UNIVERSITY OF LIVERPOOL

Big Data Analytics for Business

(EBUS633)

Individual Coursework

Professor: Gopalakrishnan Narayanamurthy

Report By:

Student ID- 201595383

Content

List of Figures

List of Tables

1.Motivation

The process of gradual but deliberate reduction of employees over time is defined as attrition. The high attrition rate in IT sector today is resulting in many issues affecting productivity and quality. Therefore, hiring the right person for a job has become crucial. A mapping of the organization's expectations and employees' expectations is essential. Mismatch between expectations and interests becomes root cause of dissatisfaction in job which leads to attrition. (Saraf and Peshave, 2020)

HR incorporates huge amounts of data into its work from pay data, employee data, engagement scores, feedback, surveys, email responses to performance reviews. Using data visualization, it can help the HR teams interact and engage with their data and metrics, and if done correctly, it facilitates faster and more accurate decision-making. The advantages of data visualization for HR teams can lead to improvement in retention rates, making better hiring decisions, elevating your workforce planning etc. which can be seen in the below images: (Kilpatrick, 2021)

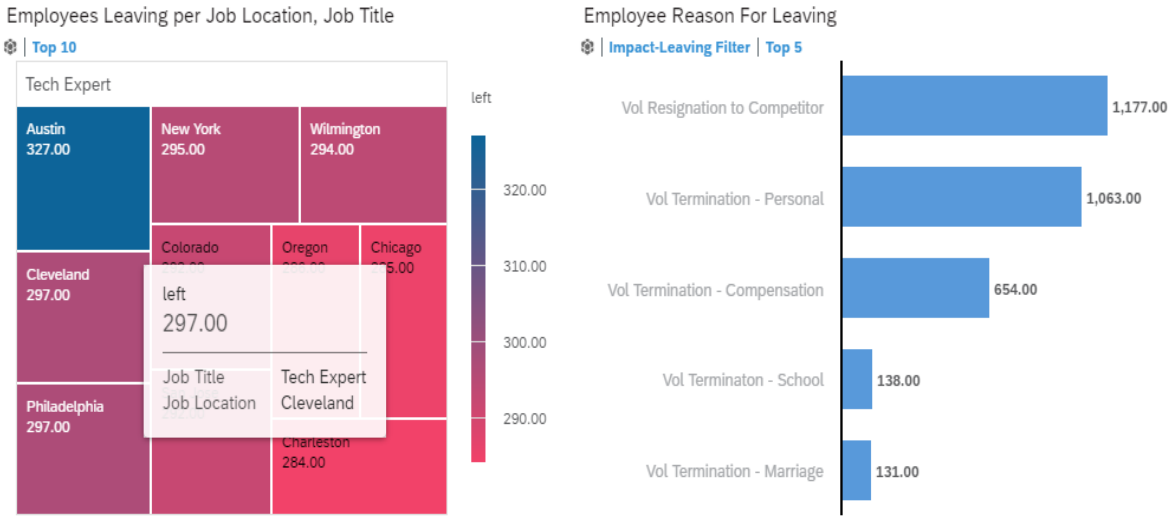

Figure 1: Employee reason and job location-Attrition (Kilpatrick, 2021)

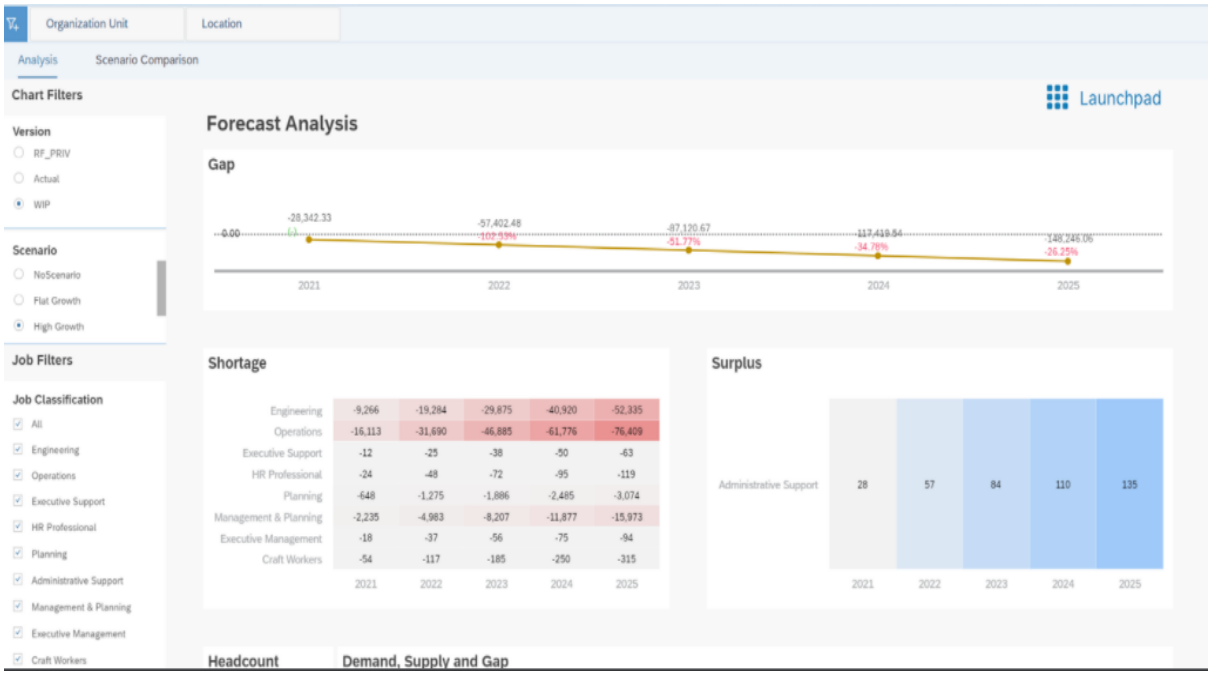

Figure 2: Forecast for work shortages and risks (Kilpatrick, 2021)

