analysis of synthetic data and noisy synthetic data shows that the model without Gaussian noise is better than noisy one.

3.5 Input and output plots of synthetic model with and without noise

To demonstrate the variations in data in terms of density, probability distribution plots have been generated. The actual target variable shows the input values of data without noise and the predicted target variable shows the output values of data with noise. The probability distribution of synthetic data has been illustrated in Figure 8.

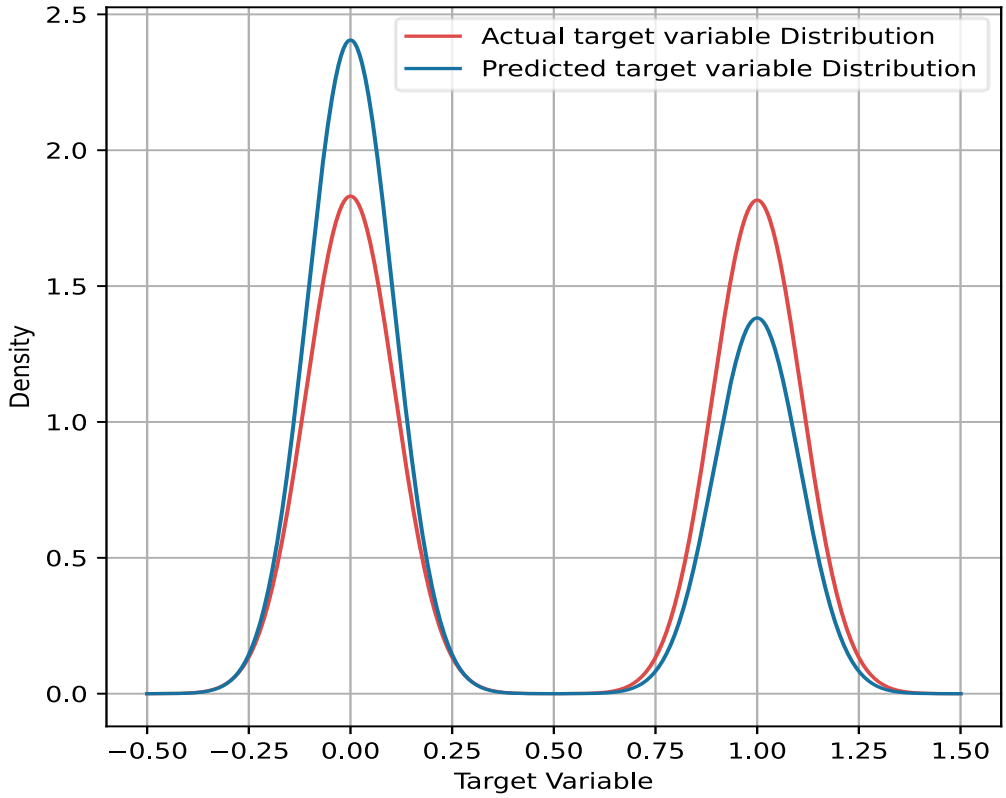

Figure 8. Probability distribution of input and output plots with little variations.

The probability distribution of synthetic data in Figure 8 shows that by changing the values of input we get the values in output that are less accurate. This shows that our model without noise is more accurate and the decision tree without noise is more reliable. The probability distribution of synthetic data by adding a little bit more input values to again verify the working of our decision tree has been illustrated in Figure 9. The probability graphs of synthetic data show that even if we add more input values again the results are the same results that our actual data without noise is more accurate and reliable than with noise.

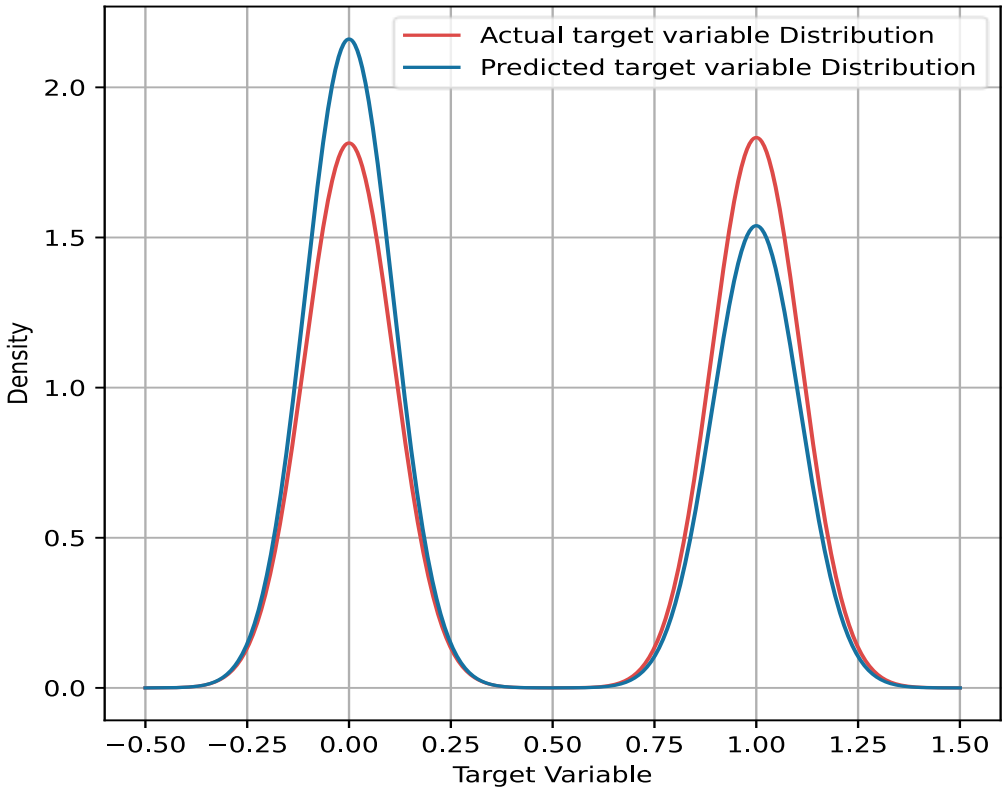

Figure 9. Probability distribution of input and output plots with more variations.

4 Case study on Kaggle data

The data was gathered from an online website known as Kaggle (Sagar, 2020). As it is restricted to the use of the actual data of a specific company, random people analytics data of a random company was selected from Kaggle. Upon gathering the data, it was then analyzed to understand if a person will actually leave the organization or not. For the prediction and analysis, it is ensured that this data does not belong to any specific organization.

- The coding part is done in Python.
- For the prediction, the machine learning classification model which is a decision tree is used for classifying whether a person will leave the organization or not.
- To visualize the model through graphs, Python.

4.1 Results

This sub-section involves the results gathered from predictive analysis. Specifically, this section includes the results that have been generated using different tests such as density plots for selected variables, decision trees, box plots, and distribution of features. Density plots have been designed for displaying the data where it has been concentrated over a period of time.

4.1.1 Satisfaction level of employees

First of all, the data related to satisfaction level of employees has been displayed in the Figure 10. The red line shows the employees who have left whereas, blue line shows those employees who have retained. The density of satisfaction level of employees is shown in the

graph below which shows that employees with a satisfaction level of 0.25 to 0.5 have left the organization and employee having a satisfaction level from 0.50 to 0.10 have retained.

