Katharina Aschbacher, BSc

The Uncertain Promise of HR Analytics: A Comprehensive Analysis of Data Utilised in Predictive Models for Employee Turnover

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme: Computational Social Systems

submitted to

Graz University of Technology

Supervisors

Supervisor: Viktoria Pammer-Schindler, Assoc.Prof.
Institute of Interactive Systems and Data Science

Co-Supervisor: Stefan Thalmann, Univ.-Prof.
Institute of Operations and Information Systems

Graz, September 2024

Abstract

The application of artificial intelligence (AI) in the field of human resources (HR) is becoming increasingly common. AI-based forecasting models enable the accurate prediction of employee turnover, facilitating the implementation of employee retention measures in a timely manner. Given the paucity of information regarding the data employed in these predictions, the objective of this thesis was to analyse the data utilised for the purpose of forecasting employee turnover. The initial section of the thesis comprises a literature review, which was conducted in order to analyse relevant studies. The analysis revealed that a significant proportion of these studies were based on the same data sets. A comprehensive examination of 30 studies revealed the use of 500 variables for predicting employee turnover. Following the removal of duplicates, the number was reduced to 286 unique variables. One challenge in the automated prediction of employee turnover is that a considerable number of identified variables, such as those related to employee satisfaction or the organisational climate, cannot be collected automatically. These variables are often critical for prediction; however, their collection frequently requires processes like surveys and interviews, making their integration into automated systems challenging. The second part of the thesis investigated the impact of reducing a data set using a mixed-methods approach to feature selection on the performance of machine learning algorithms. The findings indicated that a careful selection of relevant features enhanced the performance of predictions. By tuning the algorithms, these results were further optimised, particularly in terms of recall, which is of paramount importance for the early identification of potential leavers.

Kurzfassung

Der Einsatz künstlicher Intelligenz (KI) im Bereich der Human Resources (HR) gewinnt zunehmend an Bedeutung. Mithilfe von KI-basierten Prognosemodellen lässt sich die Mitarbeiterfluktuation präzise vorhersagen, sodass rechtzeitig Maßnahmen zur Mitarbeiterbindung initiiert werden können. Da bislang wenig darüber bekannt ist, welche Daten für diese Vorhersagen verwendet werden, war es das Ziel dieser Arbeit die Daten, die zur Vorhersage der Fluktuation eingesetzt werden, zu analysieren. Im ersten Teil der Arbeit erfolgte eine Analyse von Studien im Rahmen eines Literature Reviews. Die Analyse der untersuchten Studien ergab, dass eine Vielzahl von ihnen auf denselben Datensätzen basiert. Im Rahmen der detaillierten Analyse von 30 Studien konnten insgesamt 500 Variablen identifiziert werden, die zur Vorhersage von Mitarbeiterfluktuation genutzt werden. Nach der Entfernung von Duplikaten reduzierten sich die Anzahl der ursprünglich 500 Variablen auf 286 distinkte Variablen. Ein Hindernis bei der automatisierten Vorhersage von Fluktuation besteht darin, dass ein erheblicher Teil der identifizierten Variablen, wie Mitarbeiterzufriedenheit oder Variablen, die das organisationale Klima betreffen, nicht automatisch erhoben werden können. Diese Variablen sind oft entscheidend für die Vorhersage, jedoch ist ihre Erhebung mit Prozessen wie Umfragen und Gesprächen verbunden, was ihre Nutzung in automatisierten Systemen erschwert. Im zweiten Teil der Arbeit wurde untersucht, inwiefern sich die Reduktion eines Datensatzes mittels eines Mixed-Methods-Ansatzes der Feature Selection auf die Performance von Machine Learning Algorithmen auswirkt. Die Ergebnisse belegen, dass eine gezielte Auswahl relevanter Variablen eine Verbesserung der Vorhersageleistung bewirkt. Durch Tuning der Algorithmen konnten diese Ergebnisse weiter optimiert werden, insbesondere der Recall, welcher für die frühzeitige Identifizierung potenzieller Kündigungen maßgeblich ist.

Acknowledgements

First and foremost, I would like to thank my supervisors, Stefan Thalmann and Viktoria Pammer-Schindler, for their continued support, feedback, and valuable advice throughout the completion of this thesis. Their expertise and guidance have been invaluable to my work, and I am sincerely grateful for their supervision.

I also wish to thank Ilona Ilvonen from the University of Tampere, who, despite not being affiliated with our institution, co-supervised my thesis. Her timely feedback, willingness to answer my questions, and years of expertise have been immensely helpful in completing this research.

Lastly, I extend my thanks to my family, my boyfriend, and my friends, whose constant support and encouragement have been incredibly helpful not only during the writing of this thesis but throughout my entire academic journey.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Artificial Intelligence (AI) has become a tool that has found application in a wide range of areas and disciplines, including human resource management (HRM). While AI was originally used to analyse large amounts of data of machines or sensors, it is now also being used in various Human Resource (HR) fields (Garg et al., 2022). It is used in a number of areas, for example, in recruitment or personnel selection, but also in performance management and employee retention (Chowdhury et al., 2023).

Another potential field of application in the area of employee retention could be the prediction of employee turnover. Predicting turnover before employees have left a company, offers companies and organisations the chance of securing talent and expertise. By developing accurate predictive models, potential departures can be detected early and effective retention measures can be taken (Punnoose & Ajit, 2016).

Attracting talented employees, on the one hand, and retaining talent on the other is one of the biggest challenges for companies worldwide (Dobbs et al., 2012; Ranjan & Yadav, 2018). Katz and Kahn (2015) regard attracting and retaining skilful and capable staff as one of the primary goals of HR in general. In the last few years, these two goals have become increasingly important in HRM since organisations and companies compete globally for talented employees and their skills (Stone et al., 2015). However, attracting and retaining talented employees is exacerbated by the fact that there is a shortage of talented workers. The World Economic Forum, for example, considers it certain that there will be a talent gap in most industrialised nations as soon as the baby boomer generation retires (Wójcik, 2017).

Attracting the most skilled and talented in a certain field may come to represent a clear competitive advantage. Retaining trained employees is crucial, as the loss of in-house knowledge that may accompany the departure of workers makes it necessary to preserve knowledge and then transfer it to new workers (Loebbecke et al., 2016; Thalmann & Ilvonen, 2020) . Another negative consequence of employee turnover is that organisations not only lose knowledge, but also incur costs for recruiting and training new employees (Sutherland, 2002). As outlined by Davenport et al. (2010), leading companies are employing analytical techniques to examine personnel data with the objective of enhancing their competitive advantage. This

shows that companies are realising the importance of talented employees. A study by Lee et al. (2018), indicates that in 2016, 46% of human resources managers identified employee turnover as their primary concern.

Nevertheless, the majority of companies have not yet implemented models for prediction purposes. Tambe et al. (2019), for example, find that only 22% of companies employ HR analytics methods in their HR departments.

This underlines the relative youth of the discipline discussed and the corresponding lack of research in this area, a concern that will be discussed in the next section.

1.2 Problem Statement and Research Questions

Despite the wide range of applications of AI in HR, the evaluation of its impact and benefits in this area remains complex and provides contradictory results (Johnson & Kuhn, 2024). While some studies emphasise the benefits and opportunities of AI (Punnoose & Ajit, 2016; Ranjan & Yadav, 2018), others are more critical of the cost-effectiveness and quality of the results (Tambe et al., 2019). The heterogeneity of these results emphasises the importance of a more detailed and critical examination of AI in the field of human resources.

Johnson and Kuhn (2024) see a possible reason for this discrepancy in the treatment of AI in HR as a 'black box'. This approach means that the literature and research in this area largely focuses on the mere presence or absence of IT systems, such as AI, and does not take a closer look at the exact way they function. As a result, internal processes of such systems remain hidden and the focus largely shifts to the output and results of such systems rather than the functioning and realisation of results (Johnson et al., 2016). Therefore, analysing and examining the data used for predictive applications is an important aspect for a better and deeper understanding of AI in HR in general. At this point, it must be mentioned that the sensitivity of HR data leads to limited availability or even unavailability of such data (Chowdhury et al., 2023; Johnson & Kuhn, 2024) and the analysis of the data is therefore often associated with difficulties.

Although there are isolated, fragmentary approaches to analysing data used for predictive purposes in HR, there is a lack of systematic analysis of such data. This lack of systematic investigation can be attributed to various reasons, such as the discipline in general being rel-

atively young and the research on it suffering from fragmentation, as Pan and Froese (2023) suggest.

Nonetheless, this research gap hinders the much-needed deeper understanding of AI in HR, especially the understanding of the exact functioning of predictive forecasting models. These models are essential for retaining employees and thus securing a competitive advantage in the battle for talent within companies and organisations. Given this context, this thesis aims to analyse which types of data are used for predicting employee turnover, which can be regarded as one specific example of an AI application in HR. Furthermore, it wishes to contribute to a differentiated understanding of AI in Human Resources and to promote the scientific discourse in this field. Therefore, one objective of this work is to identify possible predictors and data sources for accurate turnover prediction via a literature review, leading to the following research question:

RQ: "Which types of data are used in predictive models for employee turnover within HR analytics?"

Furthermore, there is a lack of well-founded research to justify the selection of variables in turnover prediction models (Yahia et al., 2021). Identifying the most significant predictors is of great importance though, as selecting a meaningful set of features enhances predictive performance (Kuhn & Johnson, 2019). This underscores the necessity of thoroughly analyzing and understanding the data used for predictions. Therefore, after identifying the data types used for prediction, the next step is to determine which are the most important for predictive purposes. For example, characteristics such as salary, job satisfaction, time spent in the organisation and demographic information may appear as important predictors of employee turnover. With a limitation to a key set of important features, this work aims to replicate existing machine learning models described in the literature by using only those selected features. The objective of this is to assess whether a simplified model that is limited to the most important prediction data can achieve similar results to models that use a complete data set, which leads to the following working question:

RQ2: How does the use of a reduced dataset with feature selection influence predictive performance, considering the challenges of understanding AI as a "black box" in HR Analytics?

1.3 Overview

The present introduction provides the motivation for the topic, the research question and the related working question as well as the present overview. In the theoretical background section, important concepts and topics such as HR Analytics, Employee Turnover, Knowledge Management, Machine Learning models and the GDPR are explained. The subsequent methodology section includes the procedure of the literature review and the modelling. In the Results findings about data used and the results about the modelling will be discussed. Finally, the thesis will conclude by discussing the theoretical and practical implications of the findings, the interpretation of data under the GDPR, the limitations of this study, and potential directions for future research. A concluding summary will be provided at the end.

2 Theoretical Background

2.1 The Field of HR Analytics

Overview

In recent years, technology and its continuous advancements have transformed the field of HR (Stone et al., 2015). HR Analytics has been defined in different ways with an emphasis on different aspects of the field. Lawler et al. (2004) define HR Analytics as the process of using statistical techniques to link HR practices with the performance of an organization and by this they underscore the connection of HR practices and organizational performance. In contrast, Fernandez and Gallardo-Gallardo (2021) focus more strongly on the data dimension and define HR Analytics as the integration of internal or external HR data to support decision making in organizations. Marler and Boudreau (2017), on the other hand, offer a more comprehensive and inclusive definition and describe HR Analytics as "A HR practice enabled by information technology that uses descriptive, visual, and statistical analyses of data related to HR processes, human capital, organizational performance, and external economic benchmarks to establish business impact and enable data-driven decision-making." While there exist many different definitions of the topic (Marler & Boudreau, 2017), it is evident that the use of data analytics represents a promising possibility for organizations seeking to make strategic decisions and to further optimize their operations.