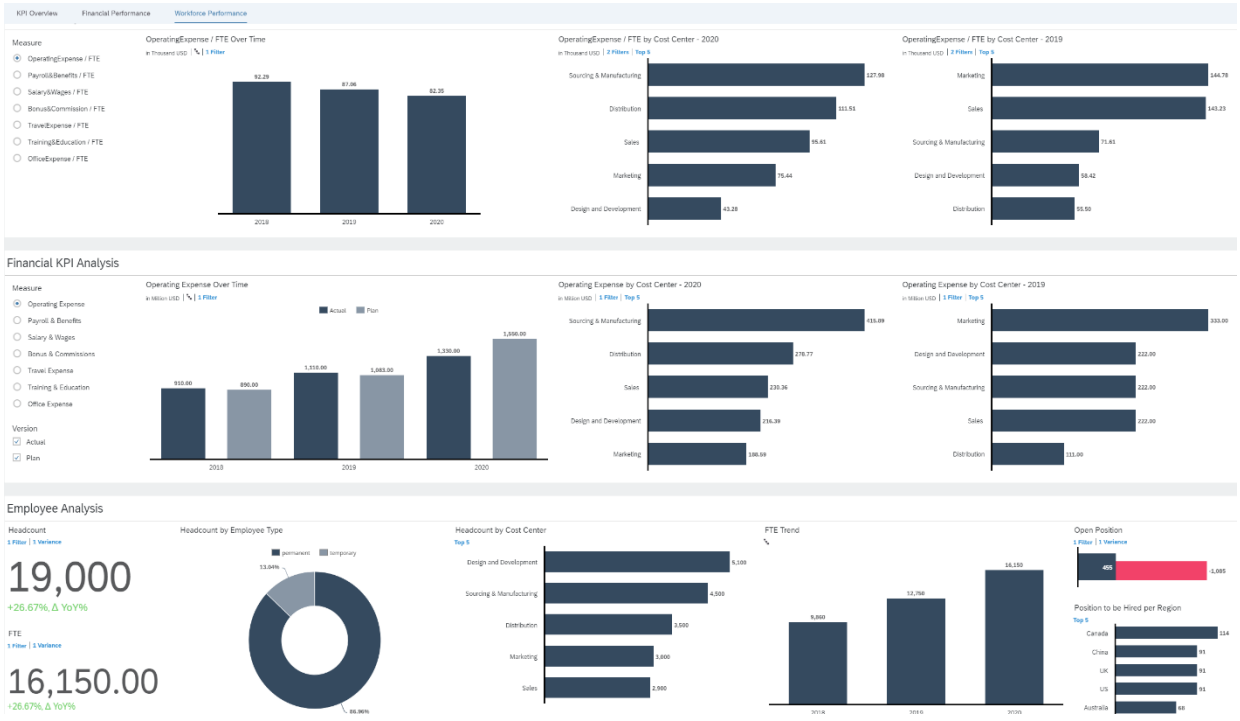

Figure 3: Workforce planning and analysis (Kilpatrick, 2021)

Another Big data analytics technique that is used by Organizations to understand the influence of employees on Enterprise Performance is Logistic Regression analysis. A survey considering Human Resource, Employee satisfaction and Enterprise performance can be conducted as shown below: (Jing, 2020)

Question	Mean of score	Variance of score
X1: Job opportunities for you	7.45	0.75
X2: Adoption of suggestions from grassroots employees	6.89	0.89
X3: Emphasis on work experience	8.75	0.74
X4: Payment by labor	9.45	0.71
X5: Funds and time for personnel training	8.45	0.68
X6: Information disclosure to grassroots employees	7.56	0.68
X7: Opportunities of promotion for internal employees	7.43	0.65
X8: Internal grievance opportunities for employees	7.23	0.78
X9: Performance assessment of employees every year	6.89	0.63
X10: Attention to performance in employee promotion	6.98	0.76
X11: Ability to resolve internal conflicts	6.78	0.74
X12: Clarity of tasks assigned to employees	7.59	0.68
X13: Span of work scope provided to employees	7.98	0.63
X14: Implementation of senior staff management	7.01	0.62
X15: Harmonious relationship between employees	7.45	0.75
X16: Benefits offered to employees	7.23	0.74

Figure 4: HR Management Survey score (Jing, 2020)

Question	Mean of score	Variance of score
Y1: Relationship between employees	7.56	0.68
Y2: Performance evaluation	7.43	0.65
Y3: Task proportion	7.23	0.78
Y4: Promotion opportunities	6.89	0.63
Y5: Role assignment	6.98	0.76
Y6: Welfare	6.78	0.74
Y7: Overall management	7.43	0.65
Y8: Training	7.23	0.78

Figure 5: Employee satisfaction survey score (Jing, 2020)

Question	Mean of score	Variance of score
Z1: Profitability and asset income	7.56	0.68
Z2: Market development and user expansion	7.43	0.65
Z3: Employee training opportunities	7.23	0.78
Z4: Customer satisfaction	6.89	0.63
Z5: Market value and industry evaluation	6.98	0.76
Z6: Operation ability	6.78	0.74

Figure 6: Enterprise performance survey score (Jing, 2020)

A significant regression relationship between two of the three group variables shows a positive correlation amongst themselves as shown below:

Relationship	Item	Nonstandard regression parameter		Standard regression parameter	T	Significant probability
		B	Error			
Human resource	constant	2.445	0.245		7.004	0.001
	Employee satisfaction	0.234	0.075	0.527	4.123	0.001
Human resource	constant	1.726	0.284		4.512	0.001
	Enterprise performance	0.254	0.099	0.648	4.425	0.001
Employee satisfaction	constant	0.732	0.124		2.789	0.001
	Enterprise performance	0.218	0.071	0.325	3.456	0.001

Figure 7: Regression Analysis (Jing, 2020)

This helps the organization to understand which variable amongst the group is the least performing and measures required to mitigate it.

2.Literature Review

Visualization implies displaying data in a picture or graphic form. To assess and extract insights from big data, data visualization needs to be interpreted formally. By extracting and displaying useful patterns, the user is able to discover simple, yet insightful insights into secret knowledge. It is used to determine the areas that require improvement, concentrating on factors that impact employee behaviour, and forecasting revenue potential. (Khalid and Zeebaree, 2021). Big data analysis requires visualization. Tables and statistics alone provide a limited view into big data. For effective analysis of big data, visualization must be integrated into analytics tools so that users from all kinds of backgrounds can access data from a wide range of sources, including clickstreams, social media, log files, and videos. (Keahey, 2013)

With data visualisation, business value goes way beyond identifying success metrics; it opens up creative discussions and collaboration opportunities at every level - from boardrooms to sales teams. Depending on expertise, departments might interpret findings differently and form unique opinions, leading to innovations that might not otherwise have been possible. Not only does data visualisation facilitate engagement, but it also allows the presentation of information in a manner that is easy to understand and comprehend. Consequently, stakeholders within an organization may use these visuals to enhance and accelerate decision-making. (O'Neill, 2017)

Logistic regression serves as a statistical method for analysing a dataset where there are one or more independent variables that determine what will happen. The use of logistic regression allows anyone to say that the existence of a risk factor makes that case more likely. The relationship between a dependent variable or feature and more than one nominal or ordinal independent variable is described by this method of description of data. (Sukhadiya, Kapadia and D'silva, 2018).

Data classification is useful for knowledge discovery and intelligent decision-making. The volume and complexity of data makes it impossible to categorize and sort effectively without proper classification technique. By organizing the data into categories, it helps in analysing the massive quantity of information. In this way, an accurate model can be developed for each defined class using the appropriate data features of the respective class. (Pramanik, Pal, Mukhopadhyay and Singh, 2021)

3.Case Example

Research work on this topic is mainly focused on creating a model that will be able to predict if an employee will leave a company or not using Visualization and Classification technique. In essence, the aim is to measure how well employee appraisals perform and how satisfied employees are with their jobs in addition to various factors that influence employee attrition rate and possible solutions. (Jain, Jain and Pamula, 2020)

Data Collection and Pre-processing

Dataset description-There are over 14,000 records in this dataset, which contains 10 features related to the attrition rate of employees. Description of selected attributes and department wise employee count is mentioned in Table 1 and Table 2 respectively.