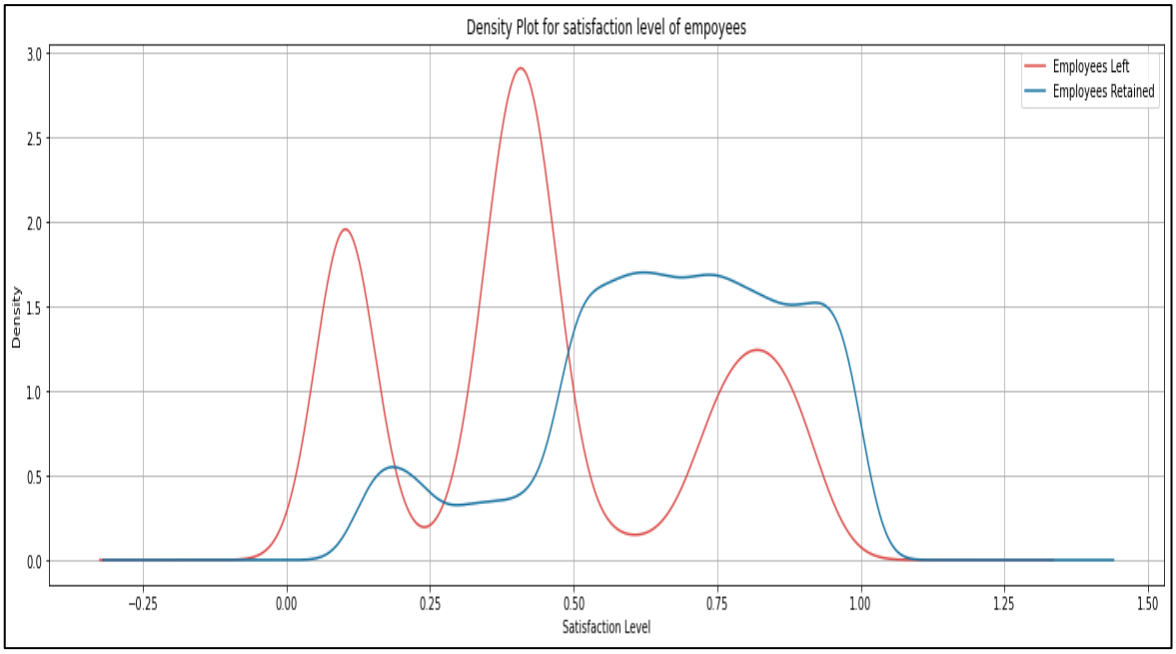

Figure 10. Employees satisfaction level.

4.1.2 Number of projects completed

The data related to the number of projects that have been completed by employees has been visualized in figure 11. The red line shows the employees who have left due to completion of a less number of projects showing that employees who have done 2, 4, 5, and 6 who have left the organization. Whereas, blue line shows those employees who have retained. The density of the number of projects completed by employees is shown in the Figure 11 below which shows that employees who have completed 3, 5, 6, have retained.

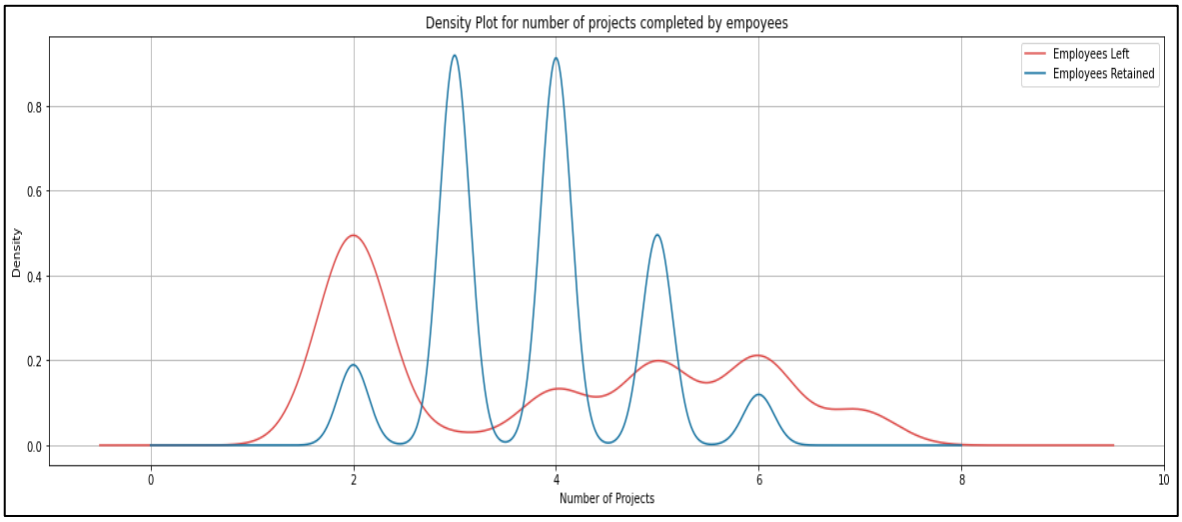

Figure 11. Employees number of projects.

This shows that number of projects completed by an employee has a significant impact on whether an employee will leave the organization or not.

4.1.3 Average monthly hours spent

The data related to the average monthly hours spent by employees on job has been shown in density plot below.

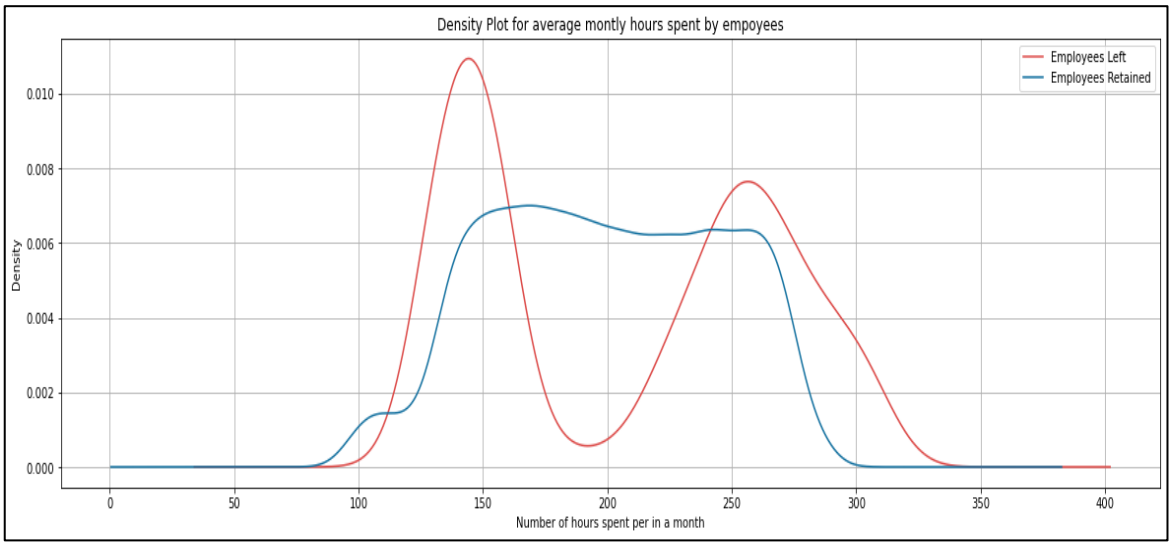

Figure 12. Employees hours spent at work.

The impact of average hours spent by employee on employee retention and quitting can be explained by peaks in Figure 12.The above peaks in graphs show that employees who have spent 100 to 150 hours at their job and those who have spent more than 250 hours tend to leave the organization. Whereas, the employees who constantly spend 150 to 250 hours per month, stay in the organization.