Regarding the history and the development of HR Analytics Marler and Boudreau (2017) mention that the first measurement in HR occurred in the early 1900s and the first publication of a book on HRM measurement in 1984 (Fitz-enz, 1984). This aligns with the view of Van den Heuvel and Bondarouk (2017) and Bondarouk et al. (2009), who see the beginnings of HR Analytics in the 1980s, when researchers noticed an early use of technology for the Automation of different HRM tasks. In the 1990s, the use of Human Resource Information Systems (HRIS) grew rapidly, mainly in the reduction of administrative tasks, while the 2000s was the time when electronic HRM applications went beyond administrative tasks and towards more strategic areas such as performance management and talent acquisition (Bondarouk et al., 2009; Van den Heuvel & Bondarouk, 2017).

Today, as organizations increasingly use Artificial Intelligence (AI) and Machine Learning

(ML) technologies (Garg et al., 2022), the role of data in HR has become even more important. These technologies can be considered efficient means for analyzing large amounts of HR data, ranging from employee performance metrics to talent acquisition applications.

Data for HR Analytics

Different data sources are used depending on the application. According to Mohammed (2019), a typical HR analytics system uses employee data, business performance records and social media data. These sources are brought together in a data warehouse, where data mining or statistical approaches are applied to discover patterns or make predictions. Falletta and Combs (2020) distinguish between public and private data sources; public data can be found in government databases, for example, and private data includes an organization's own private employee data.

A closer look at the specific application areas of HR analytics reveals different sources and approaches. In recruitment, for example, objective criteria can be extracted from a candidate's Linkedin profile (Faliagka et al., 2012) or CVs can be automatically screened (Ore & Sposato, 2022).

In the area of performance management, employee performance evaluation records are being used as a prediction tool for current employees (Jantan et al., 2010), but even the prediction of the performance of job applicants is possible (Augusto et al., 2013). In the area of employee retention, the prediction of employee turnover makes it possible to detect employees whom an organization risks losing. In this case measures can be taken to retain talent more efficiently. And again, various data sources are being used. El-Rayes et al. (2020), for example, use anonymous resumes from Glassdoor in addition to public company reviews to predict an employee's likelihood to leave a company. Similarly, employee data can be acquired through surveys (Yahia et al., 2021) or by datasets provided by HR departments of a company (Alamsyah & Salma, 2018).

However, the availability and accessibility of data are important factors to consider, when addressing data in HR Analytics.

Public data sources, such as government databases, information available on social media platforms like LinkedIn and information on data science platforms like Kaggle are generally accessible and allow users to obtain their data easily. On the other hand, private data sources,

such as internal HR records from companies, employee performance evaluations and feedback surveys are often only available to internal staff and is protected by privacy regulations. For example, in Europe the General Data Protection Regulation (GDPR) ensures that the use of employee data complies with laws. Moreover, the GDPR also ensures that privacy and ethical considerations are met. In Germany, since the introduction of the GDPR in 2018, companies face significant challenges in balancing data protection laws with developing AI applications for HR (Groß, 2021). Companies must ensure they have the permission to collect and process this data, which is usually done through employee consent. Another challenge for AI applications in HRM under GDPR is related to its strict rules for data processing, such as "purpose limitation," meaning that data cannot be collected without a specific reason, which is often difficult especially in the area of HR (Groß, 2021). HR managers first need to define clear objectives for their AI applications to ensure they have a valid purpose, which is difficult to realise in practice (Groß, 2021). Protecting the people behind the data, can also restrict or complicate the usage of data for analytical purposes. Zieglmeier et al. (2022),for example, recognise the sensibility of employee data, but also emphasise the importance of employee data for people analytics, and therefore identify appeal strategies for voluntary sharing of employee data. This is a new approach, as in the past strategies have tended to focus on data awareness. Since people analytics bring benefits for both employees and employers, Zieglmeier et al. (2022) claim that there exists a need for appeal strategies to be applied.

Despite the various strategies and regulations in place several challenges still need to be addressed before a universal adoption is possible. One of the biggest challenges is the availability and adequacy of data (Garg et al., 2022). Machine Learning models perform best with large and high-quality data, which is often limited in the HR sector. Security concerns are also cited as a challenge (Garg et al., 2022). There often remains ambiguity about how long such data is stored before being deleted, whether it is deleted at all and who has access to it. Another major point is that of the fairness and transparency of AI (Garg et al., 2022). These aspects are of the utmost importance to guarantee ethical and unbiased decision making processes. However, these are precisely the points that are often not sufficiently addressed.

Overview of Applications

Figure 1: HR Lifecycle Process. Adapted from *(Bentum, 2023; Qualtrics, 2022)*

According to Garg et al. (2022) ML is used in HRM in a wide variety of areas. For example, in recruitment, where the suitability of candidates can be used by analyzing profiles and CVs, or in personnel selection, where models can support the selection process effectively. Addressing already employed staff, developed models can help to determine training needs, but also evaluate performance and engagement and thus create suitable incentives (Garg et al., 2022).

Particularly important for this work, however, is the area of application in the field of predicting employee turnover, which is also mentioned as one of the key application fields by Garg et al. (2022).

Research on ML models in HRM is most pronounced in the areas of performance management and recruitment. Research on employee turnover is moderately advanced and can be placed in the middle of the field in terms of research progression (Garg et al., 2022).

To fully understand how intelligent solutions can be used in HR Analytics, it is important to first look at the broader HR processes to which predictions are applied. Seen from a broader perspective employee turnover is linked to different stages of the employee life cycle. Bentum (2023) distinguishes five key phases of the employee life cycle: procurement, onboarding, development, retention and exit. Figure 1 shows a visualisation of these phases.

The procurement phase includes employer branding and recruitment. Effective employer branding involves positioning a company as an attractive workplace (Bentum, 2023). For example, a positive employer image, which is formed by media, successes, products, and personal experiences, plays a major role in the selection of a potential employer (Bentum, 2023). Successful branding aids in recruitment as well as in long-term retention (Backhaus & Tikoo, 2004), underlining the importance of the procurement phase.

Storey et al. (2019) considers recruitment and selection as the the most important HR functions, especially for knowledge-based companies. Since competition for skilled personnel

intensifies (Chambers et al., 1998), it is necessary for companies to actively promote job offers through job portals, career pages, and social media (Bentum, 2023). Here a shift towards modern aid can be seen as well: While traditional recruitment events and consultants can be expensive, electronic job advertisements are cost-effective, making them an attractive alternative.

Following recruitment, the onboarding process lays the foundation for long-term retention and employment. This phase involves both technical introductions and social integration, with mentoring and regular manager meetings playing an important role (Bentum, 2023).

The development phase focuses on talent development. Strategic training can have a positive influence on company performance (Storey et al., 2019). Bentum (2023) adds that knowledge transfer through training, peer learning, e-learning, project work, and coaching are all effective methods of employee development.

Retention is a critical phase following successful recruitment and development, where predictive models can become especially useful. These models can predict an employee's intention to leave, enabling companies to set countermeasures. This thesis will explore these predictive models in more detail later.

The understanding of these HR phases is of utmost importance for effectively applying modern technologies in HRM, as they provide the context within which predictive models operate, particularly in the area of employee turnover.

2.2 Employee Turnover

Definition

Employee turnover, the withdrawal of employees from their company or organization (Mobley et al., 1979), poses considerable challenges to companies. Research by Lee et al. (2018) shows that in 2016 for 46% of HR Managers employee turnover is their most substantial concern. Although there is agreement on the negative effects of high turnover, there is disagreement on the definition of employee turnover.

First of all, there is dissent about the distinction of voluntary and involuntary turnover. In principle, voluntary turnover at this point is considered the leaving of employees on a voluntary basis by the termination of their contract, whereas involuntary turnover occurs when

the organization terminates the employment of an individual. Even though this distinction is often made in literature, there is often no consensus on which definition is used in different studies.

Abbasi and Hollman (2000), for example, understand this to mean not only the rotation of employees between different jobs, but also the status between work and unemployment. Chakraborty et al. (2021) go one step further and use both the departure of employees and the retirement of employees as turnover in their definition. Also, the literature is divided as to whether a factor such as pregnancy is part of voluntary or involuntary turnover (Mobley et al., 1979). Additionally, some authors use the term turnover to refer to both voluntary and involuntary employee turnover without clarifying whether they refer to only voluntary or also involuntary departures (Shaw et al., 1998). Shaw et al. (1998), summing up some of these difficulties, criticize the lack of differentiation in the literature in general.

In this work, the definition of voluntary turnover is used, as research on predicting turnover used as a competitive advantage, also refers to this definition.

Causes

The causes of employee turnover are manifold and various factors have been analyzed over time.

Ongori (2007) distinguishes between job-related factors and organizational factors that cause employees to leave a company. Job-related factors include, for example, stress and job dissatisfaction as well as economic reasons. Organizational reasons, on the other hand, can include instability and inefficiency in organizations and a lack of communication from management to employees (Ongori, 2007). Porter and Steers (1973) regard job satisfaction as the crucial aspect influencing employees to terminate their employment relationship or not. Job satisfaction at this point is conceptualized by them as the total sum of the met expectations of an individual and the job. Within this framework they then distinguish a number of different factors directly and indirectly influencing turnover. On an organizational level compensation, i.e. pay and promotion, seem to be a significant determinant of turnover intention (Porter & Steers, 1973). The analysis of job security as an influencing factor on the other hand provides conflicting results whereas the organizational size doesn't seem to be a factor at all (Porter & Steers, 1973). Regarding work environment factors, it was found

that the supervisory style, the size of the working unit and coworker satisfaction have an influence on turnover intentions of employees. Also job content factors seem to influence turnover, where a strong positive relationship between job autonomy and turnover has been found (Porter & Steers, 1973). Lastly, also factors focusing on the individual himself, like age or tenure, turned out to have an effect on employee turnover.

This groundbreaking analysis carried out in 1973 has still a great influence on today's research. Many studies investigated refer to the factors described by Porter and Steers (1973). However, there is still no universal framework that describes and explains the process; thus reasons and sources must be analyzed in detail.

Consequences

Employee Turnover can have far-reaching consequences for organizations, ranging from separation costs to affecting operational planning. High turnover of employees may lead to incremented staffing costs, as companies need to hire and train new employees (O'Connell & Kung, 2007; Tziner & Birati, 1996). This does not only lead to direct costs associated with the recruitment and onboarding process, but also leads to reduced productivity during the period the vacancy left by the departed employee remains unfilled. Another fact to consider is the potential effect on the morale of the remaining employees (Tziner & Birati, 1996). Employees might see an organization with high turnover as unstable and thus unattractive (Tziner & Birati, 1996).

A similar view is taken by Krausz et al. (1999), who have found that a coworker's departure triggers also psychological evaluations among remaining employees and may increase their own intention to leave. This phenomenon can be attributed to a disruption in the established routine and social life at the office. Additionally, it prompts the consideration of alternative job opportunities (Krausz et al., 1999).

Furthermore, the departure of skilled employees can disrupt ongoing projects and reduce an organization's overall efficiency, as new employees require time to reach the same level of knowledge as their predecessors (Hall & Moss, 1998). What is more, the loss of personnel can affect the organization's competitive advantage (Yamin, 2020) since the generation and transfer of knowledge represent the foundation for competitive advantages in companies (Argote & Ingram, 2000).

Prediction of Turnover

Various researchers have aimed to find the most influential factors for employee turnover. Lee et al. (2018) summarized key findings from past theoretical and empirical works such as job satisfaction, organizational commitment, job embeddedness, shocks, willingness to stay and control over the staying decision. Moreover, they found that what is deemed important by employees and consequently what influences turnover decisions has changed from the 20th to the 21st century, meaning that Millennials have a stronger focus on job purpose and personal development, than the generation before.

Using a more empirical approach, many other works have dealt with the prediction of employee turnover with Machine Learning Algorithms. Sato et al. (2019) investigated the prediction of employee churn in the restaurant industry by using anonymous data from a Japanese restaurant chain, which consisted of employees' attendance records and contained features such as age, gender or hiring channel. They found that the hiring channel and the gender had little influence.

