Exploring Patterns: A classification approach to predict the likelihood of an employee leaving a company

Semin Islamovic  
National College Of Ireland  
Dublin, Irelandx22434792@student.ncirl.ie

*Abstract*—In this Analysis, we wish to deeply explore the use of various machine learning techniques to identify and predict employee attrition using a dataset which was collected from a company’s HR department. Employee attrition contains challenges for companies, which can have consequences on the overall people morale and company progress. The analysis will begin with exploring the actual raw data and preprocessing it, applying exploratory data analysis techniques, selecting the model and finally training it for results.

In due course, we will utilize different classification algorithms such as random forests, decision trees and KNN to train models for attrition prediction. The results told us that Random Forest was the best performing model with an accuracy score of 88%.

This analysis will contribute to the HR analytics area by utilizing these machine learning methods in predicting employee attrition and demonstrating ways for companies to build on employee retention. By manipulating these predictive modeling methods, it can help companies greatly to reduce the risk of attrition and try to maintain a much more concrete and balanced work environment.

Keywords—Attrition, classification, prediction, models, random forests, accuracy.

# Introduction

In this modern time and age, where businesses compete at the highest level, employee attrition has made an appearance as a major concern for companies across the globe. Employee attrition is defined as a voluntary act of departure by an employee from a company, which slows down and greatly disrupts the work itself and poses a threat to the organization’s overall morale and performance. The consequences of attrition have a impact on significant costs which are related to recruiting workers, HR training and the total loss of produce. To this day, companies are more and more noticing and acting in making sure to address attrition accordingly to protect their competition and maintain long periods of business success.

There are a couple of different types of attrition that may occur. The first one we have already mentioned previously, voluntary attrition. The next one I want to briefly explain is the opposite which is Involuntary attrition. This takes place when the employee leaves the company due to some push factors such as performance judgement. The next type of attrition is Compulsory attrition where an employee reaches a certain age of retirement by the law rules and therefore leaves the company to go into pension. The final type is Natural attrition where factors such as death or terminal illness take place which is beyond the scale of control by the employee or the organization.

In our context of employee attrition, by using machine learning methods, we can gain great understanding from insights and eventually predict employee attrition. Organizations can benefit from leveraging heavy raw data by trying to determine the key factors and any relevant patterns of employee attrition. Therefore, this can then help them to come up with various strategies to mitigate this impactful risk. This analysis goes into the domain of HR analytics as mentioned before, by utilizing these machine learning techniques to predict employee attrition with the use of a dataset that was gathered online.

The dataset itself contains a wide range of attributes, which ranges from information to factors related to the job and different performance criteria. By making effective use of this raw data from our dataset, this classification analysis pursues to discover the wide range of key factors that potentially influence attrition and to create predictive models. In other words, this study’s main goals are to unlock useful insights from the predictive models and to provide the companies with effective direction to introduce new strategies for retention capability.

As an employee myself who is interested in the business and employee aspect, this motivated me to do this analysis to try to figure out its hidden factors. With the effective use of machine learning methodologies, can discover this and answer my main research question which is, “Does satisfaction with your job have a great influence on the likeness of employee attrition”. Following on from our research question, I also am intrigued to know to what extent can these predictive models be accurate based on other factors that are relevant to this topic in our dataset. Fundamentally speaking, our motivation lies in evidence-based intercession to address attrition challenges, accordingly, therefore boosting the maximum possible performance for an organization to ensure sustainability and worthy profits in long-term success.

In our next section, we will talk about the literature review and talk about how other people have conducted similar research analysis related to my topic of employee attrition. We seek to gain useful knowledge from this which we can use later on in our classification analysis.

# Literature Review

There are many researchers out there that have worked on and proved how Human Resource Management (HRM) has influenced employee production relationships. There is large margin of studies out there that have researched the causes, problems and various strategies which are related to the idea of employee attrition, giving us knowledgeable insights.

