PREDICTIVE ANALYTICS IN EMPLOYEE CHURN: A SYSTEMATIC LITERATURE REVIEW

Ardhianiswari D. Ekawati, Bina Nusantara University

ABSTRACT

The increase of capacity in the collection, storing, and analyzing the massive volume of data due to the rapid advancement of information technology has changed how the decision-makers in organizations approach their work. Human Resource Information Systems provide a lot of data on the employees, but there are still a few best practices on how to make use the abundant data for better decision making in Human Resources area. This limited implementation is mostly due to the decision making process in Human Resources is highly depended on the intuition, the use of advanced data analytics is still in the very early stage of implementation which is left behind in comparison to the use of advanced analytics in other areas such as Marketing, Sales or Finance. This article answers the question on how is the current implementation of predictive analytics in employee churn or employee turnover which is one of the most discussed topics in Human Resources Management, and which method of predictive analytics are better to use in predicting the employee churn. The answers of these questions are obtained from the result of a systematic literature review on the scholarly articles related to the topic published from 2000 to 2018 in major databases such as IEEE, ACM, Science Direct, Emerald and Willey Online.

Keywords: Predictive Modeling, Employees Churn, Big Data, Human Resource Analytics, Intelligent Human Resource Systems

INTRODUCTION

The advance of information technology, especially with the implementation of Human Resource Information Systems (HRIS) in organizations, has also pressured the Human Resource (HR) professionals to better use the data that have been collected throughout the years for better decision making. So far, the use of the HRIS data is mostly for reporting simple descriptive statistics and correlations. Organizations have proven the use of advanced data analytics in other areas such as marketing, sales, and finance for better results, for example, to reach a better group of customers or reduce the risk of investment. But the use of advanced data analytics in the Human Resources area is still limited. According to research performed by IBM, by interviewing 700 chief human resource officers, less than 25% are using sophisticated analytics to predict future outcomes and for decision making (IBM, 2010). The CIPD report observed that there is still a significant gap in the ability of HR professionals to be data-driven and evidence-based in HR decision-making process. This is indicated by CIPD HR Outlook survey in 2012-2013 that only 63% HR leaders think they draw insight from data, and only 21% of their non-HR business counterparts share that confidence in HR data (CIPD, 2013). Human Resource professionals still prefer to rely on their intuition in their decision-making process; they valued and felt more comfortable with the ability to interpret ambiguity and context, understanding the shifting

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

cultures of organizations and the interactions of the people in their organizations (CIPD, 2013). This still remains true from the latest CIPD HR Outlook survey in 2016-2017, that less than 50% of HR leaders using HR Analytics in the area of performance management (48%), attraction, recruitment and selection (47%), learning and development (44%), workforce planning (38%), workforce performance and productivity (22%), and most of them are still in the basic-medium level in terms of sophistication used. On average, out of all the organizations which used HR Analytics in the previously mentioned issues, only 12.8% have been using HR Analytics in advanced level (CIPD, 2017). Russell & Bennett (2015) also suggested that this is mainly because many of the relevant variables in the Human Resources area, for example, personality, are difficult to measure. Another reason is that the relationship between these variables and organizational performance is not entirely understood.

Predictive analytics covers a wide range of techniques such as statistics, modeling and data mining that use current and historical facts to make predictions about the future (Fitz-enz & Mattox, 2014). Predictive analytics have been commonly performed in the area of marketing, such as in the prediction of customer churn (Hassouna et al., 2015) or in finance, for assessing financial risk tolerance of portfolio investors (Ardehali et al., 2005), but the application in human resources are still limited in comparison to the application in marketing and finance areas. Several implementations of predictive analytics in the form of data mining have been performed in some Human Resources activities, such as prediction of employee turnover, prediction of severance pay acceptance, and prediction of employee performance (Strohmeier & Piazza, 2013).