Satisfaction level	Employee satisfaction level in the company, where 0 represents the least satisfied and 1 represents most satisfied
Last evaluation	Employee last evaluation (rating) in the company, where 0 represents the least rating and 1 represents most rating
Number of projects	Total number of the project done by an employee in his/her carrier
Average monthly hours	Mean of hours spent by the employee in the company, on monthly basis
Time spend company	Mean of hours spent by the employee in the company, on daily basis
Work accident	Numeric attribute values (0 or 1). if any accident/escalation happened with the employee in the company per month
Left	Target attribute value (0 or 1). Where 0 represents employee not left the company and 1 represents employee left the company
Promotion on last 5 years	Numeric attribute values (0 or 1). Where 0 represents employee don't get any promotion in last 5 years whereas 1 represents employees who received the promotions
Sales	The information about employees department. This is a categorical variable which has seven departments
Salary	Categorical variable divding the salary of employees in 3 broad categories (low, medium and high)

Table 1: Selected attributes (Jain, Jain and Pamula, 2020)

Department name	Employees count
Sales	4140
Technical	2720
Support	2229
IT	1227
Product_mng	902
Marketing	858
RandD	787
Accounting	767
HR	739
Management	630

Table 2: Department wise employees count (Jain, Jain and Pamula, 2020)

Data Exploration

The techniques of variable identification, univariate analysis and bi-variate analysis are applied on this dataset to further analyze the data and bring out important aspects.

Variable Identification- Firstly, predictor variables are identified as input variables and target variables as output variables. Secondly, data type and category variables are identified as shown in Table 3.

Attributes	Data type	Variable category	Type of variable
Satisfaction_level	Numeric	Continuous	Predictor
Last_evaluation	Numeric	Continuous	Predictor
Number_of_projects	Numeric	Categorical	Predictor
Average_montly_hours	Numeric	Continuous	Predictor
Time_spend_company	Numeric	Continuous	Predictor
Work_accident	Numeric	Categorical	Predictor
Promotion_last_5years	Numeric	Categorical	Predictor
Domain	Character	Categorical	Predictor
Salary	Character	Categorical	Predictor
Left	Numeric	Categorical	Target variable

Table 3: Variable identification (Jain, Jain and Pamula, 2020)

Uni-variate analysis- Statistical measures for categorical and continuous variables are applied individually.

Continuous variables- Here, the focus is on the mean, standard deviation, and spread of variable as shown in Table 4.

	Satisfication level	Last evaluation	Number of projects	Average monthaly hours	Time spend in company	Work accident	Left	Promotion in last 5 years
Count	14999	14999	14999	14999	14999	14999	14999	14999
Mean	0.61283	0.716102	3.80305	201.05033	3.49823	0.14461	0.238083	0.021268
Std	0.24863	0.171169	1.23259	49.94309	1.46013	0.351719	0.425924	0.144281
Min	0.09	0.36	2	96	2	0	0	0
25%	0.44	0.56	3	156	3	0	0	0
50%	0.64	0.72	4	200	3	0	0	0
75%	0.82	0.87	5	245	4	0	0	0
Max	1	1	7	310	10	1	1	1

Table 4: Uni-variate analysis (Jain, Jain and Pamula, 2020)

Categorical variables- Here, Figure 8 represents projects being handled by different departments.

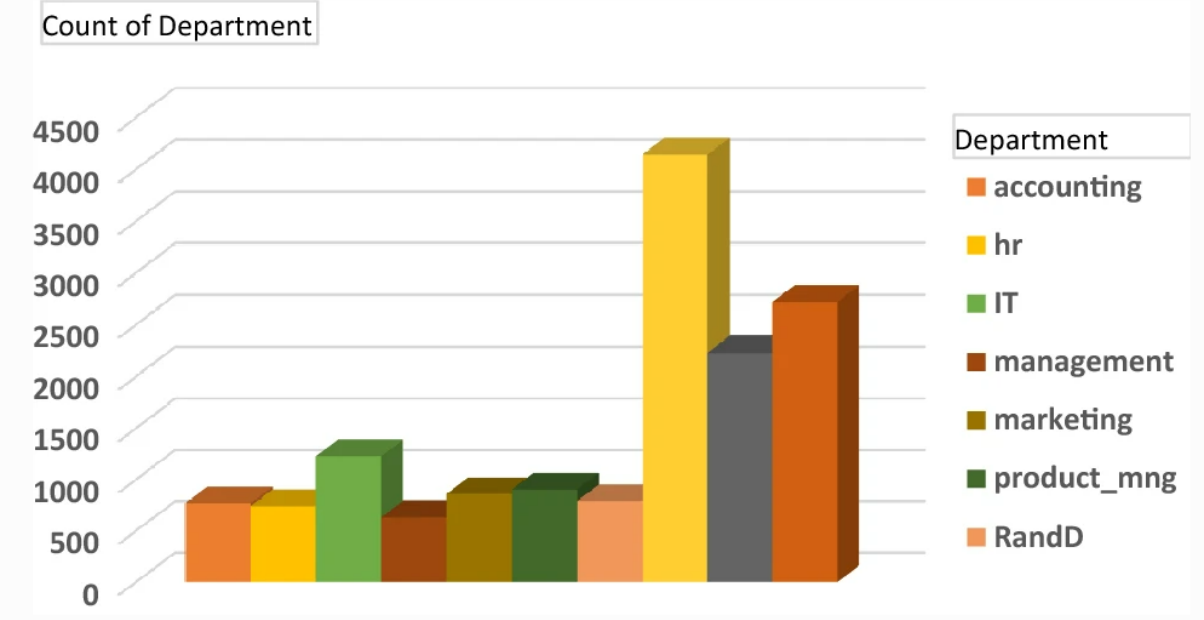

Figure 8: Departmental strength- Projects handled (Jain, Jain and Pamula, 2020)

Figure 9 shows different types of projects from type 2 to type 7 which shows most number of projects are available from type 4 and least from type 7. Figure 10 shows number of hours spent by employees.