Mahesh et al. [1] proposed and tested a model combining financial performance, employee and customer service to predict employee attrition. Their aim was to discover if there were negative effects on the relationship between customer-facing employees of the companies and their relevant customers. They found that service brand image (SBI) was entirely interposed by customers’ evaluations of service delivery (CESD). This then resulted in SBI predicting the overall future profitability. By looking at this paper and comparing it with my topic, I can find very interesting and useful information. The main one would be that loss of human capital has an very impactful effect on the businesses capital. Their sample included staff and customers with 64 different offices across the USA. It was revealed that 86% of the units had downsized.

Adhikari et al. [2] focuses on the IT department in India and its relationship with very high attrition rate. PCA was first used to gather the concealed dimensions in job attrition. The author then approached the analysis using multiple regression to find out the reason behind job changing among the workers. The data was used from numerous IT companies and was later analysed using multivariate techniques. Although regression was used in this analysis, we can still gain benefit from this paper as Principal Component Analysis is used efficiently which we could potentially later use to further improve our mode performance or reduce any irrelevant dimensions from our initial dataset.

Falluchi et al. [3] explores how objective factors can influence employee attrition. A real-life dataset was provided with 35 features and 1500 instances for the test. The authors adopted a TDSP framework methodology to build a predictive model. Throughout the article, we get to learn and visualise different classification algorithms which will help us to decide easier which ones are best appropriate for our analysis.

Raza et al. [4] proposed which organizational factors cause employee attrition in companies. Four machine learning algorithms were utilized which gives us an even greater understanding and help for our initial analysis as well. Two of them being Logistic Regression and Decision Tree Classifier which are two algorithms which are used in classification problems which is ideal in our case. The authors used an IBM employee attrition with 35 features which are similar enough to the dataset we have chosen which is also beneficial for us to know exactly what we need to predict upon. Various plots and graphs were used to give us the conclusion that the trees classifier had the best accuracy with a 93% score.

Jain et al. [5] measured the effectiveness of of employee job satisfaction within the business. The authors discovered that anyone with 0.3 or less satisfaction levels, had much more chance to leave the company. Random Forest was used for categorizing the dataset which is very beneficial for us for our classification approach to this research study. In the end, Random Forest had an accuracy of 99% in this paper which clarifies its significance. Finally, by comparing and reading various online articles related to the topic of employee attrition, we can get a good overview and idea of where to go with the classification analysis to get the best possible prediction model for our results.

# Data Description

The dataset was gathered from a free online repository known as Kaggle [6] and downloaded as an Excel CSV file for our classification analysis. The main reason I chose this dataset is because I found it most suitable for the idea of employee attrition due to it consisting of various factors that could potentially influence attrition within a company. It was becoming clear that this dataset had offered specific advantages over other online dataset because of its initial relevance to topic and variability. It stood out for me mostly also because it’s concise and well-structured format, which makes life a lot easier when trying to understand the dataset and analyze afterwards. Given below is a summary table of the main characteristics of the dataset on employee attrition.

TABLE 1: DATASET SUMMARY TABLE

|  |  |
| --- | --- |
| ***CRITERIA:*** | ***DATASET:*** |
| Source | Kaggle |
| Size | 223kb |
| File Format | CSV |
| Number of Files | 1 |
| Type | Structured |
| Number of Attributes | 35 (numerical, categorical) |
| Number of Instances | 1,470 |

The “HR employee attrition” dataset holds relevant information such as the employees’ gender, age, job type, wage, department, job satisfaction levels, hourly rate, and many others. Each of these attributes can generate insights into the demographic framework of the employees and their main roles within the company. This data file includes all aspects of the company such as job satisfaction, organization and employee performance as mentioned previously which will be crucial factors to consider in our classification analysis.