Employee churn is a big concern for organizations especially in the current competitive environment where people are the biggest asset of organizations. The cost of a voluntary employee churn is ranging from 1.5 to 5 times of the employee's annual salary depending on how difficult to fill the employee's position (Sesil, 2014). But this is actually only the small part of the overall cost to an organization due to the fact that the loss of an employee might affect the ongoing projects or services. It could, for example, lead to dissatisfaction of the customers and other stakeholders (Saradhi & Palshikar, 2011). It could lead to the loss of network or other important knowledge especially in the case of losing an experienced long-time employee. It also could affect the other employees since they have to cover the workload of the employee who left the organization. Another thing that has to be mentioned is that it takes some time for the new employee to reach a certain expected level of expertise and productivity that owned by the previous employee. The employee churn itself has been one of the essential topics in Human Resource Management, in the last one hundred years more than two thousand articles on voluntary employees churn have been published from 1920 to 2020 (Lee et al., 2017). Research on voluntary employee churn and turnover intention is mostly based on survey data, but analyzing actual employee churn requires longitudinal data to see whether the employee left in a specific time span, so in many cases, turnover intention and not the actual employee churn is investigated (Rombaut & Guerry, 2018). With the implementation of HRIS, organizations have been gathering more quantity of data that has not been used in its full potential yet. Thus, interest in performing research on using predictive analytics for employee churn has been growing in recent years.

In this article, a systematic literature review is conducted to research complete and structured literature on the implementation of predictive analytics in the Human Resource area,

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

particularly related to employee churn. The research question of this research is ‘what method of predictive analytics is better to use in predicting employee churn?' The purpose of this research is:

1. To analyze several methods of predictive analytics in the case of employee churn.
2. To find some trend in the latest progress of predictive analytics in employee churn.

It is expected from this research to contribute to the better use of data analytics in HR management decisions related to the employee churn.

LITERATURE REVIEW

According to Gartner, Inc. (Gartner, 2019), advanced analytics is the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools, typically beyond those of traditional business intelligence (BI), to discover deeper insights, make predictions, or generate recommendations. Advanced analytic techniques include techniques such as data/text mining, machine learning, pattern matching, forecasting, visualization, semantic analysis, sentiment analysis, network, and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks.

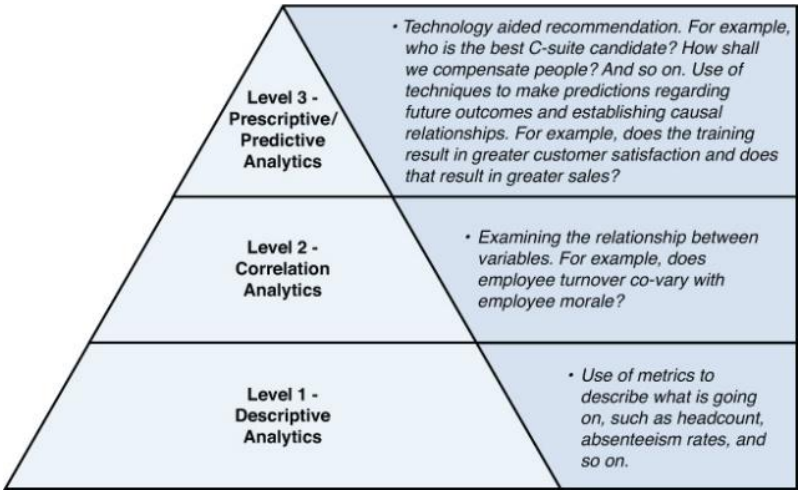

FIGURE 1
HIERARCHY OF ANALYTICS (SESIL, 2014)

Figure 1 provides an overview of a hierarchy of analytics. Level I is an organization's use of basic metrics to obtain information such as headcount, employee turnover, and even some simple statistics such as the use of means and averages. The next level is Level II, which is characterized by correlations. It also includes determining whether and when variables move relative to one another. Level III shows a focus on establishing causation and on predictions of what will happen next (Sesil, 2014).