Wang and Zhi (2021), on the other hand, also focused on challenges with the prediction of turnover and mention different human factors, the lack of training data, the imbalance of many datasets and the number of models to choose from. Their answer to the aforementioned problems was the development of a framework for a streamlined approach, while using two different publicly available datasets for their prediction. These datasets consisted of variables recurring throughout literature, variables such as the number of projects, working hours, salary, gender, age, distance from home. They found that the most important predictors were salary, age, satisfaction, the time spent at the company and the number of projects.

Zhao et al. (2019) also used two different datasets for their predictions, both a real and an artificial one. Since previous studies mainly focused on single and small datasets, they also considered the factor of size of a dataset. Hence, they use subsets of small medium and large datasets from the original datasets. Similar to Wang they used the IBM Watson dataset and got similar results regarding the importance and influence of the variables there. Variables such as ethnicity, specialized area and team, which aren't found in the publicly available datasets mentioned before, were used by their real-world bank dataset.

Contrary to Sato et al. (2019), Zhu et al. (2019), using data from Chinas biggest professional social platform, found that gender plays a key role in turnover behaviour. Additionally, they found significant effects for education and industry.

Industry focused research, like Nahar et al. (2022) work on the garment industry in Bangladesh revealed additional variables like safety satisfaction or stress level, specific to their country since Bangladesh has less worker protection rights in comparison with other countries (Al-Amin et al., 2020), making safety and stress relevant country specific predictors.

Another study which used a dataset of a communications company in China (Gao et al., 2019), found the key factors to be monthly income, overtime, age, distance from home, years at the company and percent of salary increase. Among the studies investigated, this study was the only one that included the physical condition of people as a feature. The absence of this feature in other studies might be explained by discrimination concerns and the lack of significant influence it had as a predictor according to the authors.

El-Rayes et al. (2020) used anonymous resumes from Glassdoor for their predictions together with publicly available company information. They found salary and public image to be the most important factors.

Contrary to many other papers, Yahia et al. (2021) additionally to the commonly found important features such as salary and rewards found business travel to be an important factor.

Despite the extensive range of applications of AI in HR, the assessment of their effectiveness yields inconsistent results (Johnson & Kuhn, 2024). Punnoose and Ajit (2016), for instance, emphasize the potential of AI in HR, whereas Tambe et al. (2019) are more critical of its effectiveness. Tambe et al. (2019) claim that the effective application of AI to HR problems presents different challenges compared to other areas. These challenges include algorithmic bias (Dastin, 2022; Tambe et al., 2019), the complexity of the outputs (Tambe et al., 2019), a small sample size (Johnson & Kuhn, 2024; Tambe et al., 2019), the youth of the discipline (Budhwar et al., 2022), and the black box character of AI (Johnson et al., 2016).

First, algorithmic bias is a significant issue. A notable example is Amazon's recruiting tool, which performed automated resume screening but was biased against women because it had been trained predominantly on male data (Dastin, 2022).

Second, the complexity of outputs is a concern. Topics involving human factors are inherently challenging to measure. For instance, evaluating job performance is difficult due

to potential human biases, which complicates the training of algorithms (Tambe et al., 2019).

Third, dataset size affects prediction results. Research indicates that small or insufficient datasets, especially in small and medium-sized companies, can impact the effectiveness of AI (Alzubaidi et al., 2023; Zhao et al., 2019). Additionally, the relative youth of the discipline contributes to inconsistent outcomes due to the absence of an established framework of methodologies (Johnson & Kuhn, 2024).

Lastly, the black box nature of AI must be considered. Johnson et al. (2016) view the mere presence or absence of an Information Technology System in HR as a critical factor influencing organizational success. Johnson and Kuhn (2024) argue that treating IT in HR as a black box—focusing solely on output without understanding internal processes—hinders effective utilization. Many studies discuss the effectiveness of AI in HR; for example, Madanchian et al. (2023) highlight the need for comprehensive strategies to use AI efficiently in HR. Similarly, Chowdhury et al. (2023) stress the importance of integrating human skills and knowledge with AI for effective results. Bondarouk et al. (2009) go further, suggesting that the presence or absence of a technology-based HR approach impacts employee perceptions, though its implementation did not improve these perceptions.

While some studies are contributing to the problematic described by Johnson and Kuhn (2024) and focus on the output rather than a description of feature selection or source of the data (Mittal et al., 2023; Zhang et al., 2018), others come to inconsistent findings regarding the influence of the features themselves (Sato et al., 2019; Zhu et al., 2019). These inconsistencies and the lack of a systematic analysis underline the need for further exploration in this area. Given these circumstances, this thesis aims to address the following research question:

RQ: Which types of data are used in predictive models for employee turnover within HR Analytics?

By a systematic analysis of the data types and predictors used in existing literature, this work aims to contribute to further research and knowledge of AI applications in HR and to foster an approach where the focus is not merely put on the output, but the understanding of the data used for making predictions.

2.3 The Synergy of Talent Retention and Knowledge Management

Knowledge in Organizational Success

Employee Retention in Organisational Success and Knowledge Management Davenport and Prusak (1998) define knowledge as "a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information." In their work about organisational and working knowledge, they state that it does not only become inserted into documents or databases, but also in routines, norms and practices of organisations. Moreover, knowledge can be a competitive advantage (Davenport & Prusak, 1998). Technology and products can be matched by competitors, whereas knowledge, as an intangible asset, can be used to advance new ideas of creativity or efficiency (Davenport & Prusak, 1998). Romer (1992), a leading figure in knowledge economics, argues that ideas, which can be seen as one source of knowledge, have the potential to sustained growth.

Upon this, it becomes evident that employees are significant contributors to an organization's success (Abbasi & Hollman, 2000). Their skills, knowledge and creativity drive innovation and entrepreneurial activity (Amabile, 1996). For example, when developing new and innovative products, the creativity and skills of employees are crucial (Amabile, 1996) and through this it is possible to gain a competitive advantage. Similarly, skills and expertise of employees in the field of service industry can influence customer experience and satisfaction (Hennig-Thurau, 2004).

Long-term viability and performance of an organisation depend on its ability to retain its most important employees (Das & Baruah, 2013). Retaining talented employees ensures continuity in services and in organisations in general (Venkat et al., 2023), which is vital for achieving long term business goals (Foster & Dye, 2005). Moreover, the capacity to retain the best personnel has a significant impact on a variety of organizational factors such as customer happiness, higher sales, satisfied coworkers and reporting staff, efficient succession planning, etc. (Das & Baruah, 2013) .

Recruiting and training a replacement for a departed employee costs approximately 50% of the employee's annual salary (Johnson et al., 2000). These estimated numbers can be explained through a combination of direct costs such as costs for the advertisement of the

job, the hiring process itself and the training afterwards, and indirect costs such as the lost productivity until a new employee has been found.

As highlighted before, retaining skilled employees is of great value. Nevertheless, organizations often don't realise that an employee with critical knowledge is about to leave (Hana & Lucie, 2011). Also, as soon as an employee has left it is nearly impossible to quantify the loss of their knowledge and expertise (Hana & Lucie, 2011). This loss of knowledge has especially far-reaching consequences if the departed employee had specialised skills or internal company knowledge, which is not easily replaceable.

According to Stovel and Bontis (2002), managers of Canadian financial firms do recognize the negative effects of turnover, but interestingly weren't concerned about the implications of losing employees and valuable knowledge to their competitors. This paradox underlines that even though there is awareness of the costs and problems associated with employee turnover, there is still a lack of strategic knowledge retention strategies and prevention measures against knowledge spillovers to competitors.

The War for Talent

According to Schweyer (2004), talent retention is the process used to attract, identify, recruit, develop, motivate, promote and retain employees who have the knowledge and skills necessary to fulfil the demands of a organisations and to succeed within it. As stated by McDonnell (2011) talent management remains one of the most outstanding challenges for organisations. Work force demographics and staff shortages make the fight for talent and an effective talent management strategy a competitive necessity (McDonnell, 2011). Kwon and Jang (2022) go even one step further and claim that workforce differentiation, i.e. investing disproportionate resources in employees where disproportionate returns are expected (Gelens et al., 2013), is the key to talent management. This postulation can be explained through Pareto's principle, which applied to the business area, indicates that 20% of employees are responsible for 80% of an organisation's value (Kwon & Jang, 2022; Swailes, 2013).

Already in 1998, when McKinsey & Company published the book "The War for Talent", the authors Chambers et al. (1998) claimed that talent retention is one of the most crucial factors for a competitive advantage and that holding managers responsible for their spending is just as important as holding them responsible for growing their talent pool. Moreover,

they state that already in 1998, many companies were suffering a shortage of talent. Twenty years on, organisations and companies are still facing the same issue - as evidenced by the World Economic Forum it is statistically certain that when the demographic boom generation retires, there will be a considerable talent gap in most industrialised nations (Wójcik, 2017).

Chambers et al. (1998) name three main challenges which further exacerbate the problems described above. First, the more complex the economy gets, the more sophisticated and skilled talent is required. Second, the emergence of many small and medium sized companies which target the same people, intensifies the fight for talent further. And lastly, job mobility, especially among young employees, is increasing. Similarly, PriceWaterhouseCoopers (2008)'s Annual Global Survey showed that 89% of the interviewed CEOs put talent management as one of their top priorities.

These numbers and facts underline the need for effective talent management and for effective employee retention strategies, since the fight for talent involves keeping talented employees once they have been recruited. According to Davenport et al. (2010), leading businesses are progressively implementing techniques for analysing personnel data in order to enhance their competitive advantage. Companies like Google, Best Buy, Sysco, and others are starting to replicate their accomplishments by learning exactly how to guarantee the highest levels of productivity, engagement, and talent retention (Davenport et al., 2010). At the same time, Ibidunni et al. (2016) have found that job satisfaction is key to talent retention and commitment. Thus, not only advanced analysis techniques as mentioned by Davenport et al. (2010) but also incorporating a satisfying work environment can increase employees' commitment, leading them to stay in the organisation.

Those findings align with the general consensus achieved by employee turnover prediction. Various studies which researched employee prediction models (Bandyopadhyay & Jadhav, 2021; Wang & Zhi, 2021; Yahia et al., 2021; Yuan, 2021; Zhao et al., 2019) came to the conclusion that job satisfaction is one of the contributing factors in predicting employee turnover ahead of time and conclusively in retaining talented employees. Thus, predicting employee turnover is a HR application with significant implications for the fight for talent. If organisations want to retain their most talented employees, predicting turnover helps them to tackle potential issues ahead of time, implement retention strategies, and thereby maintain

a competitive advantage against competitors that refrain from doing so.

2.4 Machine Learning for Employee Prediction

Applications and Efficacy of Algorithms in Predicting Employee Turnover

In recent years, machine learning has revolutionised predictions in the field of HRM, such as, specifically important for this work, predictions in the field of employee turnover. This change can be attributed to various factors, such as the increasing availability of large data sets (Chen et al., 2014) and the superior performance of ML algorithms, especially when processing complex and non-linear relationships in data (Khanzode & Sarode, 2020).

Prior to the widespread adoption of machine learning, organisations relied primarily on traditional statistical methods such as ordinary least-squares regression, logistic regression and hierarchical regression to predict employee turnover. These methods were popular because they were relatively easy to use and interpret. For example, Vecchio and Norris (1996) investigated whether employee turnover could be predicted by satisfaction and performance, using hierarchical regression to see if a combination of the two factors could predict turnover better than the respective factors individually. Similarly, Morrow et al. (1999) used logistic regression to examine a possible interaction between absenteeism and performance in terms of turnover. And yet others used organisational commitment and job involvement to predict turnover using ordinary least-squares regression (Huselid & Day, 1991). It can thus be stated that the conceptual approaches and factors employed to predict turnover have remained largely consistent over time.