4.1.4 Years spent at company

The number of years spent by employees have a significant impact on employee turnover as explained in Figure 13. Density plot shows that employees who have spent 3, 4, and 5 or even 6 years in the organization leave the company. The highest peak in above graph shows that maximum employees leave the organization after 3 years. Hence, it can be interpreted that employees gain experience for first 3 or 4 years of work and then leave the company. Whereas, the employees who have spent 2, 3, or 4 years, do not leave the company. This also shows that many employee do not leave the organization even after 10 years of work in the company.

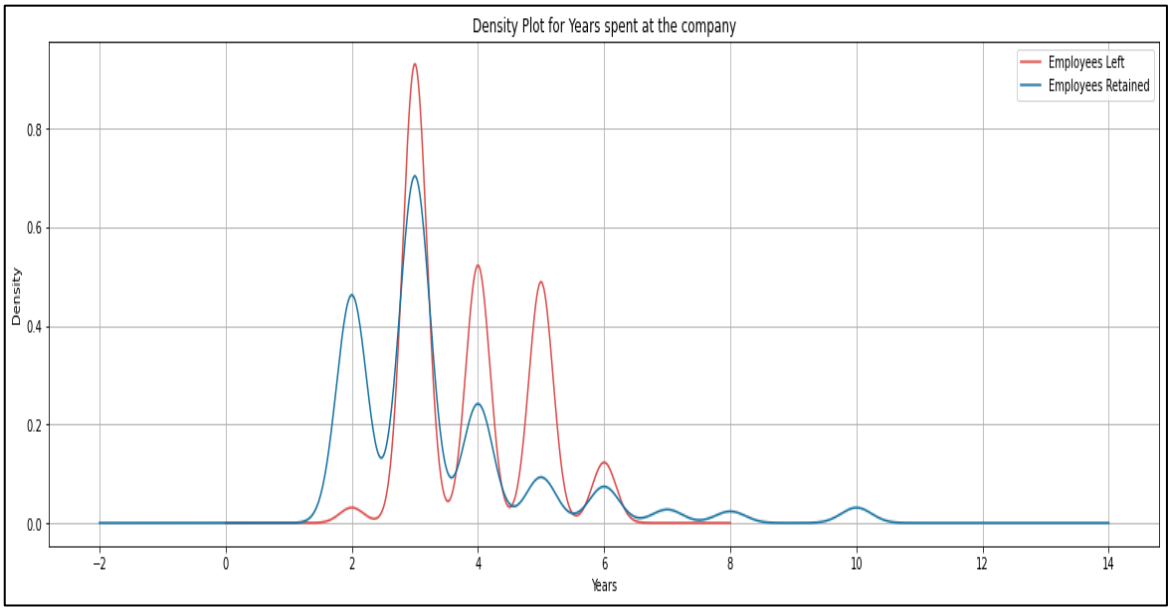

Figure 13. Years spent at the company.

4.2 Box plot analysis

Boxplots are a popular graphical tool for exploring data and better understanding the information we're dealing with. Boxplots show the first, second, and third quartiles of a data collection, as well as the interquartile range and outliers. The boxplot's information, as well as the majority of its variants, is based on the data's median. However, the mean is used in a lot of scientific applications to analyse and report data (Marmolejo-Ramos and Tian, 2010). A box plot shows how a certain variation is caused in a feature when an employee leaves an organization or stays in it. The green line in graph shows the mean value. There are two values within an inter-quantile range that is; upper quantile and the lower quantile. The other upper (upper end of whisker) and lower points (lower end of whisker) show maximum and minimum values. Here 0 means the number of employee who are still working in the company left and 1 means the number of employees who left the company. Boxplot of our data has been illustrated in Figure 14.

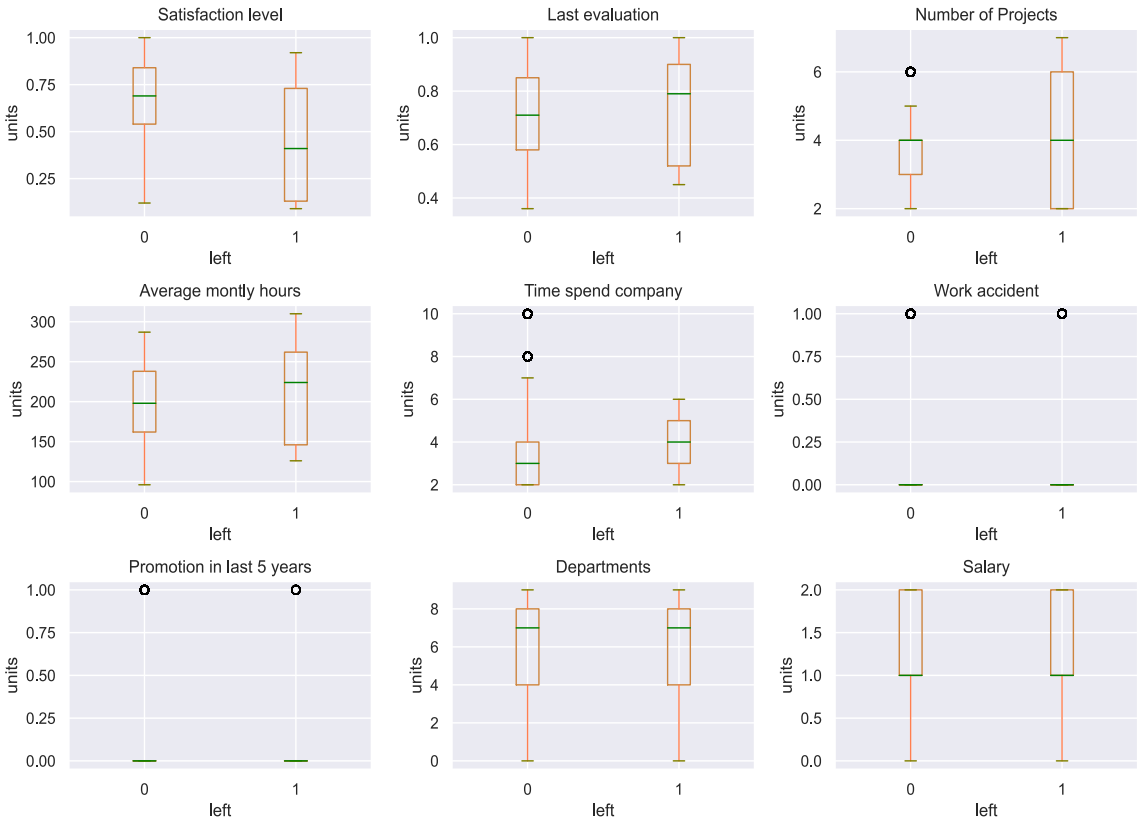

Figure 14. Box plots generated from original Kaggle data.

Box-plot of Satisfaction level of shows that an average employees having 40 percent satisfaction level for their job have left the organization and employees having 60 percent of satisfaction level with the organization and are still working for the organization. Similarly, for last evaluation Figure 14 shows that employees who scored 80% or above in their last evaluation are likely to leave the organization, whereas, the ones who scored around 70 percent do not leave in regards to their last evaluation at job. This also reveals a trend that could help identify the issues that compel the high performing employees to leave the company. Whereas, box-plot for number of project completed by employees' shows that employees who have completed less than 4 years are likely to stay at the company, however, employees whose projects are of 4+ years are likely to leave the company.