There are several common methods that are implemented for predictive analytics in employee churn (Dolatabadi, 2017):

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Decision Tree and Random Forests

The decision tree is one of the famous methods in predictive analytics because of its ease of interpretation. The tree-shaped model represents decisions and decision making. The learning algorithm is started from observations of a particular item, represented as the branches and conclusions about the target value of the item which is represented in the leaves. The algorithm forms a tree with a training dataset where each node is represented by an attribute and its branches are the attribute values. A concern with this algorithm is that any small changes in the training data might create a large variation in the classification performance, which leads to an unstable algorithm. Breiman developed random forests to solve this issue by reducing the tendency of over fitting the training dataset (Breiman, 2001). Random forests, initially introduced by Ho (1995), construct multiple decision trees on the training data and resulting in the class that is the mode of the classes of the individual tree.

Naïve Bayes

Naïve Bayes is a popular classification technique for its simplicity and efficiency. It is a simple probabilistic classifier based on the application of Bayes' Theorem. The posteriori probabilistic, computed using the theorem and naïve Bayesian independence assumptions, is used to assign churn and non-churn given an employee record in the context of employee churn.

Artificial Neural Network

Artificial Neural Network (ANN) is a machine learning algorithm that simulates the behavior of a human neuron. The human brain comprises millions of neurons connected by a unique structure known as synapses. The synapses enable the neurons to process signals from one to another. The Artificial Neural Network applies this behavior to a large number of interconnected processing units that work together to process information and produces meaningful results from it. The process to train and adjust the strength of these connections to get the intended overall behavior is known as the learning process. The ability to learn automatically from the existing data to generate predictions is the outstanding feature that made this algorithm popular. It also has the ability to enforce hidden insights into the hidden relationships (Strohmeier & Piazza, 2013). In general, the Artificial Neural Network consists of two different categories based on their structures, the Feed-Forwards, and Feedback or Recurrent. In the Feed-Forward network, the neurons comprise three layers known as input, hidden and output layers. The network only allows signals to travel one way from input to output so that the output of any layer does not affect that same layer. This straightforward network is extensively used in pattern recognition. The Feedback or Recurrent network, on the other hand, allows the signals moving in both directions by creating loops in the network. This makes the network more complicated and dynamic but also makes them powerful. The output of the previous inputs is fed back into the network and this provides them with some kind of memory. The states of Feedback network are continuously changing until an equilibrium state is achieved.

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning algorithm with each member of the training dataset is assigned to one or the other of two categories so that it is a non-probabilistic binary linear classifier (although it is also possible to use SVM in a probabilistic classification setting). In a linear classification, the input training data points are mapped in space (hyperplane), each with a different class label, which is divided by a clear gap from one another. The new data points are then mapped into the same space and predicted as part of a class according to which side of the gap they fall. SVM also can perform non-linear classification. In this case, data points are mapped using a non-linear function using the kernel trick in a high dimensional space (Cortes et al., 1995).

METHODOLOGY

This research is implemented using the Systematic Literature Review (SLR) approach that was proposed by Weerakkody et al. (2015). This approach consists of several sections, namely: defining research question that was explained on introduction, determining research sources, accomplishing the finding process by using keywords, extracting data, and analyzing the findings to answer the research question.

Figure 2 describes the process of the Systematic Literature Review:

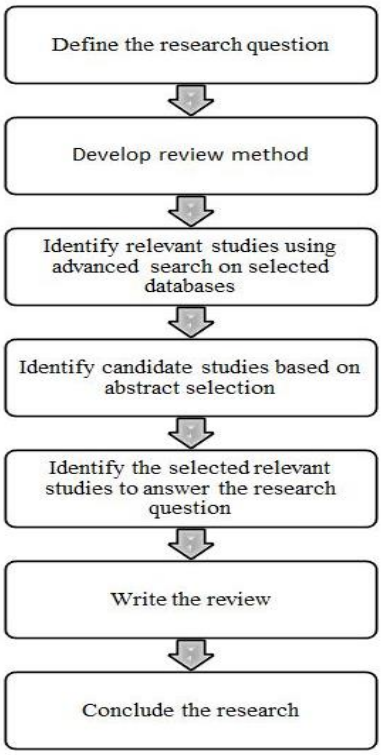

FIGURE 2
SYSTEMATIC LITERATURE REVIEW

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Search Process

From the research question that has been defined in the Introduction section, the selected sources of this SLR can be found as follow:

- IEEEXplore Digital Library (http:/ieeexplore.ieee.org)
- ACM Digital Library (https://dl.acm.org/)
- Science Direct (www.sciencedirect.com)
- Wiley Online Library (onlinelibrary.wiley.com)
- Emerald Insight (www.emeraldinsight.com)

Keyword search is applied to look for papers that are related to the defined research question, using a Boolean operator such as: AND, OR, NOT. All of the mentioned sources above own a keyword-based search engine. The defined search strings that are used for the keyword source are:

("Analytic" OR "Model") AND "Prediction" AND ("Employee" OR "Worker") AND ("Churn" OR "Turnover") AND ("Big data" OR "Data mining" OR "Neural")

The keywords consist of ‘analytic/model', ‘prediction', ‘employee/worker', ‘churn/turnover', ‘big data/data mining/neural' from the period of 2000-2018.

Inclusion and Exclusion Criteria

The results of the advanced search using the keywords are the papers related to the keywords that have been defined. The total number of papers from the search results is considered as ‘studies found'. In the next step, when the title is not sufficient to determine whether to include the paper as a candidate or not, the abstract is then read. If the title and the abstract match with the previously defined research question, this paper will be downloaded for further investigation. The number of downloaded papers is called ‘candidate studies'. All ‘candidate studies' papers' results will be read thoroughly to search for the answer to the research question. The papers that answered the research questions are the ones that will be used in the research as ‘selected studies'. The detailed result can be found in Table 1.

Table 1
NUMBER OF STUDIES IN SELECTED SOURCES

Source	Studies Found	Candidate Studies	Selected Studies
IEEE	5	3	1
ACM	13	6	2
Science Direct	97	8	3
Emerald	62	4	1
Willey Online	27	2	0
Others	2	2	2
Total	206	25	9

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Data Extraction

This systematic literature review was conducted from February 2018 and examined in a total of 204 papers. In Table 1, it can be seen that among 204 examined papers, there are 23 papers that are related to the research question based on their title and abstract. However, after being thoroughly studied there are only 7 papers that can be included in this research. Others appear in sources because there are some papers that have been mentioned in 'selected studies' by the author. Thus, these papers are not included in the previously mentioned sources.

RESULTS AND DISCUSSION

Demographic and Trend Characteristics

Source of publications

Only nine papers are considered in the 'selected studies' categories and 66.7% of papers are published in the journal. Two papers are published in the same journal "Expert Systems with Application" as can be seen in Table 2 (Quinn et al., 2002; Sexton et al., 2005; Fan et al., 2012; Sesil, 2014; Shah et al., 2016; Rombaut & Guerry, 2018; Dolatabadi, 2017; Cahyani & Budiharto, 2017; Berengueres et al., 2017).

Table 2
SOURCE OF PUBLICATIONS

Journal/Conference	Journal/Conference Name	#	%
Journal	Expert Systems with Application	2	22.2
Journal	Management Research Review	1	11.1
Journal	Journal of Technology in Human Services	1	11.1
Conference	The 2nd International Conference on Computer and Communication Systems	1	11.1
Conference	The 9th International Conference on Machine Learning and Computing-ICMLC 2017	1	11.1
Conference	2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	1	11.1
Journal	Journal of Business Research	1	11.1
Journal	Computers & Operations Research	1	11.1

Academic disciplines

Based on the authors' academic background as can be found in Table 3, interestingly although the papers were written the most by the authors from Computer Science and Information Systems background, the other disciplines such as Science and Technology and Business and Administration are following the Computer Science and Information Systems closely. Thus, basically, predictive analytics is not only the 'game' of people with the background of Computer Science and Information Systems only.

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Background of authors

Most authors come from the academic background, and even the authors from industry are all from the research and development area (with Computer Science background). There is no HR professional yet who are involved in the research in predictive analytics in employee churn (Table 4).

Table 3
DISCIPLINE OF AUTHORS

Department	#	%
Computer and Information Systems	8	32
Science and Technology	5	20
Business and Administration	5	20
Education and Social Work	3	12
Industrial Engineering	2	8
Psychology	1	4
Mathematics	1	4
Total	25	100

Table 4
BACKGROUND OF AUTHORS

Background of authors	#	%
Academic	18	72
Industry	7	28

Publication trends

From Table 5, there is a trend for more publications in predictive analytics for employee churn in 2017. The growing interest in big data and other predictive analytic methods might be the reason for the increase papers on employee churn prediction using advanced analytics methods.