However, there has been a notable shift in the methodology used to make these predictions. During the 1990s, statistical models were the predominant approach whereas in the the late 2000s and early 2010s, there was a shift towards machine learning models. The reasons and causes for this shift are explained in the following section. Machine Learning addresses many of the shortcomings of traditional statistical models. First, compared to traditional statistical models, ML models can handle large datasets and at the same time provide accurate and effective predictions. Second, ML models perform well when datasets become complex and consist of numerous factors (Khanzode & Sarode, 2020). This is particularly advantageous in the field of human resources management (HRM), where datasets

are often composed of a large number of different factors. ML models can analyse diverse inputs well and predict turnover accurately, compared to traditional methods. One of the key advantages of ML for turnover prediction is its ability to analyse non-linear and high-dimensional data. For example, ML algorithms like Random Forest and Support Vector Machines have been found to perform exceptionally well in their prediction scenario (Yuan, 2021). Zhao et al. (2019), on the other hand, state that gradient boosting trees and extreme gradient boosting perform better than other algorithms in the prediction of turnover. While several others conclude that RF work best (Aggarwal et al., 2022; Alamsyah & Salma, 2018; Marvin et al., 2021; Singh & Thakral, 2023). These findings underline that the usage and performance of algorithms is context- and dataset dependent, even if the prediction scenario, in this case the prediction of employee turnover, is the same.

By accurate turnover predictions companies expect the anticipation of employee departures in advance, which allows them to implement counter measures to retain valuable talent or to secure knowledge, before it gets lost with a departing employee. Moreover, time savings are often cited as an advantage of the usage of ML technologies in HR. For example, by implementing ML models it is possible to quickly assess thousands of job applications and shortlist suitable candidates (Mallick, 2021). Machine learning also helps HR staff to devote more time and resources to important tasks which require human interactions cannot be replaced by machines, such as working on strategic projects. At the same time repetitive, mundane, yet important tasks can be taken over by machines (Mallick, 2021). Despite frequent claims about the benefits of intelligent decision making in HR, only a small number of organisations have actually implemented big data practices - only 22% of companies state that they use analytics in their HR departments (Tambe et al., 2019). This highlights the potential but also the gap in the adoption of ML in business contexts.

Data Perspective

Those inconsistent findings can be partly explained through the usage of different datasets. Zhao et al. (2019), for instance, have stated that bigger and medium sized datasets perform better than small sized ones. By contrast, Yahia et al. (2021) propose an approach with a focus on quality rather than quantity. A frequently used dataset for the prediction of employee turnover is a dataset from IBM Watson, which contains 34 input features including

common HR features such as education related variables, satisfaction related variables and demographic data. However, there is no in-depth and substantiating study or justification on the selection of those variables (Yahia et al., 2021). Yahia et al. (2021) therefore proposed a thorough feature selection process in their study and yielded very good prediction results with a reduced dataset of eleven features. Similarly, Mozaffari et al. (2023) recognised the importance of a qualitative approach in feature selection and thus followed an approach where only the 16 most important features of the original dataset were selected, again with good results for their predictions in the pharmaceutical industry. This underlines that a thorough approach is important and that more data or features do not necessarily lead to better results (Chandrashekar & Sahin, 2014). Kuhn and Johnson (2019) also stress that finding a meaningful set of features increases the prediction performance significantly. In an even earlier study by Kuhn, Johnson, et al. (2013), it was found that the removal of non-informative predictors through feature selection is beneficial for many algorithms. This highlights that the dealing and understanding of data used for predictions is of utmost importance.

Even though there are automated feature selection methods, they often lack user knowledge which can be beneficial (Wu et al., 2022). In general, incorporating human knowledge such as domain knowledge or the understanding of data into modelling processes ("human-in the loop") leads to better performance (Wu et al., 2022).

In summary, it can be stated that the benefits of understanding the data used for prediction tasks, are evident. Therefore, the research question of this thesis is:

RQ2: How does the use of a reduced dataset with feature selection influence predictive performance, considering the challenges of understanding AI as a "black box" in HR Analytics?

2.5 Privacy Perspective within the General Data Protection Regulation

Overview

The General Data Protection Regulation, hereinafter referred to as GDPR, is a data protection regulation established by the European Union, governing the processing of personal

data. Its primary objective is to protect the rights of natural persons regarding the processing of their personal data or information. The GDPR follows a set of fundamental principles, namely that personal data must be processed lawfully, fairly and transparently and collected for a specific, legitimate purpose. Further- more, it must be adequate and limited to the minimum necessary, accurate, stored only for the duration necessary for its intended purposes and processed securely to prevent improper use. Lastly it is important, that the controller of the data is accountable for ensuring compliance with these principles (cf. Art. 5 GDPR Principles relating to processing of personal data). All member states of the EU are obliged to apply the GDPR, which has been in effect since 2018 (Krämer & Mauer, 2023) The necessity for such a data protection regulation was acknowledged as early as 2010, as the ever-growing influence of new technologies, increased interconnectivity facilitated by new media and the sheer volume of data started becoming the focus of the social and political landscape. It was in this context that the concept of a standardized data protection regulation within the EU was initially proposed in 2010, aiming to foster data protection practices among the member states. However, several years and drafts passed before the GDPR was officially adopted in 2016 and became enforceable in 2018 (Taeger & Gabel, 2022).

Implications for HR Analytics

Unlike other nations (Greenleaf, 2013), the EU has created a legal framework to protect the personal rights of individuals with regard to their data. This framework is designed to achieve a balance between the protection of personal data and the encouragement of innovation. Due to these regulations, which aim to protect individual privacy rights, organisations operating in the EU must adhere to strict compliance standards when using employee data for predictive analytics. Data types such as health data or sensitive demographic or biometric data may pose challenges to the regular use of such models within the EU, as they could lead to discrimination or misuse. Gao et al. (2019), for example, use “Physical Condition” as a predictor for employee turnover. Their study was conducted in China, a country with no comparable data protection rules (Greenleaf, 2013). In the EU, on the other hand, the use of health data for prediction purposes would be subject to strict regulations because of the sensitivity of the data. The EU classifies health data as data which require specific consent.

Moreover, users of the data would need to implement countermeasures to prevent misuse or discrimination and would need to adhere to the principles discussed before. Companies or organisations operating within the legal framework of the EU must ensure that their practices comply with the GDPR. For instance, a study by Zhang et al. (2018) mentions how the data was pre-processed and briefly alludes to the features used, but it does not become clear where the data come from. A study conducted by Punnoose and Ajit (2016) takes a similar approach with the author mentioning the features of his dataset used for prediction in only a very vague and general way. If a study does not specify where their data comes from, it may fail to comply with the principles of transparency and accountability. The implications of the GDPR on the prediction of employee turnover data will be briefly discussed in the discussion part of this thesis.

3 Methodology

3.1 Systematic Literature Review

Identification of Papers

For the first part of the methodology of the present thesis a structured literature review was conducted by following Webster's and Watson's (2002) systematic approach. Literature reviews are of paramount importance for advancing research within a field, identifying gaps, and enhancing overall knowledge. Webster and Watson (2002) highlight the scarcity of reviews in the field of Information Systems (IS), partly linked to the field's relative youth.

To analyse which data types are used to predict employee turnover and to address the lack of reviews in this area, the following review adopted a concept-centered approach (Webster & Watson, 2002), with employee turnover prediction as its central theme. Relevant literature was identified by using selected keywords such as "prediction" or "predicting" in combination with terms like "employee turnover," "employee attrition," "employee churn," and "employee retention. Additionally, keywords related to predicting, including "machine learning," "data mining," "data," and "model," were also used. It should be noted that different conjunctions were used linking these terms. An overview of the exact search terms can be found in table 1.

Furthermore, for a literature review an approach that is not based on one methodology database is recommended (Bramer et al., 2017). Therefore, different search engines and databases were used in this work. Databases and search engines used included Google Scholar, EBSCOhost (featuring Academic Source Premier and Business Source Premier) and Scopus. Given the increasing relevance of predictive analytics in recent years, the temporal scope of the literature review was limited to studies published from 2010 onwards. This time restriction aims to analyse contemporary studies while at the same time focussing on a broader timeframe.

The aforementioned search process returned 316 search results with the keyword searches outlined before. Due to overlapping sources of databases, the appearance of duplicate entries was unavoidable. After the removal of these duplicates 222 distinct records remained. Following the approach outlined by Webster and Watson (2002) and Waffenschmidt et al.

(2019), the successive screening then focused on the remaining papers, articles and book chapters and their respective relevance to the research questions.

There were several inclusion and exclusion criteria defined to ensure the selection of relevant studies. First, only studies published in or after 2010 were considered to ensure the inclusion of recent research in the field. Second, only studies which included implementations of ML algorithms specifically for employee turnover prediction were included. This required that the data obtained and processed in the studies had to be tested using commonly applied ML algorithms, such as Logistic Regression, Random Forest, Neural Networks or similar. Studies that focused solely on improving the performance of existing models without introducing new datasets or features, were excluded as they were not considered relevant. Moreover, studies were required to provide a complete description of the set of features used for prediction. Studies that only referenced a few exemplary features of their dataset without providing a comprehensive description were excluded. Conceptual papers, review papers, master's theses, and PhD theses were also excluded from the review. For example, theoretical papers that discussed the potential impact of concepts such as job embeddedness or job satisfaction on turnover, without empirical data or model implementation, were excluded from the analysis.

Additionally, the frequent use of two publicly available datasets, the IBM Watson dataset and a dataset from Kaggle, was noted. To avoid redundancy, studies using the exact same dataset without alterations were included only once. However, studies were included if they altered the original dataset, used a subset, added features, or used it next to an additional dataset. Furthermore, studies were excluded if their primary focus was not on the prediction of employee turnover- for example if the examination of turnover was only a secondary aim-, if they tested a theoretical concept as a predictor or mediator, rather than having a dataset consisting of different features or if their aim was only to improve existing models or algorithms.

Initial screening concentrated on the abstracts and titles of the literature, based on the inclusion and exclusion criteria before. Records that did not meet these inclusion and exclusion criteria underwent the full-text screening process. Subsequent to this review, 148 papers with potential significance remained. These remaining papers were analysed more closely in terms of their overall content and relevance. Following the inspection of these

works, 30 distinct papers which met the inclusion criteria were selected and taken as a source to analyse which data types were used for the prediction of employee turnover. A visualisation of the selection process can be found in Figure 2.