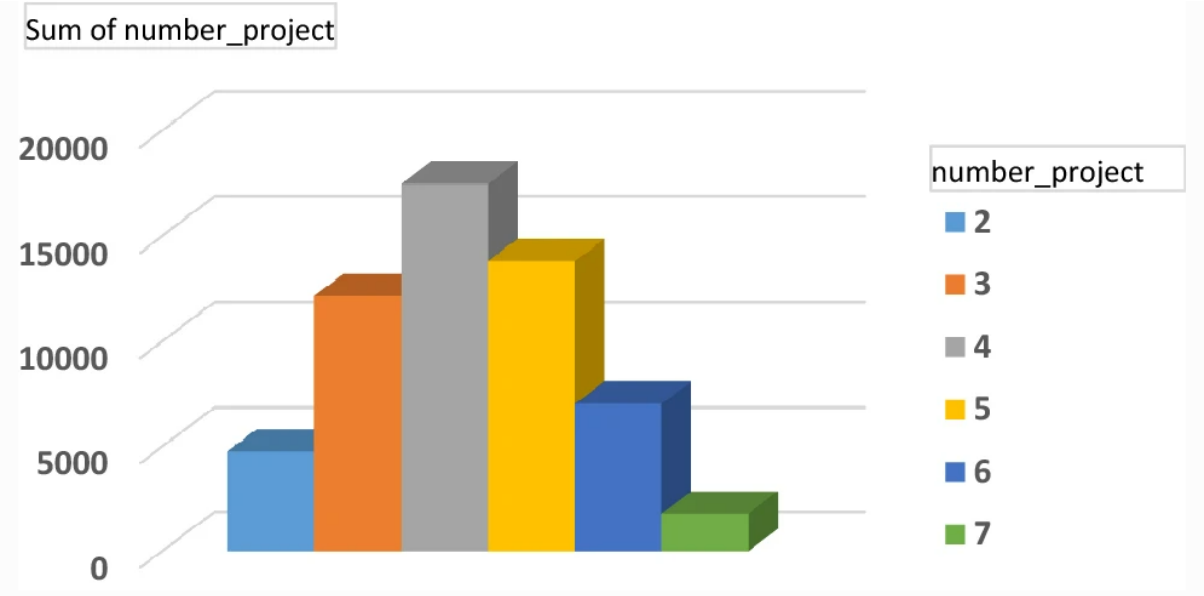

Figure 9: Types of Projects available (Jain, Jain and Pamula, 2020)

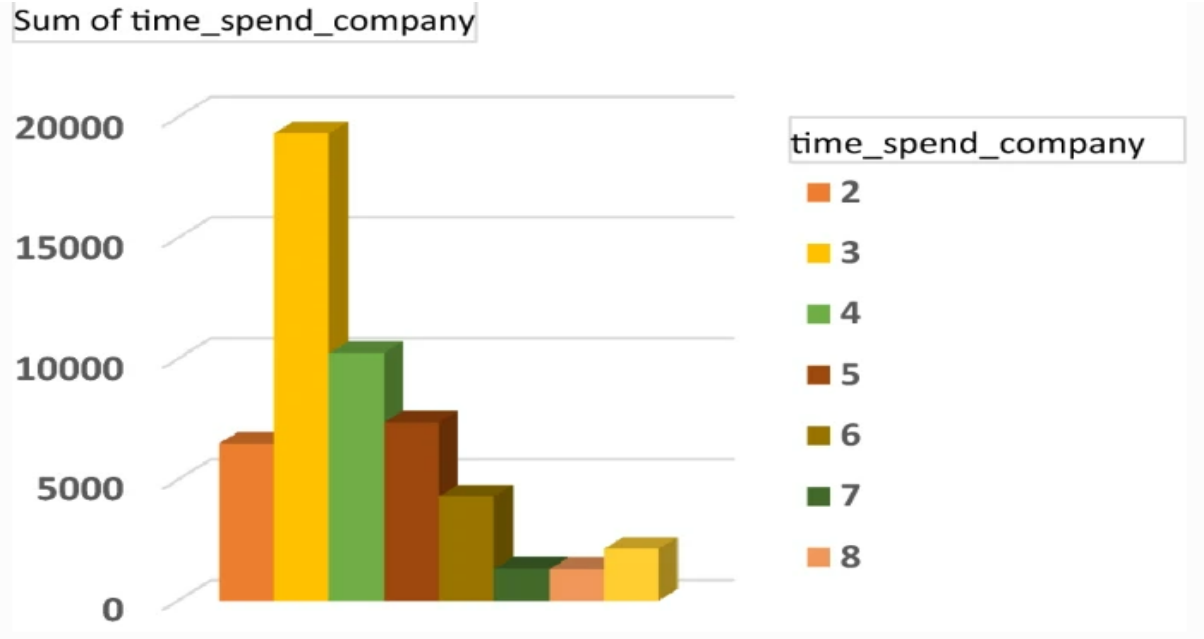

Figure 10: Hours spent by employees (Jain, Jain and Pamula, 2020)

Bi-variate analysis- To determine the relationship between two variables, bivariate analysis is performed.

Categorical and continuous- Here, the strength of relationship is determined by using correlation graph as shown in figure 11 whose values lie in the range of -1 to +1.

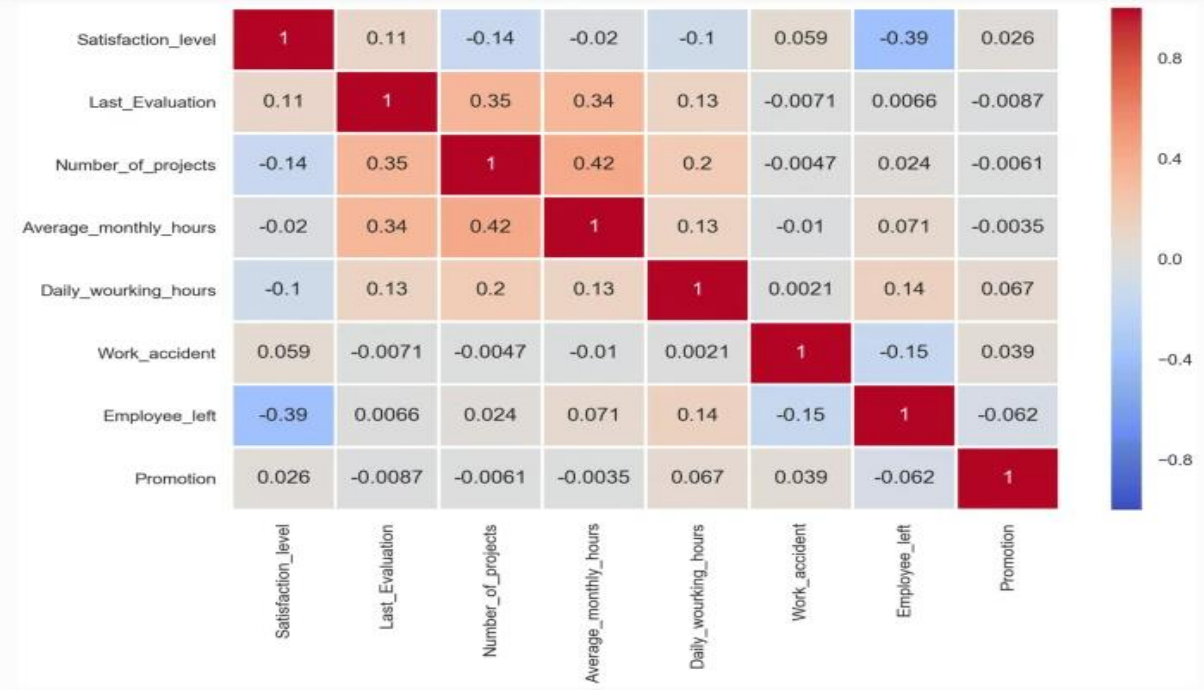

Figure 11: Correlation between different variables (Jain, Jain and Pamula, 2020)

Data Visualization

Satisfaction level versus employee left- Figure 12 depicts high chance of employees who might leave the company whose satisfaction level is 0.1 or less and those between 0.3 and 0.5.