The average monthly hours of employees can be seen in Figure 14. It can be seen that on average employees who spend 230 hours or above per month in their job have left the organization and employees who spend 200 hours per month in their job are still working for the organization. Coming to the plots of employees who have spent time at company, so the employees who have spent 3 years are likely to stay at the company, however, employees with a service of 4 and above years are likely to leave the company. Whereas, box-plot of work accident shows that it is rare that work place accident happens in a company and even if it happens so it does not effect the employee behavior to stay or leave the organization.

Similarly, box-plots of promotion in last 5 years, department and sales shows that it does not effect the employee behavior to stay or to leave the organization, so these features are not important.

4.3 Decision tree algorithm of Kaggle data

In the current study, supervised classification has been used with the help of decision tree. The accuracy of decision tree that has been calculated using actual data of this project was found to be 77% which shows that the model used for this project is accurate. For this study, decisions have been taken on the basis on entropy. The working of this decision tree model

is the same that have been explained above in the algorithm section. The decision tree of Kaggle data has been demonstrated in Figure 15.

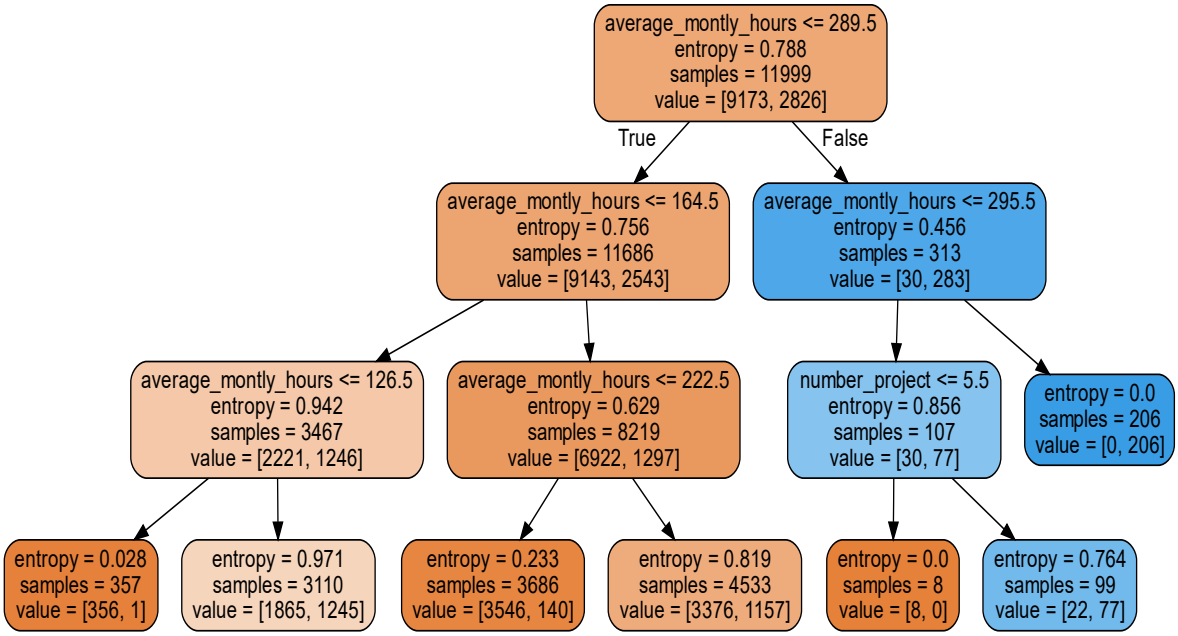

Figure 15. Kaggle data decision tree model.

The formation of decision tree shows the root, nodes and sub-nodes. To further identify the attributes in data, the following criteria for selection has been used for this research.

- Entropy
- Information Gain

Average monthly hours: The first question that the decision tree asks is if the average monthly hours less than 289.5 is. It either follows true or false based on the results.

Entropy = 0.788: Entropy is a measure of purity, and it is measured between 0 and 1. If the entropy value is less the impurity will be less and if impurity value is greater than 1 then it is difficult to decide whether a person will leave the organization or not.

Samples: This dataset contains 11999 samples so this value is set to 11999.

Value = [9173, 2826]: The value list tells us the distribution of target classes that how many people have left the company and how many people have stayed. The first element of the list shows the number of samples that belong to the people who have not left the company. While the second element of the list shows the number of samples that belong to the people who have left the company. The impurity and purity of the node are decided on the basis of entropy value.All other nodes in this decision tree model are working on the same principle and following the same steps and procedure.

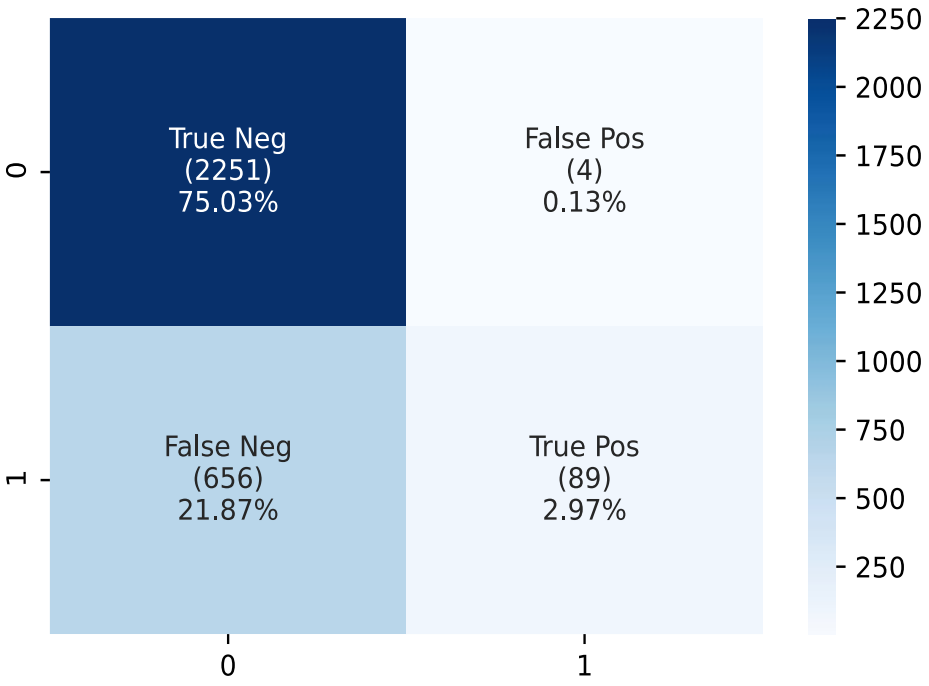

Figure 16. Confusion matrix generated for Kaggle data through python.

The confusion matrix has been generated for the original data through Python. The Figure 16 shows the findings generated from original data before adding noise, which depicts what has been predicted as well as the outcome of entire project. The confusion matrix shows the overall performance of classification algorithm used in this thesis.

4.3.1 Results of confusion matrix before adding noise

Correct predictions – basically it involves the correct decisions of classifier that is, when negative class is correctly predicted as negative and positive class is correctly predicted as positive.

- **True negatives –** where model predicted that the person will not leave, and the person did not leave. Model of this research has truly predicted 75.03% employee who are not leaving the organization.
- **True positives –** where model predicted that the person would leave, and the person did leave. It has been truly predicted that 2.97% employees who will leave the organization.