Table 5
FREQUENCY OF PUBLICATIONS

Year	#	%
2017	5	55.6
2012	1	11.1
2011	1	11.1
2008	1	11.1
2004	1	11.1

University affiliation according to country

The authors come from diverse backgrounds in terms of institutional affiliation according to the country. The authors of the nine selected papers come from 20 countries and 16 institutions. As can be seen in Table 6, most of the authors are from universities in the USA.

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Methods of predictive analytics

From Table 7, it can be seen that the Neural Network is the most common method to predict employee churn. It is also the method that has been used in the first paper in 2004 and still being used in the latest paper in 2017. Interestingly, with more papers published on employee churn prediction in 2017, there are more methods that are implemented in the recent papers.

Table 6
UNIVERSITY AFFILIATION ACCORDING TO COUNTRY

Countries	# Institutions	% Institutions	# Authors	% Authors
USA	4	25%	7	28%
Pakistan	1	6.25%	1	4%
United Kingdom	1	6.25%	2	8%
India	1	6.25%	2	8%
Belgium	1	6.25%	2	8%
Taiwan	2	12.5%	4	16%
Iran	2	12.5%	2	8%
Indonesia	1	6.25%	2	8%
United Arab Emirates	1	6.25%	1	4%
Spain	2	12.5%	2	8%
Total countries: 10	16		25	

Table 7
METHODS OF PREDICTIVE ANALYTICS

Methods*	# Papers					% Papers
	2004	2008	2011	2012	2017	
Neural Network	1	1		1	1	44.4
SVM			1		2	33.3
Naïve Bayes			1		1	22.2
Decision Tree/ Random Forest			1		1	22.2
Logistic Regression		1			1	22.2
Big Data					2	22.2
Gradient Boosting Machine					1	11.1

*One paper can use more than one methods

Predictive Analytics in Employee Churn

Saradhi & Palshikar (2011) and Dolatabadi (2017) consider that the application of predictive analytics in employee churn can be approached the same way as in the case of customer churn. Dolatabadi (2017) compares several methods: Decision Tree, Support Vector Machine, Naïve Bayes, and Neural Network for employee churn and customer churn prediction. For employee churn, Decision Tree, Naïve Bayes, and Neural Network (all 100%) result in slightly better total accuracy than the Support Vector Machine (99.55%). While for customer churn, the Support Vector Machine (99.83%) performs the best in comparison to the other methods (85.1-92.37%). Support Vector Machine also gives a very high true positive accuracy

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

compared to the Random Forest and Naïve Bayes in the study performed by Saradhi & Palshikar (2011).

IMPLICATIONS AND LIMITATIONS

With more methods of predictive analytics, a more reliable and accurate prediction of employee churn can be obtained. This will help the Human Resource professionals to focus more on a certain group of employees to ensure that they will stay at the organization.

This paper has a limitation in that the number of databases is limited much because of restricted access. The amount of papers needs to be augmented mainly excavated from a credible database and published in the last 17 years.

CONCLUSION

From this review, it can be concluded that the study on predictive analytics for employee churn is still very attractive for both the people from the academic and industry background. Support Vector Machine, in general, provides a reliable result to predict the employee churn accurately. However, there are still lots of things to learn from all the methods, even if currently some methods already give a good accuracy for employee churn prediction. It is difficult to compare which method is the best from the papers that have been studied since the datasets come from different cases especially considering there are only a few available studies exist at the moment. With more studies on predictive analytics in employee churn, it is expected that the best method to predict the employee churn in a certain case of an organization can be found.

FUTURE RESEARCH

There is still a need to improve the accuracy of the employee churn prediction models and understanding which method is best used in a certain situation. A study with several distinctive cases (e.g. different sizes and types of organizations) with several methods can be performed to find which method results in better accuracy in each case. Some factors also still need to be investigated whether or not it makes a distinctive case from the implementation of a prediction model in employee churn in one organization to another.