The dataset itself is comprised of a total of 1,470 instances across the one file, to certify a relatively sufficient sample size for our classification analysis. The attributes of this dataset contain both categorical and numerical data types, proving the assorted essence of the data. By utilizing all this different information we are given by the file, it opens a fantastic opportunity to explore and research various and unique patterns across different dimensions. It is vital to mention that this dataset was obtained by an online website called Kaggle as mentioned before with essential ethical approaches taken into narration. References to the raw data sources are provided to make sure to enable reproducibility in succeeding analyses.

In summary, the gathered dataset provides an in-depth and diverse resource for researching the topic of employee attrition with a collective context. By utilizing the various dimensions included in the dataset and applying effective classification algorithms, our goal is to discover valuable insights that can unveil effective and sufficient decision-making processes and reveal mitigation techniques for this attrition problem to help companies globally.

# Methodology

This study heavily focuses on evaluating the performance of feature selection and classification algorithms on the employee attrition dataset. The research follows a structured approach as per the KDD methodology for data mining and machine learning. KDD (Knowledge Discovery in Databases) is a popular process that consists of extracting previously undiscovered and potentially valuable information from very large datasets [7]. It is an iterative process, and we will be going through each important step of the KDD process in this methodology section of the report with selecting the raw data itself being the first and most obvious step of the entire process. I have chosen to use the KDD process over the CRISP-DM process due to KDD allowing for a broader range of jobs and not just data mining itself. KDD is also associated with real-world applications, making it suitable for my analysis on employee attrition. Below shown is a visual representation of the KDD process [8].

A diagram of data processing process

Description automatically generated

Figure 1: Visual representation of the KDD steps

## Data selection and Understanding:

The first and major step of the KDD process is known as Data selection. It is defined as the process where raw data which is related to the analysis topic is gathered and finally decided on. As mentioned before, the dataset I selected for this analysis was the employee attrition dataset collected from Kaggle. After initially finding and selecting the dataset I am happy to continue my analysis on, the next crucial step is understanding and familiarizing myself with the dataset itself.

This involves exploring and understanding the structure, data content and any problems that may be included with the selected data. In our case, we first loaded the dataset into the Data Frame using a powerful data manipulation library called Pandas and explored the attributes of the dataset and tried picking out the key ones such as job roles, employee ages, satisfaction levels and so on. This is very important to reveal early so we can get a clear idea of what we are dealing with and know exactly what our target predictors are. While employee attrition datasets have been used in numerous studies, we aim to apply a comprehensive analysis with the use of various machine learning algorithms to receive higher and better accuracy and gain further insights into the driving factors that influence employee attrition. Our approach differed from past articles by focusing on a dataset that consists of various aspects of employee information such as both professional status and personal attributes also.

## Data Pre-processing and Transformation:

In our next essential step of pre-processing, we prepared the initial dataset for our analysis by cleaning the raw data to maintain high quality and standard in our dataset. Data pre-processing includes checking and handling missing values or errors, identifying outliers, encoding categorical attributes, and standardizing numerical attributes if necessary. With the help and use of specific functions in python, it was revealed to us that our dataset was clean and consisted of no missing values. This was an advantage to us and saved us time to focus on the rest of the pre-processing required to fully clean our dataset for analysis.

By viewing the head of the dataset in python it gave us an overview of the first five rows along with all the features included in the dataset. We straight away identified features that were deemed irrelevant to our classification analysis such as “Employee count”, “Over 18”, “Standard Hours” and “Employee Number”. We simply removed these columns in python and simplified our dataset overall. With the effective use of a Boxplot visual representation, it was very clear to us to identify any outliers that were present in the dataset. The “Monthly Income” column contained outliers and had to be dealt with accordingly.

The Interquartile Range (IQR) is another way of identifying outliers by calculating the difference between the 75th percentile (Q3) and the 25th percentile (Q1) and by applying this in python, we removed the outlier and updated our final clean dataset which was now ready for use and analysis. This pre-processing approach consisted of advanced methods to encode categorical attributes and deal with outliers effectively and accordingly. Data transformation is a key step to take in the KDD process to boost the performance and consistency of the analysis. Our transformation focused on a specific encoding technique known as one-hot encoding. This converted the categorical attributes to match a binary {0, 1} format to make it more effective and easier when using classification algorithms to perform models to the highest and best possible standard.