Incorrect predictions –it involves the incorrect decision of classifier that is, when positive class is falsely predicted as positive and negative class is falsely predicted as negative.

- **False positives –** where model predicted that the person would leave, and the person did not leave. It has been observed that 0.13 % employees were not leaving the organization which have been falsely predicted that they will leave the organization.
- **False negatives –** where model predicted that the person will not leave, and the person did leave. It has been observed that 21.87 % employees were leaving the organization which have been falsely predicted.

4.4 Sensitivity analysis on Kaggle model

Sensitivity analysis has been performed for this model. For this purpose, Gaussian noise with small variance was added to the original data to check the variations in actual data and find out the change in accuracy and performance of our model due to this Gaussian noise.

4.4.1 Decision tree algorithm after adding Gaussian noise

The decision tree algorithm generated after adding Gaussian noise in the data has been depicted in Figure 17.

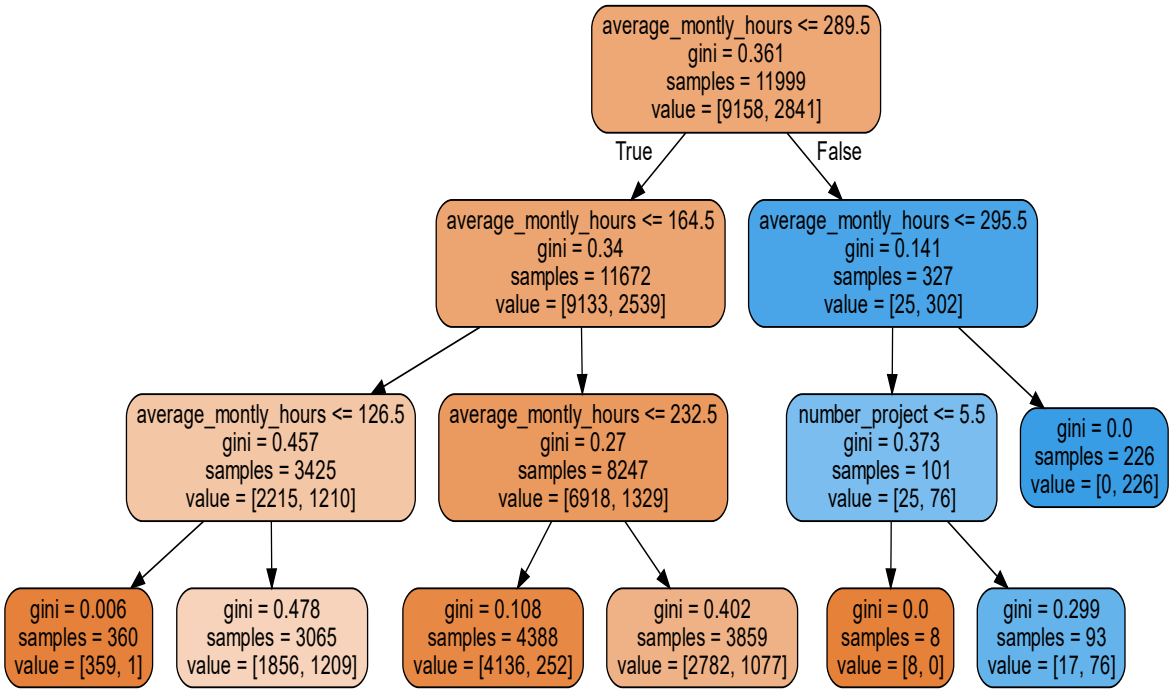

Figure 17. Decision tree of perturbed Kaggle data.

It can be observed from Figure 17 that the accuracy of decision tree algorithm after adding noise has been changed to 74% which is 3% less than the accuracy of original data. Similarly other features like average monthly hours and number of projects values also slightly changed when we added the noise in the data. The working of decision tree has also been slightly changed by adding some noise. This shows that the model we used without noise in this thesis is more accurate.

The confusion matrix has been generated after adding Gaussian noise in the data to check the sensitivity analysis has been depicted in Figure 18. The confusion matrix in figure 18 demonstrates the sensitivity of this model and how the values have slightly changed after the

addition of Gaussian noise. The value of true negative is 75.37% and that of true positive is 2.33% which are correctly predicted. Whereas the value of false positive is 0.30% and that of false negative is 22%. Which shows that confusion matrix values with noise data is less accurate than without noise data and our model with original data is better than synthetic one.

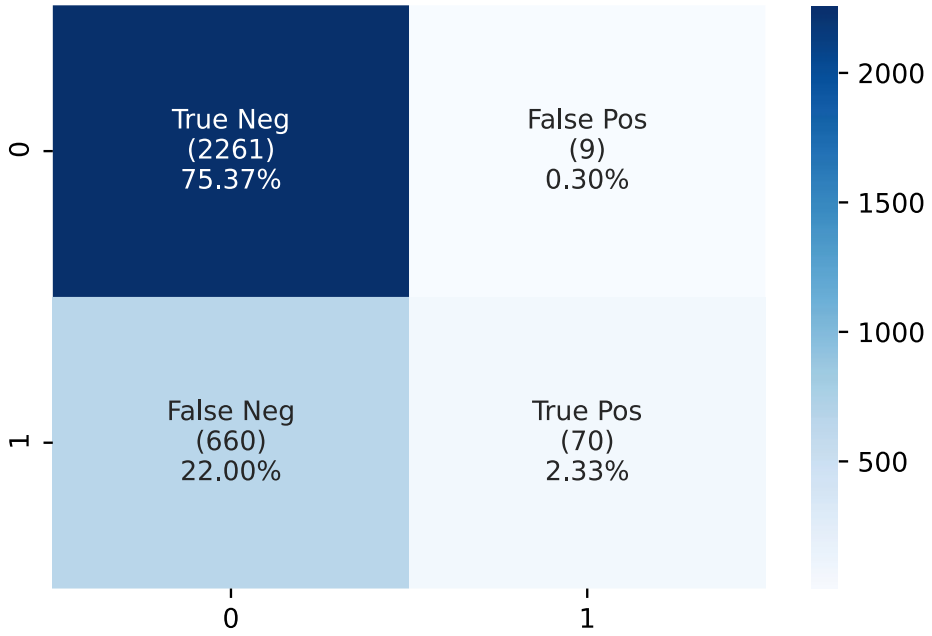

Figure 18**.** Confusion matrix of perturbed Kaggle data.

4.4.2 ROC curve of Kaggle model and noisy model

We are computing ROC curves to analyze the sensitivity effect of Gaussian noise on the predictor It shows the ROC curve, depicting a comparison between true positive rate (on y-axis) and false positive rate (on x-axis) of original and noisy data. The curve of original data has been represented by orange line, whereas the curve of noisy data has been denoted by blue line. The curve which is closer to true positive rate (on y-axis) have better accuracy and results then the curve which is closer to false positive rate (on x-axis). matrix data which has been shown in Figure 19. The accuracy of ROC curve from original data is 55%, while that

of noisy data is 53.9% which is slightly different from original one and because of adding slightly noise the curves have very slightly differences.