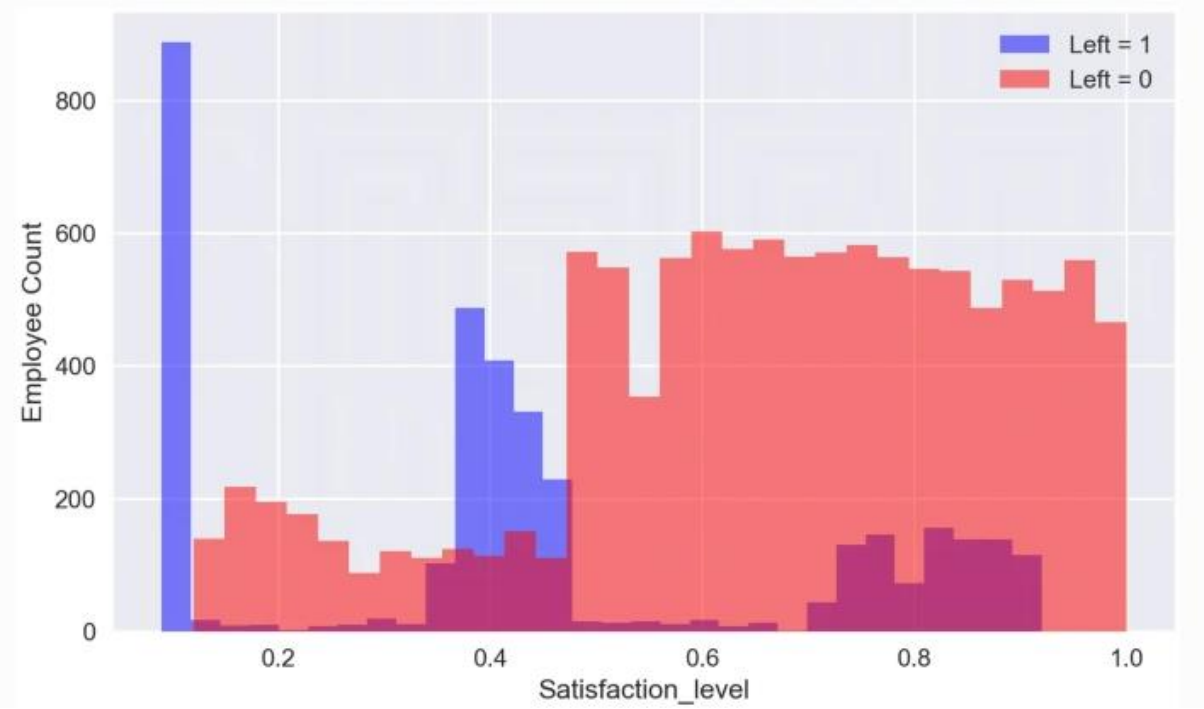

Figure 12: Satisfaction level v/s Employee count (Jain, Jain and Pamula, 2020)

Monthly hours versus number of projects- Figure 13 shows as an increase in project count increases monthly hours proportionally.

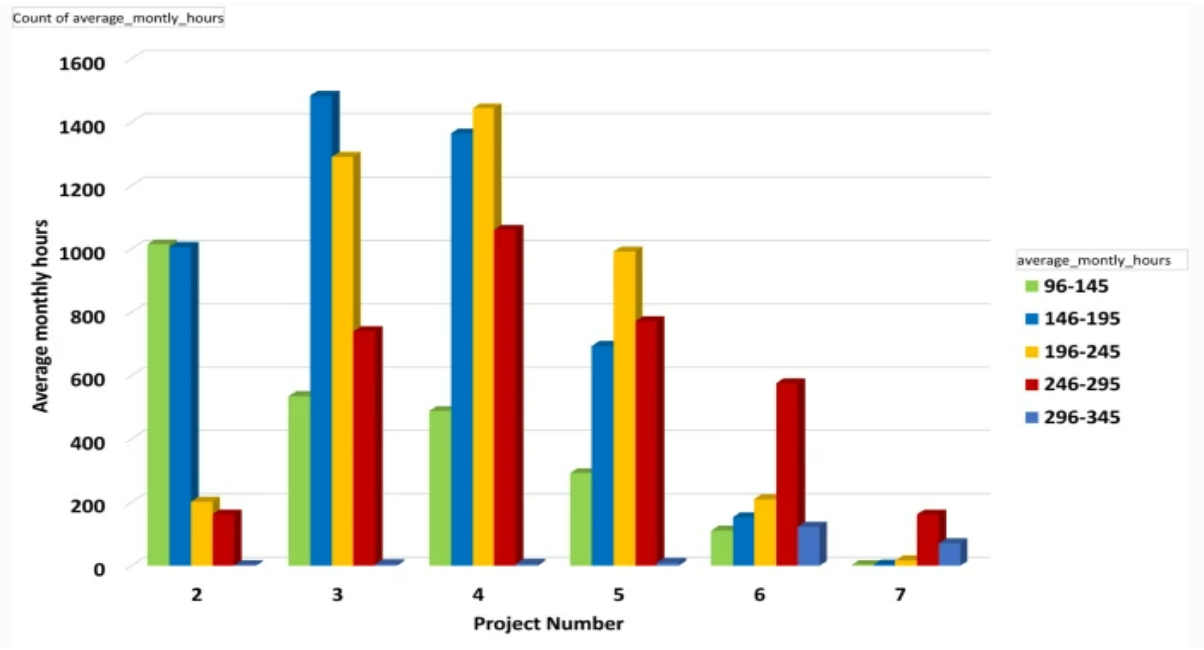

Figure 13: Monthly hours v/s no. of projects (Jain, Jain and Pamula, 2020)

Last evaluation versus number of projects- Figure 14 depicts last evaluation is directly related to number of projects.