Key Word Search	**Source**	**Number Results**
Employee turnover AND Prediction OR Predicting	Scopus	72
Employee turnover AND Prediction OR Predicting	Google Scholar	43
Employee turnover AND Prediction OR Predicting	EBSCO Host	18
Employee attrition AND Prediction OR Predicting	Scopus	59
Employee churn AND Data mining OR Data OR Machine learning	Scopus	14
Employee AND Attrition AND Model	Scopus	13
Employee AND Attrition AND Machine AND Learning	Scopus	45
Employee AND Attrition AND Machine AND Learning	Google Scholar	19
Employee retention AND Prediction OR Predicting OR Data OR Machine learning	Scopus	17
Employee churn AND Prediction OR Predicting	Scopus	16
Total		316

Table 1: Search Results Across Multiple Academic Databases

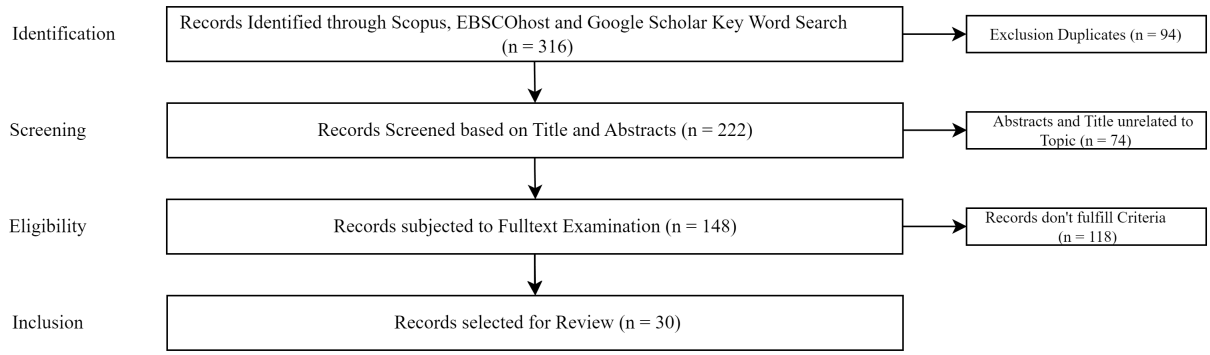

Figure 2: Visualisation of Selection Process of Literature Review

Analysis of Data

For the analysis of the data, a list of all features of the datasets found in the selected studies was created manually. In the next step, a Python program was executed to remove duplicates, which was then again followed by a qualitative check in order to avoid duplicates. The features found were checked for similar feature names and then included only once per category (e.g. Distance from Home vs. Distance Home or Department vs. Departments). Following this approach, a set of 286 distinct features was obtained. These features were grouped into overarching categories which align conceptually. A qualitative and human led

approach of grouping of features into conceptual categories offers flexibility in the analysis but at the same time can deliver reliable and insightful results when applied correctly (Braun & Clarke, 2006; Nowell et al., 2017).

Reduced Dataset

For the second part of this work the adoption of a reduced dataset was necessary. In the context of the 'black box' problem in HR Analytics, Kuhn, Johnson, et al. (2013) highlight the importance of meaningful feature selection. Consequently, the objective was to identify a meaningful set of features for the IBM Watson dataset. As an original dataset the IBM Watson dataset was used, which then was reduced according to the following principles. A full overview of the IBM dataset can be found in the Appendix.

A mixed-methods approach inspired by the one of Yahia et al. (2021) was adopted. First, the selected papers were analysed for their feature importance and a preselection of relevant features according to literature was carried out. In a second step those features were analysed quantitatively with feature selection algorithms.

In the first step, the importance rankings of the analysed literature were employed to derive an initial result of relevant features. Of the 30 papers analysed, 12 provided a feature importance ranking. Of these, three rankings were specific to particular features and did not merely refer to general features. Five of those papers give feature importance rankings but did not use them to reduce the dataset or do not say how they reduced the dataset exactly. The study conducted by Wang and Zhi (2021), which served as a reference point for this part of the thesis, was also affected by this, which precluded a direct adaptation of its feature importance. Therefore, also these papers were excluded from the analysis and 4 studies, which give feature importance and reduce their datasets accordingly, were used to analyse feature importance.

Of those 4 papers the features which were used for the respective prediction task, selected by feature importance ranking, were taken as a guide of which features can be considered important (full list of features can be found in the Appendix).

These features served as a basis for determining which features to include from the IBM Watson Dataset. Out of the 35 features of the IBM dataset 24 features (with one being the target variable Attrition), were equivalent or conceptually closely related (e.g. 'Graduated

Major' was found in literature and was equated as 'Education Field' in the IBM dataset) to the features detected in the most important features according to the feature importance rankings of the literature discussed before. This led to an initial reduction of the original dataset to the following features: Age, Attrition (Target Variable), Business Travel, Department, Distance from Home, Education, Environment Satisfaction, Gender, Job Involvement, Job Level, Job Satisfaction, Marital Status, Monthly Income, Overtime, Percent Salary Hike, Education Field, Performance Rating, Standard Hours, Total Working Years, Training Times Last Year, Years at Company, Years in Current Role, Years Since Last Promotion and Years with Current Manager. Some features, such as the feature Education Field, were not found as the exact same feature in the analysed literature, but showed characteristics which were conceptually related (such as Graduated Major), which were deemed as important and were thus also kept in the initial reduced dataset.

In a second step, those pre-selected features were further analysed using two feature selection techniques: *Recursive Feature Elimination* and *SelectKBest*.

The Recursive Feature Elimination (RFE) technique involved using five widely used and accurate classification algorithms: Random Forest, XGBoost, Logistic Regression, Support Vector Machine and Decision Tree. Parameters were tuned individually for each algorithm, since the dataset is imbalanced, though (1233 No Attrition, 237 Yes Attrition), all algorithms were adjusted and incorporated a parameter to balance the classes. Following the application of RFE to these algorithms, a majority vote was used to find the most relevant features. Specifically, a feature was kept if it was found to be important by at least three out of the five algorithms (majority vote), otherwise it was eliminated. An overview of the results can be seen on Table 2.

To evaluate the consistency of the feature selection method, the SelectKBest algorithm was also applied to the pre-selected features from the literature and the results were compared with the ones obtained with RFE. When applying SelectKBest, the imbalanced dataset was addressed using Synthetic Minority Over-sampling Technique (SMOTE). Additionally, 10-fold cross-validation was applied to ensure the reliability of the process. Only features which were selected across all 10 folds were considered as important and thus selected.

SelectKBest achieved comparable results to RFE, therefore further supporting the reliability of the approach. The selected 13 features were: Age, Business Travel, Environment

Table 2: Feature Selection Results Using RFE

Feature	**RF**	**XGB**	**LR**	**DT**	**SVM**	**Vote**
Age	Yes	Yes	No	Yes	No	Keep
Business Travel	No	Yes	Yes	No	Yes	Keep
Department	No	Yes	Yes	Yes	Yes	Keep
DistanceFromHome	Yes	No	No	Yes	No	Eliminate
Education	No	No	Yes	No	No	Eliminate
EducationField	No	Yes	No	No	No	Eliminate
EnvironmentSatisfaction	Yes	Yes	Yes	Yes	Yes	Keep
Gender	No	No	Yes	No	Yes	Eliminate
Job Involvement	No	Yes	Yes	Yes	Yes	Keep
Job level	Yes	Yes	Yes	Yes	Yes	Keep
JobSatisfaction	Yes	Yes	Yes	Yes	Yes	Keep
MaritalStatus	Yes	Yes	Yes	Yes	Yes	Keep
Monthly Income	Yes	No	No	Yes	Yes	Keep
Overtime	Yes	Yes	Yes	Yes	Yes	Keep
PercentSalaryHike	Yes	No	No	Yes	No	Eliminate
PerformanceRating	No	No	No	No	No	Eliminate
StandardHours	No	No	No	No	No	Eliminate
TotalWorkYears	Yes	Yes	No	Yes	No	Keep
TrainingTimeLastYear	Yes	No	No	Yes	No	Eliminate
YearsAtCompany	Yes	Yes	No	No	Yes	Keep
YearsInCurrentRole	Yes	Yes	No	Yes	Yes	Keep
YearsSinceLastPromotion	Yes	No	No	Yes	Yes	Keep
YearsCurrentManager	Yes	Yes	No	No	Yes	Keep

Note. RF = Random Forest, XGB = XGBoost, LR = Logistic Regression, DT = Decision Tree, SVM = Support Vector Machine.

Satisfaction, Job Involvement, Job Level, Job Satisfaction, Marital Status, Monthly Income, Overtime, Total Working Years, YearsAtCompany, YearsInCurrentRole and YearsWithCurrentManager. These features largely correspond to those found through RFE.

RFE identified the same features as SelectKBest and additionally found two more: Department and YearsSinceLastPromotion. When reviewing the literature, YearsSinceLastPromotion was ranked among the Top 10 important features according to Gao et al. (2019) and was selected by three algorithms in RFE, leading to its inclusion in the final dataset. The feature Department was found important in 3 out of the 4 papers used for the literature importance ranking and was selected by four algorithms in RFE, accounting to its inclusion as well.

Consequently, the final reduced dataset consists of 15 features selected through a combination of literature review, RFE and SelektKBest along with the target feature, Attrition.

An overview of the reduced dataset is presented in Table 3.

3.2 Predictive Modeling

Modeling was executed in Python 3.8.6 using the Machine Learning library *scikit-learn*.

Dataset

Following the methodology described in Section 3.1, 'Reduced Dataset', the dataset utilized for predictive modeling was prepared accordingly and can be found in table 3.

Table 3: Reduced IBM Dataset

Feature	**Data Type**	**Description**
Age	Numeric	Age of the employee
Attrition	Categorical	Whether the employee has left the company
Business Travel	Categorical	Frequency of business travel
Department	Categorical	Department in the company
Environment Satisfaction	Numeric	Satisfaction with work environment
Job Involvement	Numeric	Level of job involvement
Job Level	Numeric	Level of job position
Job Satisfaction	Numeric	Satisfaction with job
Marital Status	Categorical	Marital status of the employee
Monthly Income	Numeric	Monthly income
Overtime	Categorical	Whether the employee works overtime
Total Working Years	Numeric	Total years of working
Years at Company	Numeric	Years spent at the company
Years in Current Role	Numeric	Years in the current role
Years Since Last Promotion	Numeric	Years since last promotion
Years with Current Manager	Numeric	Years with current manager

Data Preprocessing

The data was inspected for missing values and duplicates. Object columns were transformed into categorical variables and One-Hot Encoding was applied to the categorical columns 'BusinessTravel', 'Department', 'MaritalStatus', and 'OverTime'. The target variable 'Attrition' was encoded as 1 for 'Yes' and 0 for 'No'.

Exploratory data analysis was conducted, focusing on descriptive statistics and visual inspection to identify irregularities or outliers.

When addressing numerical columns, a distinction was made between ordinal columns ('EnvironmentSatisfaction', 'JobInvolvement', 'JobLevel', 'JobSatisfaction'), for which standardization was deemed unnecessary, and columns with continuous data. Continuous data columns were assessed for skewness using both visual and statistical methods. Features which were skewed were normalized using the log1p function, which significantly reduced the skewness.

Modeling

For the modeling task, the data was split into training and test sets with a ratio of 70% for training and 30% for testing.

In the modeling part two distinct modeling tasks were performed. The first involved applying the 15 algorithms described by Wang and Zhi (2021), without specific parameter tuning. The second task involved tuning the hyperparameters of these algorithms through a grid search and addressing the imbalance through SMOTE technique.

To enhance the reliability of the results ten-fold cross-validation was employed and the average scores for accuracy, precision, recall, and F1 score were calculated to evaluate the performance of the models.

4 Results

4.1 Literature Review

Systematic Literature Review

For the systematic literature review a systematic keyword search as described in the methods part of this work was conducted. When using the different search engines Scopus, Google, Scholar and EBSCOhost, 316 search results could be obtained. After removing duplicates and scanning the papers for their relevance and eligibility for this literature review, 148 distinct papers remained, which were subjected to a full text analysis.

Only 30 of the 148 papers reviewed in the full text examination fulfilled the criteria to be included in this review. This underlines the problematic nature of the so-called black box approach, which as emphasised at the beginning of this paper, is a prevalent issue in HR analytics. Many of these studies do not disclose the data used for predicting employee turnover. They often only describe a few examples of the features used as an overview, rather than providing a full and transparent description of the data sets. This lack of transparency is incompatible with the principles of open science, for example, even the studies included in this review often did not fully describe their feature pre-processing methods (Abdullah et al., 2023).

Furthermore, even studies that were considered highly informative (Yahia et al., 2021; Zhao et al., 2019) did not describe the parameters used for their modelling processes. This lack of transparency hinders the reproducibility of the studies. This is another example of the black box approach, where the focus is on the mere application of AI rather than its thoughtful and accurate application. The study of Alamsyah and Salma (2018), for example, which was also included in the selected studies, does not provide its pre-processing steps in detail and does not perform feature engineering or selection.