## Data Mining:

In our next stage of Data Mining, we applied Exploratory Data Analysis (EDA) techniques such as descriptive statistics and visualizations to reveal any relationships or patterns among the data itself. By visually representing these variables with the effective use of histograms and scatter plots, we gained insights into the distribution of the data and identified any potential predictors of attrition itself. A correlation matrix is a useful statistical method of seeing how strongly or weakly variables are related to each other [9]. Without a doubt, we applied this to our analysis and by creating a colored correlation matrix heatmap, it was much easier to spot out the most correlated variables in our dataset.

Attributes such as years at company, current role, with current manger, monthly income and total working years were found to be most correlated with each other. So, we picked out all these specific variables and conducted another correlation matrix just on these selected variables to further analyze the relationships. Pair plots, line plots, scatter plots and histograms were all effectively utilized on these highly correlated variables to further give us a clear idea and understanding of the leading factors of attrition. Other attributes such as “Marital Status” were appealing to me and therefore, conducted analysis on it also to see if we can reveal any patterns or relationships with this intriguing variable.

We then moved on and experimented with numerous machine learning algorithms to see for ourselves which one is most suitable and appropriate for predicting employee attrition itself. As part of our initial task, we applied three classification algorithms to further gather knowledge and extract trends or patterns from the clean and transformed dataset. The algorithms we finally decided to select were Random Forest, K-Nearest Neighbors (KNN) and Decision Trees to predict the leading cause of employee attrition. The reason behind choosing these algorithms is because they have the capability to handle complex relationships, they are very interpretable and produce great performance in handling real-life datasets with various and unique attributes which was ideal in case. They are all relatively simple and easy to understand and implement also, which is why I mainly chose to work upon them and complete our classification analysis on employee attrition. Our novelty involves experimenting with different algorithms and choosing different ones to those that were used in previous papers.

We then trained these machine learning models using a library in python known as scikit-learn. Our next essential step was to split the dataset into training and testing sets. We split the dataset in a 80-20 ratio where 80% of the data is used for training and 20% is used for testing. We specifically chose this ratio because by allocating 80% of the initial dataset to the training set, it allows for a sufficient amount data for the model to be trained from. Similarly, with a larger training set, overfitting will be less likely, which is important as this leads to poor performance on the data.

## Evaluation and Implementation:

For the final stage of the KDD methodology, we evaluated the performance of the trained models and implemented the outcomes to extract useful insights. We determined the model performance using powerful metrics such as recall, F1-score, precision, and accuracy. In our case of employee attrition, the accuracy measures the preciseness of the model’s predictions. The recall and precision provide us with insights to clearly identify the type of employees that are in risk of attrition.

Hyperparameter tuning is an essential part of the mode creation process, as it consists of optimizing the hyperparameters to improve the model’s performance. We applied hyperparameter tuning to each of the selected algorithms and used the random search technique to finalize it. I chose random search over grid search as it has many advantages over it such as random search allowing the ability to cover a wider range of values and reveal better outcomes and it generally requires fewer computational resources which is useful for complex models. During hyperparameter tuning, we got scores for KNN (0.84), Decision trees (0.83), and Random Forest (0.85). After tuning we ran the python on the test set again to see if we got any score improvements. We seen an increase in KNN with a score of 0.85, Random Forest with a final score of 0.88 but a slight decrease in the decision tree algorithm with a score of 0.82. This slight decrease may be because of the algorithm’s sensitivity to certain hyperparameter settings but will not have a huge effect on our outcome.

# Implementation

In the world of Data Mining and Machine Learning, the implementation section is extremely important to achieve accurate and knowledgeable insights and targeted predictions from the raw data. In this section of our report, we will primarily be focusing on the use of Python and Weka, the selection of components, technical challenges faced during the analysis and various forms of analytics. For our classification analysis on the selected employee attrition dataset, we mainly used python as our source of analyzing and creating predictive models and using Weka to reproduce these steps and results for comparing in the end.