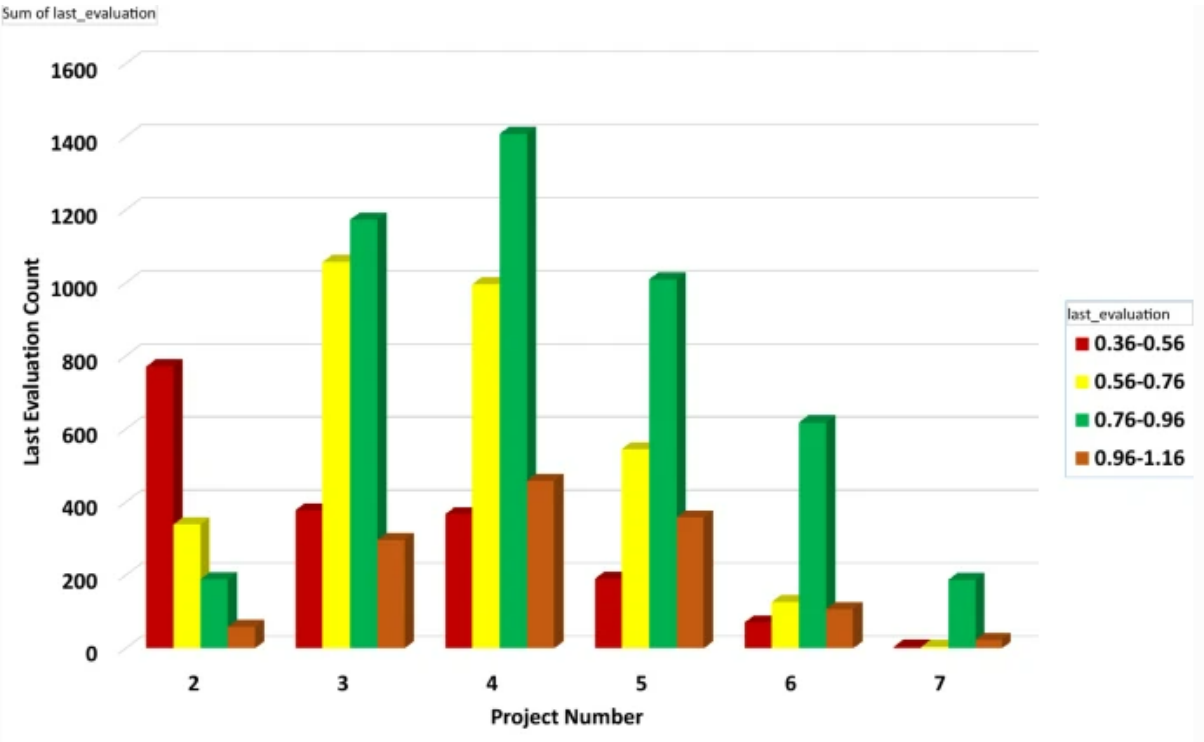

Figure 14- Last evaluation v/s no. of projects (Jain, Jain and Pamula, 2020)

Employee's count versus last evaluation- Figure 15 shows low performance and high performance are two main reasons that employees may leave the organization.

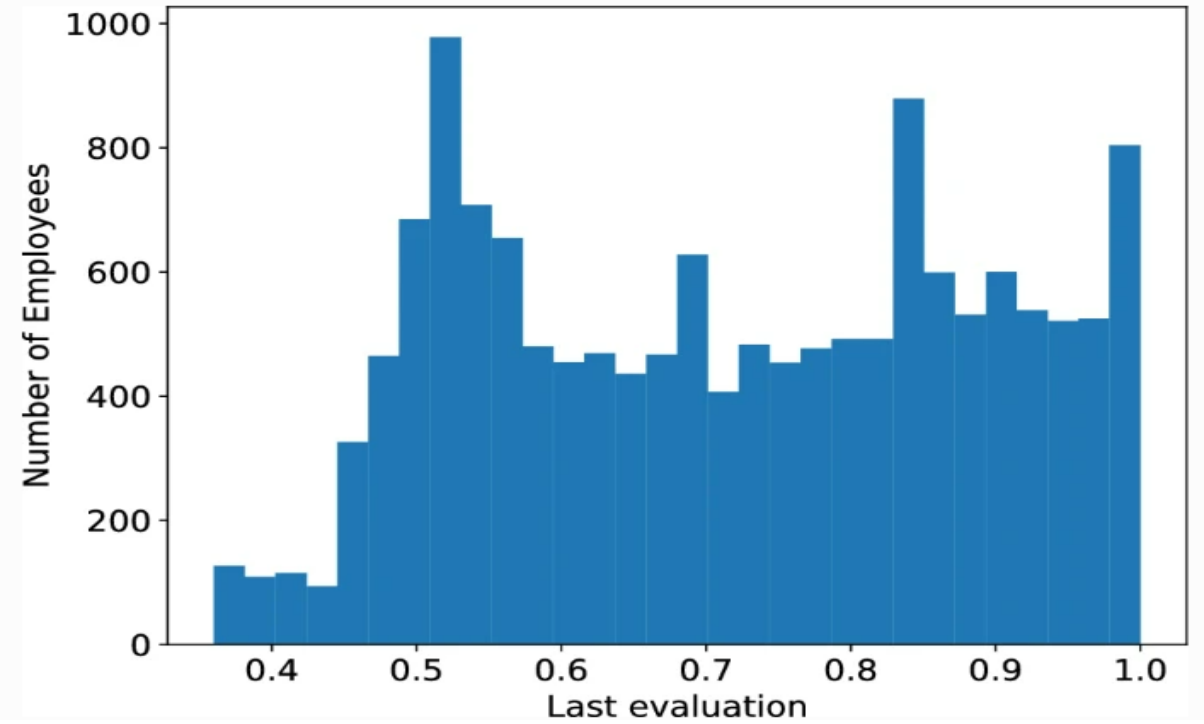

Figure 15: Employee's count v/s last evaluation (Jain, Jain and Pamula, 2020)

Salary versus employees left- Figure 16 suggests that majority of employees who left were in low to medium salary range.