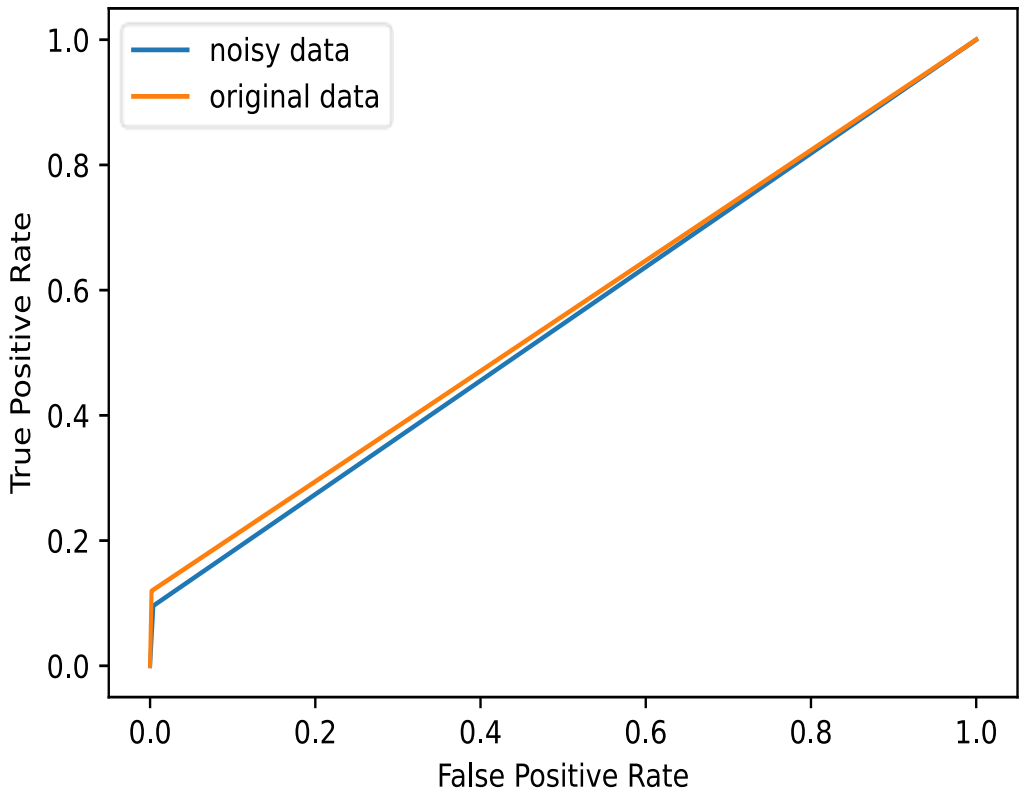

Figure 19. ROC curves w.r.t decision tree and decision tree with Gaussian noise.

In conclusion, the accuracy of our model from original data which is 77% is better than that of the accuracy of noisy data which is 74%. The sensitivity analysis of actual and noisy data shows that the model without Gaussian noise is better than noisy one.

4.4.3 Input and output plots of Kaggle model with and without noise

To check variation in data, the probability distribution of input and output plots of Kaggle data has been generated. The probability distribution of Kaggle data has been shown in Figure 20.

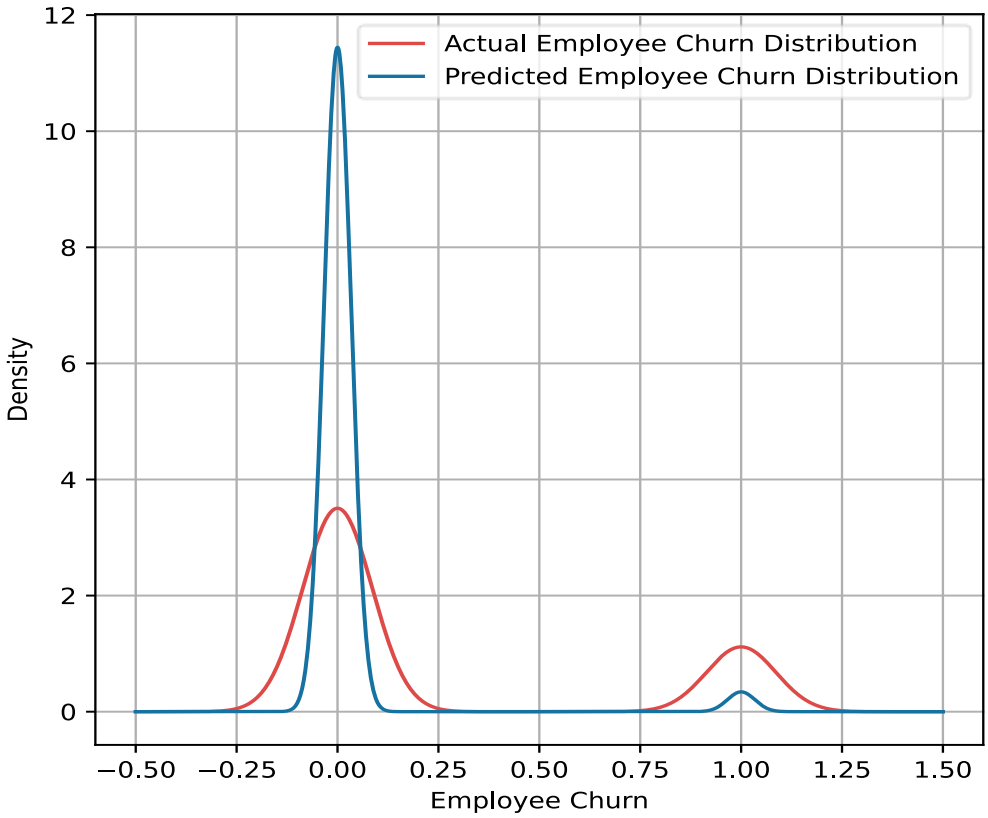

Figure 20. Probability distribution of input and output plots with little variations.

The actual target variable shows the input values of Kaggle data without noise and the predicted target variable shows the output values of data with noise. The probability distribution of Kaggle data in Figure 20 shows that by changing the values of input we get the values in output that are less accurate. This shows that our model without noise is more accurate and the decision tree without noise is more reliable. The probability distribution of synthetic data by adding a little bit more input values to again verify the working of our decision tree has been illustrated in Figure 21.