While these practices show an interest in conducting HR Analytics, they often lack a systematic approach to understanding and explaining the underlying methods. This tendency makes reproducibility challenging and also undermines potential advances in HR analytics. The widespread black box approach suggests that despite the prevalence of AI tools in HR Analytics, their application is frequently superficial, with a primary focus on implementation

rather than a detailed analysis of the methods.
Another interesting finding was that of the 148 papers which were subjected to a full-text analysis, 82 used one of the publicly available datasets. Out of these, 59 used the dataset published by IBM Watson and 23 used the HR Analytics dataset from Kaggle, which is also freely accessible. It should be noted at this point that although some of the studies analysed changed the variable names of the datasets slightly (Pratibha & Hegde, 2022), used a slightly different version (Yadav et al., 2018) or did not specify where they had obtained their dataset from (Duan, 2022); on closer analysis it became apparent that one of the publicly available and commonly used datasets was utilised.

This frequent use of publicly available datasets highlights the sensitivity and difficulty of obtaining HR data due to its sensitive nature. HR data often contains personal and confidential information about employees, such as gender or monetary data, which companies are reluctant to share.
Additionally, data protection regulations and ethical considerations prevent the further disclosure of such data. The frequent use of HR analytics data sets provided by IBM Watson and Kaggle emphasises these challenges. The straightforward availability of these datasets allows easy access to such data for research, which explains their frequent usage as well.
Furthermore, the use of available datasets can significantly reduce the time and resources required for data collection. This may be particularly beneficial in academic settings, where resource constraints often are a limiting factor.

Overview Selected Studies

When analysing the 30 selected studies, different approaches in using models and datasets, can be found. Key findings from this analysis are described in the following; a detailed table of the papers and their main findings can be found in the Appendix.

In general, the data used can either be related to a specific industry or country or they used more general data that can be used across industries (see table **??** and table **??**). More specifically, some studies used proprietary company data, others publicly available datasets and yet others used industry-specific data. For example, the dataset from a Japanese restaurant chain was used to develop an LSTM model for early turnover prediction (Sato et al., 2019), whereas Zhu et al. (2019) used data from China's biggest professional social

platform data to combine their results of survival analysis with ensemble learning.
The dataset from Japan covering the restaurant industry, for example, provides detailed information specific to this industry. Key variables here include the day of the week, the start time and the end time of the working days as well as the working hours of an employee. These variables are crucial to understanding operational dynamics and management in an industry characterised by irregular working hours and changing customer demands. Data from China's professional social media platform, on the other hand, provide insight into the engagement and influence of users on digital platforms. Interesting variables include, for example, the number of comments, the number of likes, the influence on others or the interactions on this platform, next to more general variables such as organisational or personal variables. These variables are central to understanding the digital behaviour, social influence and engagement patterns of these users in China. This data however, which showed the influence of social media, is specific to this study from China. However, it was found that data such as gender and GDP and not the social media variables, such as the number of interactions, had the strongest predictive power.

A different, innovative approach was used by El-Rayes et al. (2020). They used data from anonymous resumes submitted on Glassdoor together with publicly available data from companies to predict turnover. At this point, data such as the year the company was founded, the rating of the company by employees and the industry to which the company is assigned were analysed. This data, for example, was used to benchmark company performance, to analyse the ratings of former employees towards their employers and to filter out industry trends in order to draw conclusions about turnover intention.
Other studies focused on specific industries, such as the garment industry in Bangladesh, where Nahar et al. (2022) predicted turnover using industry-specific variables, and Mozaffari et al. (2023) predicted employee turnover with data from the pharmaceutical industry in Iran.
These studies highlighted some industry- and country-specific variables. For instance, the study related to the garment industry in Bangladesh used unique data types not found in the other selected studies, such as 'Safety Satisfaction', 'On-Time Salary' and 'Annual Refreshment Facility'. In the study, which took place in the pharmaceutical industry, for example, the following factors were considered to be most important: having children and

having a long-term employment contract reduced the likelihood of turnover.

Studies using more general data have focused on various standard employee metrics. These variables typically include demographic variables and other descriptive measures such as age, gender, department, education, job satisfaction, monthly income and work-life balance. They also cover aspects such as promotion history, job performance and tenure. This approach allows a comprehensible and generally valid overview that covers various personal and organisational factors and can effectively predict turnover and make the results more generalisable.

Studies from Iran (Esmaieeli Sikaroudi et al., 2015) and the United Arab Emirates (Alshehhi et al., 2021) again showed country-specific variables that could not be found in other studies. For example, the study from Iran mentions variables such as the compatibility of the body with the job or knowledge of labour rights as well as veteran status as predictors, while the study from the UAE mentions the last action taken by the employer against the employee. These studies might highlight country-specific employment variables due to unique cultural, legal and social factors.

The understanding of the data used is related to understanding the goals of the different studies analysed and understanding the motivation behind the research.
Regarding the purposes of the studies, most studies mentioned they wanted to understand employee turnover further and wanted to analyse the key factors influencing it (Bandyopadhyay & Jadhav, 2021; Dolatabadi & Keynia, 2017; Saradhi & Palshikar, 2011). Other goals were the early detection of turnover (Sato et al., 2019), the comparison of different dataset sizes and algorithms (Zhao et al., 2019), test models with longitudinal data (Zhu et al., 2019), test new algorithms (Gao et al., 2019) or use a novel approach for predicting employee attrition with ensemble learning and fuzzy inference (Sharma et al., 2022).
Understanding the approaches and types of data used for the prediction of employee turnover in the analysed studies emphasises the complexity of turnover prediction. By examining the data used in these studies, it is possible to identify the variables which play a critical role for predicting turnover. An overview and description of commonly found variables used for turnover prediction can be found in the following section.

Data Used for the Prediction of Turnover

One of the goals of this research was to identify the data used for predicting employee turnover. Table 4 shows the frequency with which the most commonly used data variables appear across the reviewed papers. For example, the frequency of 23 for the variable 'Gender' indicates that gender appeared in 23 out of the 30 analyzed papers as a predicting factor.

Table 4: Frequency of Variables in Analyzed Literature

Variable	**N**	**Data Type**	**Variable**	**N**	**Data Type**
Gender	23	Categorical	Years at Company	8	Numeric
Age	22	Numeric	Job Involvement	6	Categorical
Salary	20	Categorical/Numeric	Total Working Years	6	Numeric
Education	16	Categorical	Distance from Home	6	Numeric
Department	15	Categorical	Years in Current Role	6	Numeric
Promotion	15	Categorical/Numeric	Average Monthly Hours	6	Numeric
Job Satisfaction	13	Categorical	Number of Companies Worked	5	Numeric
Tenure	13	Numeric	Number of Projects	5	Numeric
Marital Status	12	Categorical	Overtime	5	Categorical
Experience (Years)	10	Numeric	Business Travel	5	Categorical
Training	10	Categorical/Numeric	Work Accident	4	Categorical
Performance Rating	9	Categorical	Years with Current Manager	4	Numeric
Environment Satisfaction	8	Categorical	Relationship Satisfaction	3	Categorical

Table 4 integrates various closely related terms and synonyms used for each variable into a single representative name. Variations such as plural versus singular forms, different spellings, or slightly different names are also grouped together under one variable.
For brevity, only exemplary terms and not the full list of synonyms and related terms are mentioned at this point. The complete list of features per paper can be found in the appendix.

For instance, the feature 'Salary' consists of a range of related terms such as compensation, monthly income, yearly salary increment, payment tier, Employee CTC Level or rewards. The term 'Rewards', as used by Yahia et al. (2021), includes pay and organizational rewards but was counted only under 'Salary' and not included in the 'Promotion Frequency'.

In the same way, "Education" includes terms like education background, education field, education qualification or highest education, for example. Other features are summarized in the same way: "Department" covers,department type, division, and department of working; "Promotion" includes terms such as the promotion last year, promotion in the last 5 years, the last pay raise, the promotion status, percent salary hike, the time since salary increase,

and the number of promotions within the organization. The "Tenure" feature encompasses time spent at the company, years at the company, and number of years at the company. "Experience" also includes first-grade experience, experience in the current domain, relevant experience, and experience at the current company. "Training" covers general training, training times last year, training count, and training hours. "Performance Rating" includes terms such as last evaluation, performance rating last year, and job performance.

In order to gain a more comprehensive understanding of the features, a qualitative and systematic description will be provided in the following section. The most frequently cited variable was 'Gender', followed by 'Age' and 'Salary', indicating these are the top three features according to their frequency. It appears evident that demographic features appear in most studies since they are normally readily available to companies. Similarly, more descriptive measures about employees, such as the education of an employee or the department where an employee works can also be accessed easily. Regarding the encoding, age, for example, was represented across all studies as a continuous numeric variable. It encompassed ages ranging from 18 to 50-60 (depending on how old the oldest employee at the time was). Gender was always a categorical feature with two distinct values: 'female' and 'male' (or 0/1, 1/2). The option for alternative genders was not found in the analysed studies. For the salary feature, both approaches were found: a categorical distinction with salary as low, medium or high, or numerical features with the amount of salary given. Since these values are often country specific (for example Gao et al. (2019) from China distinguished values from 5,000 to 97,938) a categorical distinction seems to offer better comparability. Education, on the other hand, was always measured as a categorical feature. Values included the different education levels such as 'Below College', 'College', 'Bachelor', 'Master', 'Doctor', and 'Other', or Academy, Bachelor, Master, Doctor, or Primary school, higher school, higher diploma, bachelor, master, diploma, and PhD. Department was again always classified as categorical. Here, different distinctions were made: Sales, Technical Support, IT, and Product Management, or HR (1), R&D (2), and Sales (3). These distinct departments reoccurred across literature, which could hint at that many studies were inspired by each other or took one of the publicly available datasets as an inspiration.

Salary on the other hand, can be highlighted as the most frequent predictor that is not related to demographics. This finding is highlighted by a study (Iqbal et al., 2017), which

showed that salary is one of the main indicators of employee satisfaction, and employee satisfaction is one of the most important indicators for retaining employees.
The promotion variable is closely related to salary, as promotions often involve salary increases, making it reasonable for it to appear frequently among the top features and shortly after salary. Promotion in the Last 5 Years was recorded as either 0 or 1, representing a binary variable, indicating whether there was a promotion in the last 5 years. Additionally, it was treated numerically, indicating the number of years since the last promotion or it was encoded as Percent Salary Increase, which was then again a numerical value.

Job satisfaction is one of the most important predictors of turnover intention (Samad, 2006). Some studies also include more specific types of satisfaction, such as environment satisfaction and relationship satisfaction, which indicate different aspects of an employee's satisfaction overall. In general, satisfaction variables appear frequently among the analysed studies. Despite their prevalence across studies and their predictive power (Bandyopadhyay & Jadhav, 2021; Srivastava & Eachempati, 2021; Wang & Zhi, 2021; Yahia et al., 2021), it must be noted though, that these variables are often not readily available to employers and typically require collection through surveys. This contradicts the idea of a fast and automatic assessment of an employee's intention to leave, which is the goal of modern HR approaches. Relationship Satisfaction, Environment Satisfaction, and Job Satisfaction were always encoded as categorical features with values like Low, Medium, High, and Very High.

Also closely related and important are the variables training and performance. Increased investment in the training of an employee offers companies the opportunity to increase their performance (Elnaga & Imran, 2013). Increased performance may also result in higher turnover because employees with a high human capital are also more likely to leave the company to look for better job opportunities (Wei, 2015). Training was represented as either a numeric feature indicating the hours of training or as the number of times training occurred. The Performance feature was always encoded categorically with values like Low, Good, Excellent, and Outstanding, indicating different levels of performance.

Moreover, variables such as "Years at Company," "Years in Current Role," "Total Working Years," "Experience," and "Years with Current Manager" all are related to the tenure of an employee and experience both within the company and their career in general. This indicates that tenure and experience might be significant factors in predicting employee turnover.