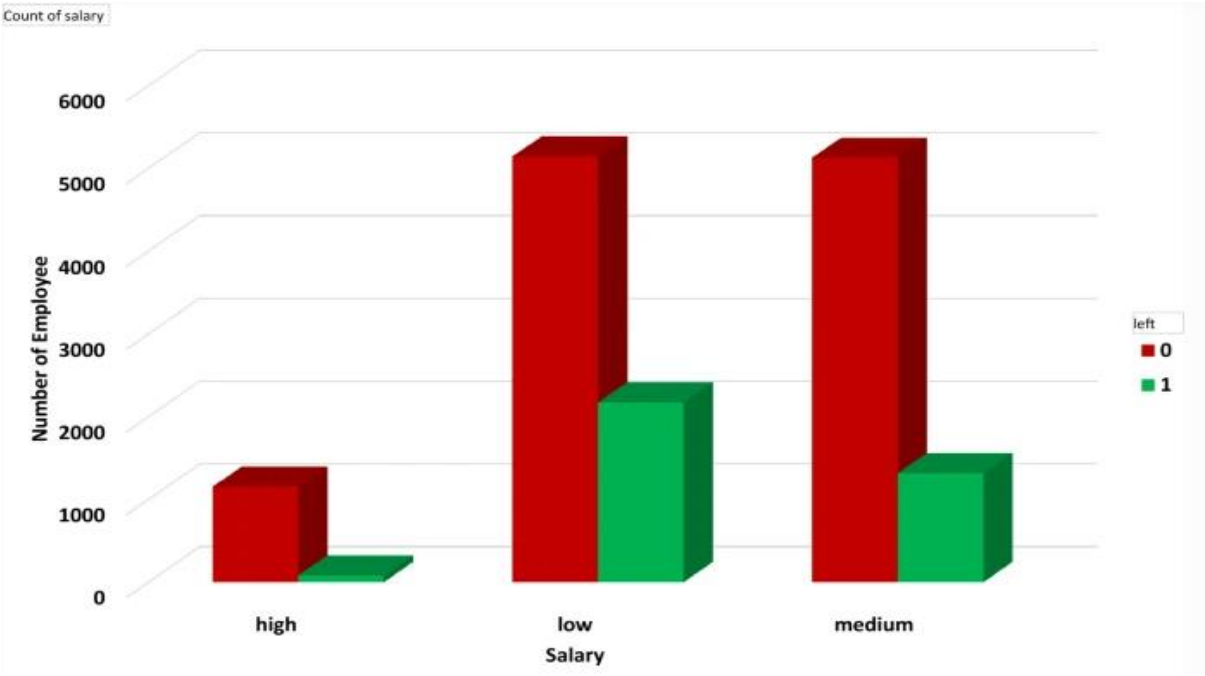

Figure 16: Salary v/s employee left (Jain, Jain and Pamula, 2020)

Department versus employee left- Figure 17 indicates that the technical, sales, and support departments had the highest levels of employee turnover.

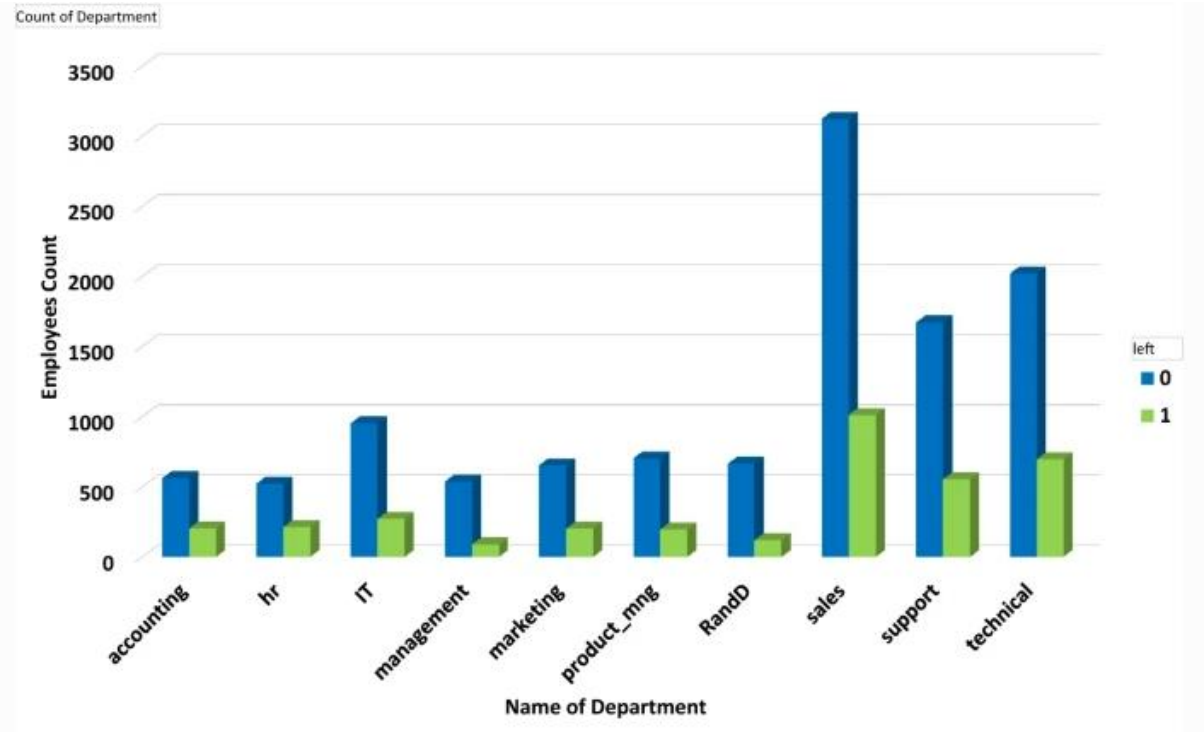

Figure 17- Department v/s employee left (Jain, Jain and Pamula, 2020)

Model formulation

Various predictive models such as Decision Tree, Random Forest and Support Vector Machine as in this case can be used and one with the best classification results can be employed for prediction. They are used on training and testing data by using the cross-validation operator.

Model evaluation

Confusion matrix provides a visual representation of performance of a classifier, giving information about number of true positives, false positives, true negatives, and false negatives. The classification report discusses precision, recall and F-1 score for the model as shown in Table 5.

Department	Algorithm	Precision	Recall	F1-score
Sales	DT	95	98	97
	SVM	88	92	90
	RF	99	98	98
Technical	DT	89	98	93
	SVM	83	91	87
	RF	97	97	97
Support	DT	90	97	93
	SVM	89	95	92
	RF	98	97	98
IT	DT	97	98	97
	SVM	86	83	84
	RF	99	98	99

Table 5- Department wise result analysis (Jain, Jain and Pamula, 2020)

Findings

The departments that contain more than 1000 employees were selected to undergo classification technique for prediction. The characteristics of classifiers are more likely to detect the majority class and less likely to identify the minority class, which in this case is an employee not leaving the company. When evaluating the classifiers with the confusion matrix, the Random Forest outperforms all the other classifiers. One reason Random Forest performs better than Decision Tree may be that the latter picks up decision boundaries at random in each step, rather than choosing the best one.

Performance Impact

Based on this analysis, the organization can target the employees most likely to leave the organization, then provide them with confined incentives. There could be case of false positive where the management thinks that employee would leave but they do not and also case of false negative, where the management neglects the employee and they leave the organization. This condition needs to be balanced and type of treatment to employees should be weighted accordingly. (Jain, Jain and Pamula, 2020)

4.Barriers

The limitations and challenges of implementation of visualization can be identified as:

Integration of data- Creating effective dialogue processes often involves trying to visualize and then act upon all the parameters relevant to a planning process. It is important to balance qualitative, social values with more quantifiable physical values. As such, big data is associated with groups rather than individuals. In order to integrate traditional datasets with crowdsourced data, where control is less, we need methods.

Representation of data- The challenge is how to represent information visually using digital models. It is imperative that interactive visualizations have a deeper understanding of the use of light and colour, and of the appropriate level of detail and realism.