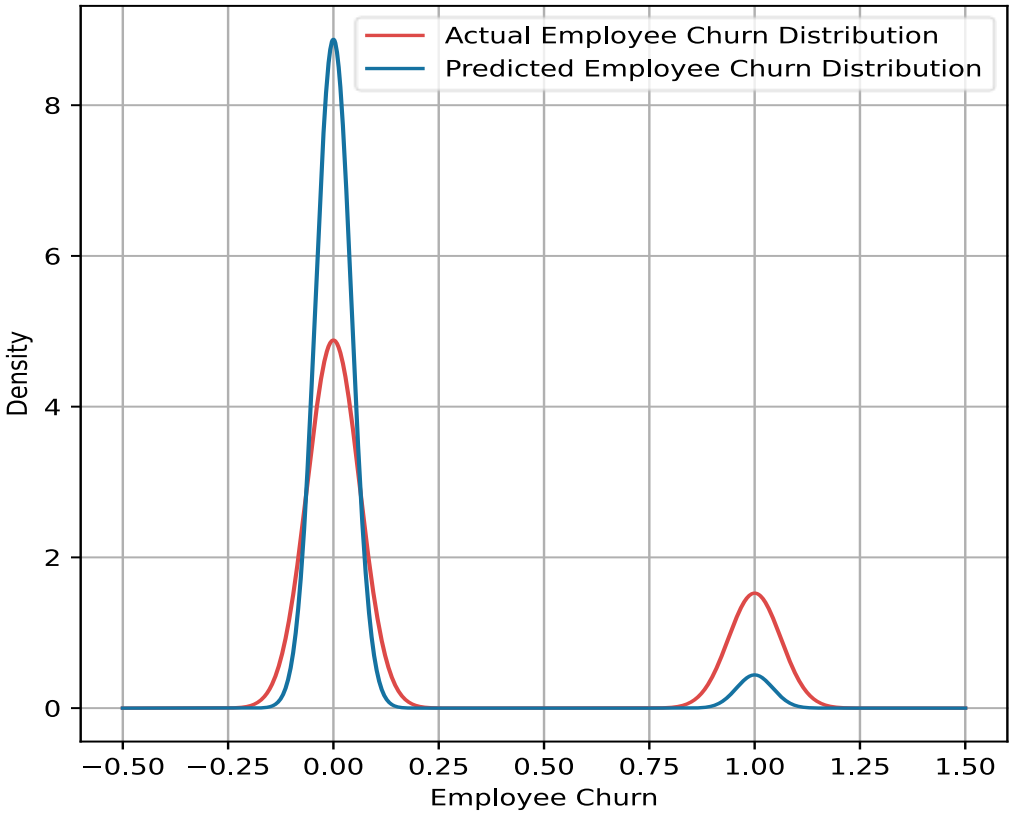

Figure 21. Probability distribution of input and output plots with more variations.

The figure shows how variations have occurred by adding a little bit more noise to the data. It can be observed from probability distribution plots that the accuracy of the model after adding noise is less than the data without noise and the decision tree model is also less reliable, which means that our model without noise is model is more accurate.

4.4.4 Prediction of employee distribution from actual distribution

Then the employee distribution has been predicted from actual distribution. The purpose of this distribution is to comprehend how many employees have actually left the organization in actual distribution and how many people have been predicted who will leave the organization. Figure 22 shows the predicted and actual distribution graphs.

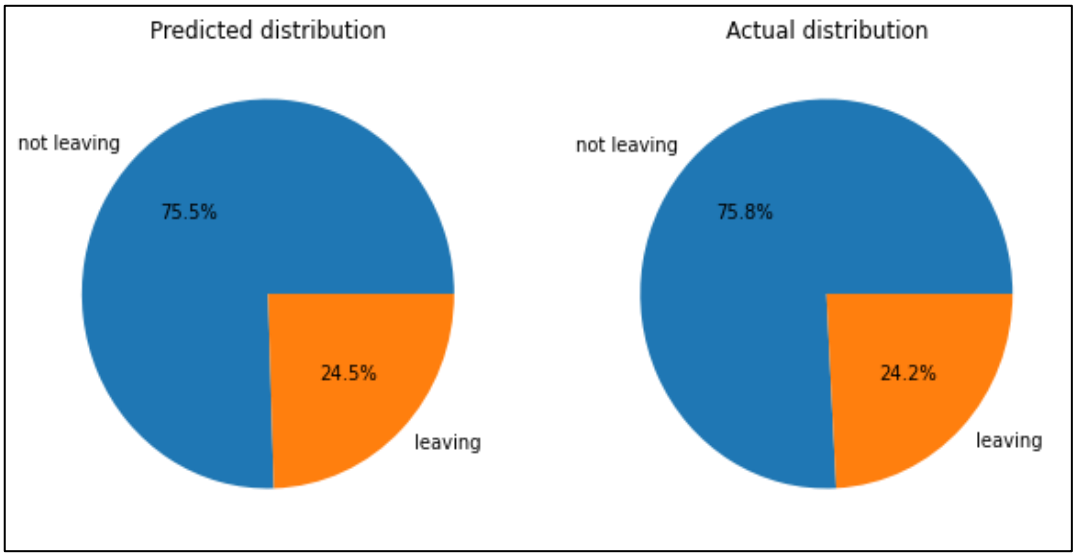

Figure 22. Pie chart - predicted distribution.

The actual distribution shown in Figure 22 depicts that 75.8 employees have not left the organization and 24.2 percent have left. Whereas, in predicted distribution it can be observed that 75.5 percent employees are predicted to leave an organization whereas, 24.5 percent have been predicted not to leave the organization. It shows that our model is acceptable and more accurate.

5 Discussion

In the past, organizations found it difficult to envisage whether a person will leave an organization or not but with the passage of time and advancements in modern world, people analytics has turned out to be an important way for determining and predicting the attrition rate of employees all over the world. With the results data analysis, it has become easier for companies to predict the intention of employees to quit an organization on the basis of data related to job and variables, their behavior and progress. The results of data analysis show that there is a significant impact of employee satisfaction, number of projects completed, hours spent at work monthly, time spent at company per day, promotion of employees in past 5 years, type of department and work accident. It can be depicted from the results that employee satisfaction is directly related to worker's satisfaction. The density plot was created for finding the impact of satisfaction level of employees on employee turnover. It has been found from data analysis that employees who have satisfaction level of 50 to 100 do not leave the organization which those who have a lower level of satisfaction leave the organization.

The second variable that has been selected for analysis through density plot is the number of projects that have been done by an employee. The results of density plot of employee turnover and number of projects completed show that those employees who have completed a desired number of project do not leave the organization whereas, those employees who have completed less projects are more likely to leave the organization. This demonstrates that the number of completed projects have a significant impact on the employee turnover.

The third variable which has been used is the average hours that are spent by the employee in a company per month. The results of hours spent by employees in an organization have shown that the more hours an employee spends in an organization on average, the more chances he/she has of leaving the organization. The results have shown that the employee who leaves the selected organization has spent more than 250 hours in an organization per month. On the other hand, the employees who tend to spend fewer hours in the organization,

remain in the organization. The employees who spend 150 to 250 hours in an organization per month, do not leave the organization intentionally.

Next, the impact of number of years spent by employees in an organization shows that the employees who spend 5 years at the organization, are more likely to leave the organization which may be taken in a way that these employees may work for the organization to gain relevant experience in the field and when they have gained the desired knowledge, they eventually leave the organization. In contrast, the employees who work more than 5 years at an organization are more likely to stay in the organization for a longer period of time. These employees are more loyal to the company.

After that, analysis has been performed to understand the distribution of employee on the basis of salary bracket, nature of department in which the employee works, promotion, and work accident. The results analysis of these features show that salary is a major impact on the satisfaction level of an employee. The employees who leave the organization have medium to low salary range whereas, those who have a higher salary, do not leave the organization. This shows that salary is a good predictor of turnover intention of an employee. In addition to the impact of salary on employee turnover, department also plays an important role in determining the turnover intention of employees. It can be explored from the results of data analysis that employees who work in the sales department have the highest turnover rate, second comes technical department. The other departments in this sequence include support department, IT, accounting, human resource, marketing and product management, research and development and management department.

The results show that the employees from management and R&D department would have a higher satisfaction and hence, they will not leave the organization. Promotion also plays a greater role in determining the impact of employee turnover. The employees who are promoted more often, tend to stay in the organization except for leaving. Lastly, workplace accidents do not seem to be a major cause of employee turnover. This shows that the accidents in workplace do not lead an employee to leave the organization. This can be put in

a way that workplace accidents are natural hazards which could be caused to anyone, anywhere in the world.

Companies from around the world have realized the importance of people analytics in people management. It has been identified that most of the organizations in today's world make use of people analytics and other human resource management software which shows that the application of human resource information technology has been increasing intensively. Similarly, the buy-in of managers in today's business environment is also increasing intensively which has been identified as the companies are more data driven in modern world. In contrast, it takes a lot of resources and additional costs for applying people analytics in the business operations of companies around the world.