These variables were always measured numerically, indicating the duration an employee has been with the company, usually in years.

The variables "Overtime" and "Average Monthly Hours" highlight the amount of work-life balance an employee experiences. Higher levels of overtime and many working hours can be possible criteria for turnover due to overwork or job dissatisfaction. Work Accident was encoded as a binary feature, indicating whether an accident had occurred, whereas the average monthly hours were measured numerically in hours.

During the analysis, it was observed that all variables, except for "Number of Projects" and "Work Accident," listed in Table 4 are variables used in exactly the same way or conceptually closely related to the IBM Watson dataset, which was described in an earlier part of this work. This wide usage suggests that the IBM Watson dataset, also when not explicitly used in studies, still has influenced the selection of data in other datasets. This prevalence underlines the dataset's impact on the field of employee turnover prediction and its reliability and relevance. Therefore, it can be considered, a significant dataset for understanding and predicting employee turnover dynamics and thus it is also used in the second part of this work as a initial dataset.

4.1.1 Categories

The explanation of the categorisation process is described in the following. Initially, the variables of the selected studies were identified and in the following, extracted manually. This provided a comprehensive overview of variables used in literature for turnover predictions. In the following step, duplicates were removed, which reduced the number of variables a lot. In the next step the found variables were carefully examined to understand their meanings and the contexts in which they were used. At this point, it is important to note that unfortunately this was not possible for all variables, since not all papers gave a description of their datasets. Variables were then grouped based on conceptual similarities. For example, salary and promotion-related variables were grouped together, since they are conceptually related by remuneration (compare with Schema Theory Anderson and Pearson (1984)). This conceptual grouping facilitated the creation of a preliminary list of categories, which was the next step. Initially, ten categories were identified, which were: Remuneration and Compensation Progression, Employee Characteristics and Demographics, Time Spent at Company, Role

Related Factors, WorkLife Balance, Level of Qualification, Company Factors, Evaluation and Training, Experience and Education, Satisfaction and Organisational Climate. In the next step, the variables were systematically placed into the preliminary categories (cf. Figure 3) and allowed an initial overview of categories with variables. It should be noted, however, that the categorisation of variables was not entirely clear-cut, as many lacked a clear definition and interpretation. Nevertheless, this approach enabled an overall understanding of the categories to be gained.

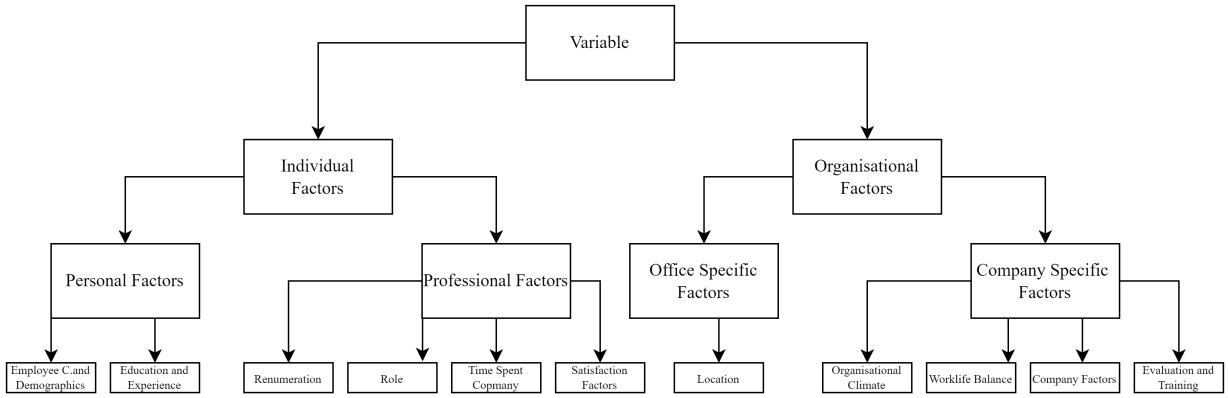

Figure 3: Process Flow for Variable Classification

The preliminary categorization was then reviewed and discussed. This discussion aimed to evaluate the categorization and to identify potential misunderstandings. Thereafter, a review, further consultation in literature and a comparison of variables in different studies. This iterative approach resulted in 11 categories. This refinement led to the addition of two new categories: "Location" and "Organisational Climate." Additionally, the "Level of Qualification" category was merged into "Education and Experience" based on their close relationship. After discussion and iteration, the final 11 categories were established. A visualisation of the categories process can be found in Figure 4.

This categorization aims in helping to identify patterns and drawing insights that are essential for developing targeted retention strategies. A categorisation was necessary because of the sheer amount of variables. Across all studies analysed 500 variables were identified, when considering only distinct variables this number is reduced to 286 variables. For example, variables such as turnover or employment status were not counted in the instinctive count of predictor variables because they represent the target variable. In addition, slightly

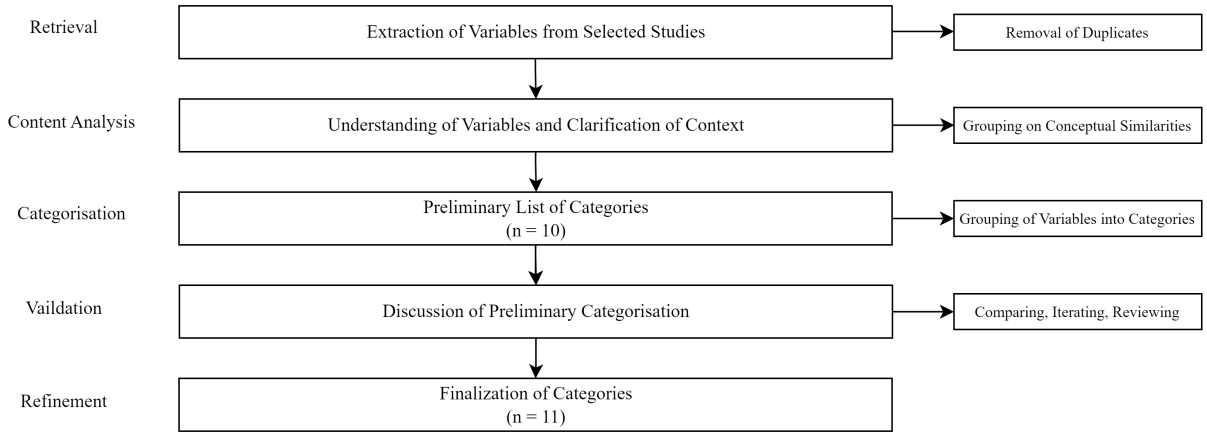

Figure 4: Categorization Process

different spellings of the same variable were counted only once (e.g. number of projects and number of projects or distance from home and distance from home (km)). The feature 'employee number' was counted twice, since it once appeared as the employee count and once as an employee id. However, it is important to note that these numbers depend strongly on the methods in how features are defined. For example, factors such as an additional descriptive word can result in the same feature being counted multiple times. If only conceptually distinct features were counted, the number would be even lower, as many features are synonyms or closely related. A full overview of all variables of all papers and all distinct variables can be found in the appendix.

The following sections provide detailed definitions and descriptions of these categories. An overview of the 286 variables assigned to the different categories can be found in table 5. The meaning of the variables 'winning count', 'transfer cost' and 'opportunity' were not described in the respective papers- therefore there meaning had to be assumed and they were assigned accordingly. To provide a comprehensive overview and to understand which variables are within each category, key variables per category and the number of their occurrence are listed under the respective definition. It should be noted that for this count, only variables that are exactly equivalent are included, which may result in a discrepancy in the number of variables compared to that presented in Table 4. These variables were obtained by reviewing the analyzed papers, which provide feature importance rankings as it was done for the reduction of the IBM dataset. For the key variables of the category 'Organisational Climate' the work of Bandyopadhyay and Jadhav (2021) was consulted. An overview can

be found in the appendix as well.

4.1.2 Definitions of Categories

Remuneration and Compensation Progression Related Factors: Variables related to the financial compensation and career progression of an employee within an organisation. Includes variables such as the monthly income, salary increases and benefits. In the context of Herzberg's Two-Factor Theory, financial incentives are considered hygiene factors (Herzberg, 1966) and thus if not fulfilled lead to dissatisfaction which then again may lead to attrition. This category consists of 37 variables.

Key Variables	Number of Occurrences
Percent Salary Increase/Years Since Last Promotion	7
Monthly Income/Income Base	6
Mission Cost	1
Rewards	1

Employee Characteristics and Demographics: Contains variables related to the personal and demographic characteristics of employees. Demographic factors, such as age, gender, marital status, and ethnicity, are important for understanding employee turnover because they are readily available to most companies on the one hand and on the other hand also have strong predictive value in various turnover prediction studies (Schlechter et al., 2016; Sibiya et al., 2014). This category consists of 37 variables.

Key Variables	Number of Occurrences
Gender	23
Age	22
Marital Status	12

Time Spent at Company: Contains variables related to the duration of the employment and the time spent in specific roles within the organisation. This category contains variables directly related to the overall tenure such as the total years spent at the company or the time in the current role. Also variables related to the job and contract duration, such as the engagement date, the last new job and the duration of previous contracts, are included. Employee tenure can be considered a critical factor in turnover studies. Longer tenure often correlates with higher organizational commitment (Allen & Meyer, 1993), which is related

with a lower turnover intention (Ramalho Luz et al., 2018). This category consists of 30 variables.

Key Variables	Number of Occurrences
Years at Company/ Tenure	10
Years in Current Role	4
Years With Current Manager	4
Last Contract Duration	1

Role Specific Factors: This category includes variables related to the specific roles and positions of employees within the organisation. Variables here include variables connected to role and title, to the contract and job nature ('Contract Type', 'Is Client Facing Role') and variables related to management and team structure ('Direct Manager', 'Team', 'Management Level'). The Job Characteristics Model (Hackman & Oldham, 1974) suggests that the nature of the job itself can impact employee satisfaction and conclusively turnover, thus justifying the grouping of role specific variables together. This category consists of 45 variables.

Key Variables	Number of Occurrences
Business Travel	5
Direct Manager	1
Office Base	1
Position Category	1
Manager	1
Job Title	1

Location Factors: Includes variables related to the geographical context of an employees' work. Variables directly related to the location such as 'Location' and 'City' or variables describing the location further ('City Development Index') can be found in this category. Location factors are important to consider since they, for example, affect employees' commute times which can influence the turnover intention of an employee, particularly in long commutes (Kim & Park, 2017). This category is among the smaller categories and consists of 8 variables.

Key Variables	Number of Occurrences
Location/Employee Location	3
Distance from Home	6
City of Residence	1

Worklife Balance: This category encompasses variables that measure the balance between work responsibilities and personal life. Variables include variables such as the number

of projects or the standard working hours, but also variables connected to flexibility and leave ('Leave Entitlement', 'Flexible Working Hours'). Neglecting the variables just described results in poor work-life balance, which leads to job dissatisfaction (refer to Herzberg (1966)) and turnover. This category is also among the smaller ones and consists of 9 variables.

Key Variables	Number of Occurrences
Average Working/Monthly Hours	6
Overtime	5

Company Factors: This category includes variables related to the organizational context and structure. Variables such as organisational divisions ('Department', 'Unit'), size and type of the organisation and other organisational variables ('On- Time Salary', 'Competence Center' = Specialized Unit, 'Children Cost' = Benefits For Employees With Children) are found in this category. Organisational divisions such as the department for example are found to be significant predictors of turnover in different studies (Gao et al., 2019; Gurler et al., 2023; Zhao et al., 2019). This category consists of 28 variables.