Avoiding misinterpretation- It is possible to misinterpret visualizations in ways in which they were not intended. The inclusion of too much detail and visual realism in visualizations at the beginning of a planning process is inefficient and even misleading, since that information will not be decided until later on. (Billger, Thuvander and Wāstberg, 2016)

The barriers on the adoption of classification depends on characteristics and the nature of the data and the performance of learning algorithms. While examining real-world data, a more thorough investigation of data collection methods is needed. It is also possible for the historical data to contain ambiguous values, missing values, outliers, and meaningless information. Thus, for the associated application domains to use the classification technique effectively, existing pre-processing methods need to be modified or enhanced or new data preparation techniques proposed. (Sarker, 2021)

5.Recommendations & Roadmap

An organization that uses data analysis to make strategic decisions has a significant competitive advantage in overcoming current challenges and preparing for the future. In many organizations, however, access to data and the skills to analyze it are restricted to IT and business intelligence teams. Data visualisation and Classification are more than skills with software, it's a way to interpret and communicate data's significance to others. (Stokes, 2021)

A four-step action plan can help organisations to reduce attrition and retain top talent in accordance to Visualisation and Classification techniques.

Lead- This includes the leaders of the organisation to showcase their commitment to the importance of visualising the data at meetings to participating in training alongside staff. This in turn trickles down to frontline managers and HR leaders is making use of the data effectively.

Train- Self-service learning, role-based training and advanced skill development can improve their productivity and foster innovation. The data capturing and pre-processing thus forms an integral part of the entire process, in addition to data exploration which helps in understanding the different types of variables stored in the data.

Measure- Success indicators such as performance metrics, job satisfaction and compensation can help the organisation to understand the correlation between employee's expectations and organisation's interests and quantifying the data visualization and classification technique results can help encourage buy-in from top executives. Skill assessments can also be useful for managers to understand the strengths and weaknesses and implement new training techniques for enhancing employee's productivity.

Support- The outcomes that are generated after Visualization and Classification techniques need to be worked upon with the help of intuitive software. The processed data should be presented in a meaningful manner and attrition trend may be represented graphically so that the results can deduce the chances of an employee leaving the organisation. (Srivastava and Nair, 2017)

6.Conclusions

The key focus of this report was to mainly invest in the future of HR analytics as an integrated part of the HR department to assist managers in making predictive decisions based on statistical evidence, HR analysis data, and literature. The focus was also on examining the IT infrastructure and technological interventions, such as those that impact the way data is mined, stored, and made, which are crucial to successful HR analytics implementation and are necessary in order to be efficient.

A predictive model that resembles the one outlined in the case example can be applied to raw data to ensure accurate inferences and insights that assist organisations in reaching their growth goals. However, there is further potential for improving the existing literature resources through case studies of predictive models for determining the relevance and utility of each model created for this specific industry sector. (Mohammed, 2020)

References:

Billger, M., Thuvander, L. and Wästberg, B., 2016. In search of visualization challenges: The development and implementation of visualization tools for supporting dialogue in urban planning processes. *Sagepub*, [online] 44(6), pp.1012-1035. Available at: https://journals.sagepub.com/doi/10.1177/0265813516657341

[Accessed 24 March 2022].

Jain, P., Jain, M. and Pamula, R., 2020. Explaining and predicting employees' attrition: a machine learning approach. *SN Applied Sciences*, [online] Available at: https://link.springer.com/article/10.1007/s42452-020-2519-4#citeas

[Accessed 23 March 2022].

Jing, C., 2020. Regression Analysis of the Influence of Human Resource Management on Enterprise Performance. In: *International Conference on Management Science and Industrial Economy (MSIE 2019)*. [online] Macau: Atlantis Press. Available at: https://www.researchgate.net/publication/341200805_Regression_Analysis_of_the_Influence_of_Human_Resource_Management_on_Enterprise_Performance/references

[Accessed 22 March 2022].

Keahey, T., 2013. *Using visualization to understand big data*. [ebook] Available at: https://dataconomy.com/wp-content/uploads/2014/06/IBM-WP_Using-vis-to-understand-big-data.pdf

[Accessed 22 March 2022].

Khalid, Z. and Zeebaree, S., 2021. Big Data Analysis for Data Visualization: A Review. *International Journal of Science and Business*, [online] 5(2), pp.64-75. Available at: https://www.researchgate.net/publication/348750308_Big_Data_Analysis_for_Data_Visualization_A_Review

[Accessed 22 March 2022].

Kilpatrick, B., 2021. Why Data Visualization is Now Essential for HR. [Blog] *SAP*, Available at: https://blogs.sap.com/2021/05/11/data-visualization-for-hr/#:~:text=Data%20visualization%20encourages%20your%20HR,high%2Dstakes%20decision%2Dmaking.

[Accessed 21 March 2022].

Mohammed, A., 2020. HR ANALYTICS: A MODERN TOOL IN HR FOR PREDICTIVE DECISION MAKING. *Journal of Management*, [online] 6(3), pp.51-63. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3525328

[Accessed 24 March 2022].

O'Neill, P., 2017. *The business value of visualisation*. [online] ITProPortal. Available at: https://www.itproportal.com/features/the-business-value-of-visualisation/

[Accessed 22 March 2022].

Pramanik, P., Pal, S., Mukhopadhyay, M. and Singh, S., 2021. Big Data classification: techniques and tools. *ScienceDirect*, [online] pp.1-43. Available at: https://www.sciencedirect.com/science/article/pii/B9780128202036000023

[Accessed 24 March 2022].

Saraf, V. and Peshave, M., 2020. AN ANALYSIS ON EMPLOYEE-ATTRITION IN IT INDUSTRY. *Mukt Shabd Journal*, [online] IX(VII). Available at: http://shabdbooks.com/gallery/280-july2020.pdf

[Accessed 21 March 2022].

Sarker, I., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, [online] Available at: https://link.springer.com/article/10.1007/s42979-021-00592-x#citeas

[Accessed 24 March 2022].

Srivastava, D. and Nair, P., 2017. Employee Attrition Analysis Using Predictive Techniques. *SpringerLink*, [online] 1, pp.293-300. Available at: https://link.springer.com/chapter/10.1007/978-3-319-63673-3_35#chapter-info

[Accessed 24 March 2022].

Stokes, N., 2021. Why your workforce needs data literacy. [Blog] *Tableau*, Available at: https://www.tableau.com/en-gb/about/blog/2021/7/why-your-workforce-needs-data-literacy

[Accessed 24 March 2022].

Sukhadiya, J., Kapadia, H. and D'silva, M., 2018. Employee Attrition Prediction using Data Mining Techniques. *International Journal of Management, Technology And Engineering*, [online] 8(X), pp.2882-2888. Available at: http://ijamtes.org/gallery/369%20oct%20ijamte%20-%201127.pdf

[Accessed 22 March 2022].