Python is a coding language that offers a wide range of powerful libraries for data manipulation, data analysis, and model making which is ideal in our case for applying classification algorithms to predict the main cause of employee attrition. On the other hand, Weka is also a powerful GUI (Graphical User Interface) tool that provides you with many machine learning algorithms and relevant tools for data cleaning, classification, regression, clustering, etc. For data pre-processing, python includes libraries such as Pandas and NumPy, while Weka has a pre-process module. Both libraries were efficiently utilized to clean our data and make it appropriate for our analysis. Python offers great data manipulation abilities while Weka’s pre-process module offers different methods such as normalization and transformation.

SciKit-Learn is another powerful Python library that heavily focuses on creating new features for feature engineering and model selection, training, and evaluation. Weka’s Attribute Selection offers very similar capabilities for feature selection based on various criteria’s such as correlation and chi-square. For working on the model itself, Weka contains a classify module which provides us with a wide range of classification algorithms to consider for our analysis on attrition. Model selection involves choosing the most appropriate machine learning algorithm for the given task and its dataset. In our case, as mentioned before, we chose the Decision Tree, Random Forest and KNN. Both Python and Weka were well able to utilize these algorithms and models were created accordingly for results to compare and finally decide upon which is most sufficient.

For the final step of the entire KDD process, which is the evaluation of the final model, both python and Weka also include libraries to accommodate this procedure. Python continues to use SciKit-Learn while Weka uses the evaluation module. Model evaluation looks and analyses the overall performance of the trained models using different metrics such as accuracy, recall and precision as mentioned before. Python offers various functions for cross-validation and confusion matrices while Weka provides similar functions for evaluating models and some of these include cross-validation and various statistical tests.

Some certain Python web frameworks like Django can be used for creating web applications that showcase machine-

learning models as APIs which stands for Application Programming Interfaces. On the other hand, Weka made models can also be deployed using Java for overall interoperability with other related platforms.

During the entire implementation of the DMML workflow and KDD process, technical challenges did arise and caused some problems to solve to complete the final classification analysis. Some of these challenges included imbalanced raw data, feature engineering, selecting the model and overfitting.

The employee attrition selected dataset inherited class imbalance which can often lead to lower performance when modelling. Techniques such as oversampling the minority class can help with this class imbalance problem. Evaluation metrics such as F1-score can also provide us with a more accurate overall assessment of the model performance which was applied to help us with the imbalance.

Dealing with a load of different and unique features to extract useful and informative insights from them can be challenging such as salary rate and job role, etc. This means trying to decide which features to include or exclude requires general knowledge and a lot of testing and experimenting with. The solution we came up with is to try our best to understand relationships between the features through exploratory data analysis (EDA) and come up with our final target variable which in our case was “attrition”. With the use of python, we removed irrelevant features such as “Over 18” and others as they were deemed unnecessary for our classification analysis. By analyzing the correlation among features, we also got a vital amount of information regarding employee performance.

Selecting the most suitable classification models for our attrition dataset was also time-consuming and hard decision-making. Various algorithms can perform differently on imbalanced data which means we had to take extra precautions to avoid any unnecessary problems. The solution to this challenge would be by experimenting with all the different algorithms and applying ensemble methods to improve overall performance to pick out which is most suitable. As mentioned before, in our case we chose Decision Tree, Random Forest and KNN as our main three classification algorithms.

# Results and Conclusions

In this final section of our report, we will present the results of our classification analysis on the employee attrition dataset and interpret our findings to see if we have revealed what our goal was in the first place which was, “What are the leading factors of employee attrition?”. We will break the results down into meaningful subsections and present tables and figures showcasing the main highlights of our overall analysis.