While more researches on the implementation of prediction models of employee churn are still needed, in reality, it is still difficult to find organizations that are willing to participate in the studies. The reason is mostly because of the human-resources-related data are highly sensitive and confidential.

REFERENCES

Ardehali, P. H., Paradi, J. C., & Asmild, M. (2005). Assessing financial risk tolerance of portfolio investors using data envelopment analysis. *International Journal of Information Technology & Decision Making, 4*(3), 491-519.

Berengueres, J., Duran, G., Castro, D. (2017). Happiness, an inside job ? Turnover prediction using employee likeability, engagement and relative happiness, *In* IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 509-516.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.

Cahyani, A. D., & Budiharto, W. (2017). *Modeling Intelligent Human Resources Systems (IRHS) using Big Data and Support Vector Machine (SVM)*. *In* Proceedings of the 9th International Conference on Machine Learning and Computing-ICMLC 2017, 137-140.

CIPD. (2013). *Talent analytics and big data-the challenge for HR*. CIPD Research Report November 2013.

CIPD. (2017). *HR outlook-winter 2016-17*. Retrieved from https://www.cipd.co.uk/knowledge/strategy/hr/outlook-reports.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Dolatabadi, S. H. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor, in The 2nd International Conference on Computer and Communication Systems, Krakow, Poland, 0-3.

Fan, C. Y., Fan, P. S., Chan, T. Y., & Chang, S. H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications, 39*(10), 8844-8851.

Fitz-enz, J., & Mattox, J. (2014). *Predictive analytics for human resources*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Gartner. (2019). *Advanced Analytics*. Augmented Analytics and Artificial Intelligence in the Spotlight at Gartner Data & Analytics Summit, February 18-19 in Sydney, Australia.

Hassouna, M., Tarhini, A., Elyas, T., & Abou Trab, M. S. (2015) Customer churn in mobile markets: a comparison of techniques. *International Business Research, 8*(6), 224-237.

Ho, T. K. (1995). Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, *1*, 278–282.

IBM. (2010). *Working beyond borders*. Insights from the Global Chief Human Resource Officer Study. Executive Summary, IBM Global Business Services, Somers, NY, USA.

Lee, T. W., Hom, P. W., Eberly, M. B., & Mitchell, T. R. (2017). On the next decade of research in voluntary employee turnover. *The Academy of Management Perspectives, 31*(3), 201-221.

Quinn, A., Rycraft, J. R., & Schoech, D. (2002). Building a Model to Predict Caseworker and Supervisor Turnover Using a Neural Network and Logistic Regression. *J Technol Hum Serv, 19*(4), 65-85.

Rombaut, E., & Guerry, M. (2018). Predicting voluntary turnover through Human Resources database analysis. *Management Research Review, 41*(1), 96-112.

Russell, C., & Bennett, N. (2015). Big data and talent management: using hard data to make the soft stuff easy. *Business Horizons, 58*(3), 237-242.

Saradhi, V. V., & Palshikar, G. K. (2011). Expert systems with applications employee churn prediction. *Expert Systems with Applications, 38*(3), 1999-2006.

Sesil, J. C. (2014). *Applying advanced analytics to hr management decisions*. Pearson Education.

Sexton, S., Mcmurtrey, S., Michalopoulos, J. O., & Smith, A. M. (2005). Employee turnover : a neural network solution. *Computers & Operations Research, 32*(10), 2635–2651.

Shah, N., Irani, Z., Sharif, A. M., & Job, A. (2016). Big data in an HR context : Exploring organizational change readiness, employee attitudes, and behaviors Readiness. *Journal of Business Research, 70*, 366-378.

Strohmeier, S., & Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research, *Expert Systems with Applications, 40*(7), 2410-2420.

Weerakkody, V., Irani, Z., Lee, H., Osman, I., & Hindi, N. (2015). E-government implementation: A bird's eye view of issues relating to costs, opportunities, benefits and risks. *Information Systems Frontiers, 17*(4), 889-915.

Citation Information: Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences, 22*(4), 387-397.