The current study has also proposed different ways for finding whether an employee will leave an organization or not using machine learning algorithms such as decision tree, entropy etc. In the current study, supervised classification has been used with the help of decision tree. The accuracy of decision tree for the model used in this project was found to be 97% which shows that the model used for this project is accurate. For this study, decisions have been taken on the basis on entropy. Train and test data sets have been used for splitting the decision tree attributes. Train data set was used for training the model. Test dataset has been used to identify if the predicted model is accurate or not. The test size was set at 20%.

On the basis of this research, it can be discussed that people management is important for every company. The ultimate goal of HR professionals is to make use of latent skills and competencies of employees effectively. The crucial part of human resource department is to measure the key attributes and performance of its employees. The ultimate tasks in an organization can be unraveled with the use of new technology such as people analytics which helps in the transformation of raw data into valuable insights and measurable outcomes. With the use of people analytics in an organization, people planning techniques can be formulated in a more strategic way. This allows the organizations to compete with their competitors as the human resource department is considered as a vital department in any organization.

6 Conclusions and future work

The Data analytics has become an important aspect for addressing the unique employee features which are critical for sustaining and developing a competitive advantage. Companies have been using data analytics in many different ways to predict the propensity of customers for buying a certain product, employee turnover for managing employees effectively and understanding the overall business environment. By using analytics in company's processes, it becomes easier for companies to develop a diverse range of products, maintain a healthy culture in the organization, improve, make quality decisions and retain its key employees with knowledge and skills.

People management function is creating steps for combining experience, intuition and beliefs with the increasing trend of data analytics. As defined by Boudreau and Marler, people analytics is a human resource practice which makes use of visual, descriptive and statistical analysis of data in relation to the key human resource processes, organizational performance, human capital, and external economic benchmarks for establishing business impacts and permit data focused decision making. People analytics can therefore be used for solving the pressing issues in business environment. In conclusion, people analytics helps in solving the critical issues that might lead to employee turnover.

The objectives of this research have been successfully achieved as the primary goal of this research was to focus on the impact of people analytics in reducing employee turnover. From the findings of this research, it can be concluded that people analytics helps in identifying whether an employee will leave the organization or not. Also, it identifies the significant features that lead to cause that turnover of employees in the organization. This shows that employee turnover can be diminished by organizations by using people analytics. The key aspects that have been identified which lead to definite employee turnover include satisfaction level of employees, promotions at workplace, hours spent at the job, number of projects completed, accidents at workplace and years spent at a specific job.

The other major objective of this research was to analyze the analytical skills and abilities which are required by human resource specialists to effectively manage employee data and integrate it in the process of the company. The core types of people analytics that are used by companies all over the world include descriptive, predictive and prescriptive analysis. This research has used predictive analysis to identify if an employee leaves the organization or not. In future, prescriptive analysis can be used which will involve the use of findings from this research to measure and understand how this data and its results can be used and how it will benefit the companies in future.

People analytics has many advantages for organizations and also leads to an increased business performance. From the results of entire research, it can be concluded that people analytics are a great measure which can be used effectively in order to identify the best candidates. Hence, helping in the hiring process. For instance, during the screening process of recruitment process for a specific role, people analytics can be used for recognizing the core competence within a specific program.

In future studies, the researchers can make use of machine learning algorithms other than decision tree such as; linear regression, SVM algorithm, Naïve Bayes algorithm, KNN algorithm, or gradient boosting algorithm etc. can be applied in model. There is an opportunity for the researchers to gather company specific data and analyze the impact of people analytics in managing employee turnover. For this research, the generalizability of model has not been defined. For finding the generalizability of this model, a large number of datasets maybe included and tested in the research in future. Also different limits on decision tree and entropy can be applied in the model to make it more generalized such as; applying threshold limit for measuring entropy. A major limitation of this research is that, it does not involve the company specific data rather uses random data from a random website. Moreover, people analytics can be applied more specifically to other areas of human resource as well. For instance, future researches can consider the application of people analytics on identifying salary and rewards, recruitment and selection as well as to engagement of employees.

References

Aslam, H. D., Aslam, M., Ali, N. and Habib, M. B., 2014. Importance of Human Resource Management in 21st Century: A Theoretical Perspective. International Journal of Human Resource Studies, 3(3).

Becker, B. and Gerhart, B., 2010. The Impact of Human Resource Management on Organizational Performance: Progress and Prospects. The Academy of Management Journal, 39(4).

Boakye, A., 2020. The Rise of HR Analytics: Exploring Its Implications from a Developing Country Perspective. Journal of Human Resource Management, 8(3).

Brown, J.B., 2018. Classifiers and their metrics quantified. Molecular informatics, 37(1-2), p.1700127.

Dutta, A. and Chaudhry, S., 2021. Managing people more effectively: Challenges and best practices. Journal of Management Research and Analysis, 8(1).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

Jones, D.C., Kalmi, P., Kato, T. and Mäkinen, M., 2008. The effects of human resource management practices on firm productivity: Preliminary evidence from Finland (1121), ETLA Discussion Papers.

Kamiński, B., Jakubczyk, M. and Szufel, P., 2018. A framework for sensitivity analysis of decision trees. Central European journal of operations research, 26(1), pp.135-159.

Kaufman, B. E., 2014. Theorising determinants of employee voice: An integrative model across disciplines and levels of analysis. Human Resource Management Journal, 25(1).

Kingsford, C. and Salzberg, S.L., 2008. What are decision trees?. Nature biotechnology, 26(9), pp.1011-1013.

Lamrini, B., 2020, Contribution to Decision Tree Induction with Python: A Review, in D. Birant (ed.), Data Mining - Methods, Applications and Systems. IntechOpen, London. 10.5772/intechopen.92438.

Marler, J. H. and Boudreau, J. W., 2016. An evidence-based review of HR Analytics. The International Journal of Human Resource Management, 28(1).

Marmolejo-Ramos, F. and Tian, T.S., 2010. The shifting boxplot. A boxplot based on essential summary statistics around the mean. International Journal of Psychological Research, 3(1), pp.37-45.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill International. pp. 57-58

Mitchell, T.M., 1999. Machine learning and data mining. Communications of the ACM, 42(11), pp.30-36.

Peeters, T., Paauwe, J. and Van De Voorde, K., 2020. People analytics effectiveness: developing a framework. Journal of organizational effectiveness: people and performance, 7(2), pp. 203–219.

Sagar, D., 2020. HR Analytics: People Management. [Online]. Available at: Kaggle Dataset, https://www.kaggle.com/datasets/dineshsagar66/hr-analytics-people-management, [Accessed 2.3.2022].

Shrivastava, S., Nagdev, K. and Rajesh , A., 2018. Redefining HR using people analytics: the case of Google. Human Resource Management International Digest, 26(2), pp. 3-6.

Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130.

Thakre, N., 2021. Human Resource (HR) Analytics For Organizational Transformation And Effectiveness. Training & Development Journal, 50(47-51).

Tursunbayeva, A., Di Lauro, S., and Pagliari, C. (2018). People analytics—A scoping review of conceptual boundaries and value propositions. International Journal of Information Management, 43(1), 224–247.

Vermeeren, B. et al., 2014. HRM and its effect on employee, organizational and financial outcomes in health care organizations. Human Resources for Health, 12(1).