### Model Performance

As our initial assignment task was to choose three classification algorithms, we have chosen the Decision tree, KNN and Random Forest models to predict employee attrition. The accuracy results shown Table 2 below, tell us that the Random Forest model achieved the highest accuracy out of all the models after hyperparameter tuning with an accuracy of 88%, which was then followed by KNN with a score of 85% and finally Decision Tree with a final accuracy score of 82%.

##### TABLE II: Model Performance

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy before**  **HPT** | **Accuracy after**  **HPT** |
| Random Forest | 87% | 88% |
| KNN | 85% | 85% |
| Decision Tree | 76% | 82% |

### Feature Importance

Using the essential aspect of feature importance analysis, we carefully studied and identified the most significant factors which contribute heavily to employee attrition. After conducting an initial correlation matrix, we easily spotted out the most correlated features such as “Years since last promotion”, “Total working years”, “Years at company” and a few others as shown in Figure 2 below which is a further conducted correlation matrix based upon only highly correlated features from that were identified from the original correlation matrix.

A screenshot of a computer screen

Description automatically generated

Fig 2: Correlation matrix on highly correlated features

### Model Interpretability and final conclusion

A graph of a bar graph

Description automatically generated

Fig 3: Histogram visualizing job roles

A graph of different types of data

Description automatically generated with medium confidence

Fig 4: Correlation of attrition analysis

In conclusion, our analysis of the employee attrition dataset revealed worthy insights into the factors that influence it. We discovered that job satisfaction with regards to how many years the person is at the certain company and job and the salary, and career advancement opportunities are all main determinants of employee attrition possibility.

If I was to do this project again, I would explore further data sources such as performance reports or employee surveys to further boost our accuracy of the models. I would also look into the initial dynamics of attrition by identifying major changes in employee behavior and ongoing performance over a period of time. If given more time, we would also definitely look at inducing advanced machine learning methods such as deep learning to also enhance our model performance.

# References

[1] M. Subramony and B. C. Holtom, ‘The Long-Term Influence of Service Employee Attrition on Customer Outcomes and Profits’, *J. Serv. Res.*, vol. 15, no. 4, pp. 460–473, Nov. 2012, doi: 10.1177/1094670512452792.

[2] A. Adhikari, ‘Factors Affecting Employee Attrition: A Multiple Regression Approach’, *IUP J. Manag. Res.*, vol. 8, no. 5, pp. 38–43, May 2009.

[3] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, ‘Predicting Employee Attrition Using Machine Learning Techniques’, *Computers*, vol. 9, no. 4, Art. no. 4, Dec. 2020, doi: 10.3390/computers9040086.

[4] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, ‘Predicting Employee Attrition Using Machine Learning Approaches’, *Appl. Sci.*, vol. 12, no. 13, Art. no. 13, Jan. 2022, doi: 10.3390/app12136424.

[5] P. K. Jain, M. Jain, and R. Pamula, ‘Explaining and predicting employees’ attrition: a machine learning approach’, *SN Appl. Sci.*, vol. 2, no. 4, p. 757, Mar. 2020, doi: 10.1007/s42452-020-2519-4.

[6] ‘Employee Attrition’. Accessed: Apr. 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/colearninglounge/employee-attrition

[7] ‘KDD Process in Data Mining’, GeeksforGeeks. Accessed: Apr. 18, 2024. [Online]. Available: https://www.geeksforgeeks.org/kdd-process-in-data-mining/

[8] Unknown, ‘BLOG DHINDA AFSARYNA AWANYS: Konsep Data MIning contoh kasus KDD’, BLOG DHINDA AFSARYNA AWANYS. Accessed: Apr. 18, 2024. [Online]. Available: https://dhindaawanys.blogspot.com/2018/04/konsep-data-mining-contoh-kasus-kdd.html

[9] Q. Collaborators, ‘Correlation Matrix: What is it, How It Works with Examples’, QuestionPro. Accessed: Apr. 23, 2024. [Online]. Available: https://www.questionpro.com/blog/correlation-matrix/