Key Variables	Number of Occurrences
Department/Units	15
Children Cost	1
Competence Center	1

Evaluation and Training: This category covers variables related to employee performance evaluation and training activities. Variables are related to the frequency and duration of training activities or include different performance metrics and appraisal results. Performance evaluations and training opportunities are vital for the development of an employee and his or her career progression, since they provide feedback as well as an improvement of skills which are crucial for job satisfaction and motivation. Studies show that employees who feel valued and get offered opportunities for personal and professional development are less likely to leave the organisation (Memon et al., 2016). This category consists of 15 variables.

Key Variables	Number of Occurrences
Training/Training Times Last Year	10
Job Performance/Peformance Rating	6

Education and Experience: This category includes variables related to the educational background and work experience of employees. Variables cover academic qualifications, disciplines, specialisations, attended institutions, professional experience, the work history as well as professional records. This category is the largest and consists of 47 variables.

Key Variables	Number of Occurrences
Education*	14
Total Working Years	6
Grade	4
Graduated Major/ Major Discipline	4
Job Level	4
University Graduated/Enrolled in	2
Record by Day and Month	2
First Grade Experience	1
Previous Experience	1
Total Experience	1

***Note:** Education contains Education Level, Education Qualification, and Education Status.

Satisfaction Related Factors: This category encompasses variables that measure various aspects of employee satisfaction. General and more specific satisfaction variables are included as well as job involvement. Employee satisfaction is a well-documented predictor of turnover as already discussed (Samad, 2006). This category consists of 20 variables.

Key Variables	Number of Occurrences
Environment Satisfaction	8
Job Satisfaction	8
Job Involvement	6

Organisational Climate: This category includes variables related to the overall organizational environment and culture. Here variables include organisational aspects and climate, support structures, interpersonal relations and managerial actions. The organizational climate can significantly affect employee turnover. A study from Madden et al. (2015), for example, shows that a positive work environment and a supportive management can enhance employee retention. This category consists of 10 variables.

Key Variables	Number of Occurrences
Challenging Work	1
Work Recognition	1

Table 5: Variables by Category

Category	Variables
Remuneration and Compensation Progression	Final Salary, Base Salary, Compensation, Average Salary (OriginalJob/NewJob), CTC Level (Annual Remuneration of the Employee, Measured in Cost to Company = CTC), Daily Rate, Hourly Rate, Monthly Rate, Employee Promotion Status, Gross Salary, Income, Income Base, Last Pay Raise, Median Salary, Monthly Income, Mission Cost (Payment for Mission), Payment, Payment Tier, Percentage Salary Hike, Percent Salary Hike, Percent Salary Increase, Promotion, Promotional Chance, Promotion Last 5 Years, Promotion Last Year, Overtime Bill Allowance, Rewards, Salary, Salary Level, Stock Option Level, Years Since Last Promotion, Yearly Salary Increment, Transfer Cost (Allowance for Leaving? No Explanation), Time Since Last Grade Increase, Time Since Last Salary Increase, Months Since No Grade Increase, Months Since No Salary Increase
Employee Characteristics and Demographics	Age, Ethnicity, Date Of Birth, Gender, CDL (Commercial Driver's License), Behaviour (Of Employee), Designation, Employee ID, Employee Name, Employee Number (ID), Have Children, Hiring Channel, Main Financial Support, Marital Status, Native Place, Over 18, Physical Condition, Veteran Status, Separation Type (Voluntary/Involuntary), Probation, Column Number, Comment Number, Impression Tag Number, Influence Number, Likes, Number Interactions, Number Posts, Percentage Influence On Others, Percentage Information Perfection, Profession Tag Number, Recent Feeds, Views, Standpoint Numbers, Designation Client Organisation, ID, Stress Level, Perseverance and Interest to Work
Time Spent at Company	Action Date (Year Of Leave), Employment Month, End Year, Engagement Date, Job Tenure, Last Contract Duration, Last New Job, Number Job Changing, Number Years Company, Previous Working Periods, Starting Year (New Job), Starting Year (Original Job), Start Year, Tenure, Termination Date, Time Spent Company, Year Of Joining Company, Years At Company, Years In Current Role, Year Hired, Year, Time Worked, Start Date, Month Of Observation, Month Of Starting Date, End Date, Effective Date, Length Working Time, Months Before Termination Date (Number Of Months Between The Effective Date And The Termination Date), Years With Current Manager
Role Specific Factors	Business Travel, Contract Type, Direct Manager, Employment Nature, Employment Title, Job Title, Job Role, Level Of Position, Management Level, Manager, Office Base (Are Employees' Job Locations Based In Office), On Site/Off Site, Position, Position Category, Position Level, Role In Restaurant, Safety Measure (Is Work Environment Safe), Working Type, Team, Work Accident (Yes, No In This Company Per Month), Winning Count, Title, Billed/Not Billed (Number Of Billed And Not Billed Months), Compatibility Of Body With Job, Is Client Facing Role, Finishing Time (Restaurant), Day Of The Week (Restaurant), Starting Time (Restaurant), Scope Of Service, Request Channel, Service Class, Service Line, Service Type, Average Time Assigned to each Service, Duplicate Service or Not, Miss Call Service or Not, Number of Duplicate Service, Referring to the Software Development Team or Not, Resolved Service on the First Calling or Not, Resolved Service on the first Day or Not, Time Spent on Projects, Cumulative Incidents, Cumulative Responsible Accidents, Challenging Work, Number of Projects
Location Factors	Distance From Home, Assigned Garage, GDP (Of China), City Development Index, City Office Where Posted, City, City Of Residence, Location
Worklife Balance	Average Monthly Hours, Avg Work Hours, Leave Entitlement, Flexible Working Hours, Overtime, Standard Hours, Work Life Balance, Workload, Working Hours
Company Factors	Department, Department Of Working, Department Type, Employee's Department, Company Size, Company Type, Branch, Children Cost (Benefits For Employees With Children), Competence Center, Division, Employee Count, Employee Number (Count), Human Resource Policy, Industry, Industry Type, Job Vacancy (Sub Division), Number Employees (New Job), Number Employees (Original Job), On-Time Salary, Overall Employee Rating (New Job) Of Company, Overall Employee Rating (Original Job) Of Company, Sales (Department), Units, Year Company Founded (New Job), Year Company Founded (Original Job), Transport Facility, Annual Refreshment Facility, Opportunity
Evaluation and Training	Appraisal Rating, General Training, Job Performance, Last Evaluation, Performance Rating, Performance Rating Last Year, Performance Score, Performance (2015, 2016), Training, Training Count, Training Hours, Trainings Last Year, Training Times Last Year, Average Number Of Services Given Per Year, Service Time
Education and Experience	Education, Education Status, Education Background, Education Field, Education Level, Education Qualification, Degree Level, Degree Major, Experience, Experience Client Organisation, Experience Parent Organisation, Experience Current Company, Experience In The Current Domain, Area Of Expertise, Duration Relevant Work Experience, Enrolled University, First Grade Experience (Number Of Companies Already Worked With And Associated From Grade 1 (Highest Rank)), Graduated Major, Grade, Highest Education, Job Level, Knowledge About Working Conditions And Laws, Level Of Education, Major Discipline, Number Companies Worked, Number Turnovers, Qualification, Previous Experience, Record By Day, Record By Month (Qualification Related), Relevant Experience, Social Interaction Skills, Specialized Area, Technical Skills, Total Experience, Total Working Years, University Enrolled In, University Graduated, Years Of Experience, Work Experience, Working Experience Current Company, Communication Skills, Compatibility Experience With Job, Max Degree (Highest Education), Max School Type (Level Of Highest School), Relating Certificate To Employee Service, Average Working Experience Other Companies
Satisfaction Related Factors	Career Growth Opportunities, Environment Satisfaction, Job Autonomy, Job Environment Satisfaction, Job Involvement, Job Satisfaction, Job Search Behavior, Cumulative Absences, Relationship Satisfaction, Safety Satisfaction, Salary Satisfaction, Satisfaction, Satisfaction Level, Turnover Intention, Continuous Commitment, Affective Commitment, Positive Affectivity, Negative Affectivity, Normative Commitment, Organizational Commitment
Organisational Climate	Discrimination, Distributive Justice, Encouragement, Goodwill, Last Action Type Taken Against An Employee By The Employer With Respect To Their Employment, Manager Support, Ever Benched Or Not, Peers Leaving, Work Recognition, Work Environment

[1]Note: Variables specific to Chinese social media platforms: Interaction Number, Posts Number, Standpoint Number, Column Number, Comment Number, Likes, Viewed Number, Recently Received Feeds Number, Influence Number, Percentage Influence Over Others, Percentage Information Perfection, Impression Tag Number, Profession Tag Number.
[2]Note: Variables specific to Employee Client Services: Duplicate Service or Not, Miss Call Service or Not, Number of Duplicate Service, Referring to the Software Development Team or Not, Request Channel, Resolved Service on the First Calling or Not,Resolved Service on the First Day or Not, Scope of Service, Service Class, Service Time, Service Type

4.2 Modeling Results

4.2.1 Exploratory Research

In the following, different relationships in between commonly cited variables and attrition get visualised in an exploratory research step.

Figure 5 shows the relationship between an employee's job satisfaction and their attrition status. It can be seen that employees who are satisfied with their job are more likely to stay at the company, while those who are less satisfied are more likely to quit. Interestingly, a satisfaction level of 3 is associated with higher attrition compared to a satisfaction level of 2. This finding could be influenced by other variables which affect the turnover decision. Nevertheless, the general trend indicates that lower job satisfaction leads to more employees leaving.

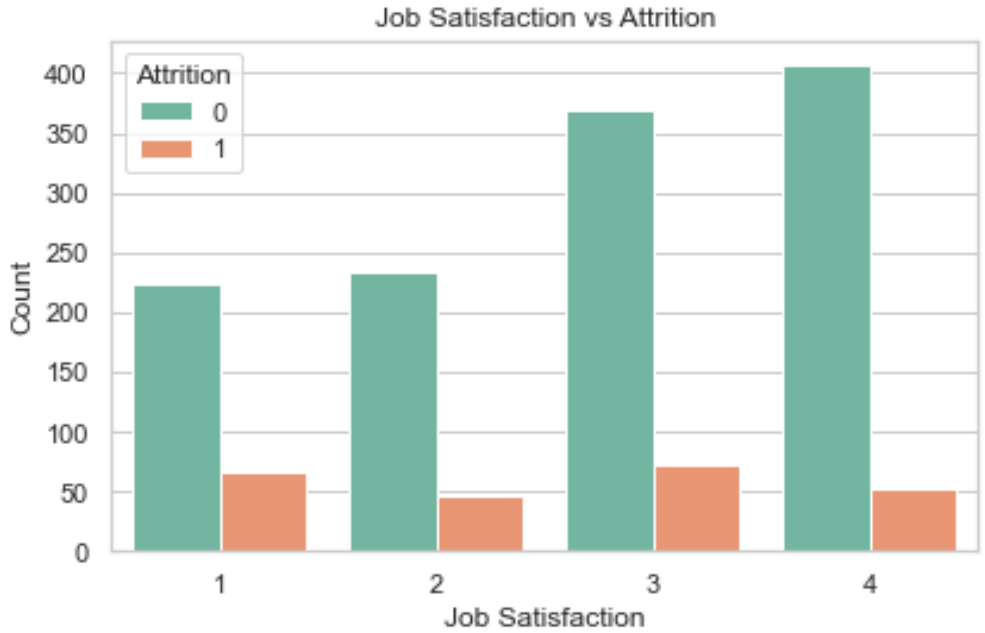

Figure 5: Variations in Job Satisfaction and Attrition

Figure 6 shows the monthly income and its relationship with attrition. It can be seen that employees who have higher salaries are more likely to stay at the company, which makes sense from a theoretical perspective. Relatively high salaries can also be found in the employees who left, underlining again that many variables together and not one alone lead to attrition.

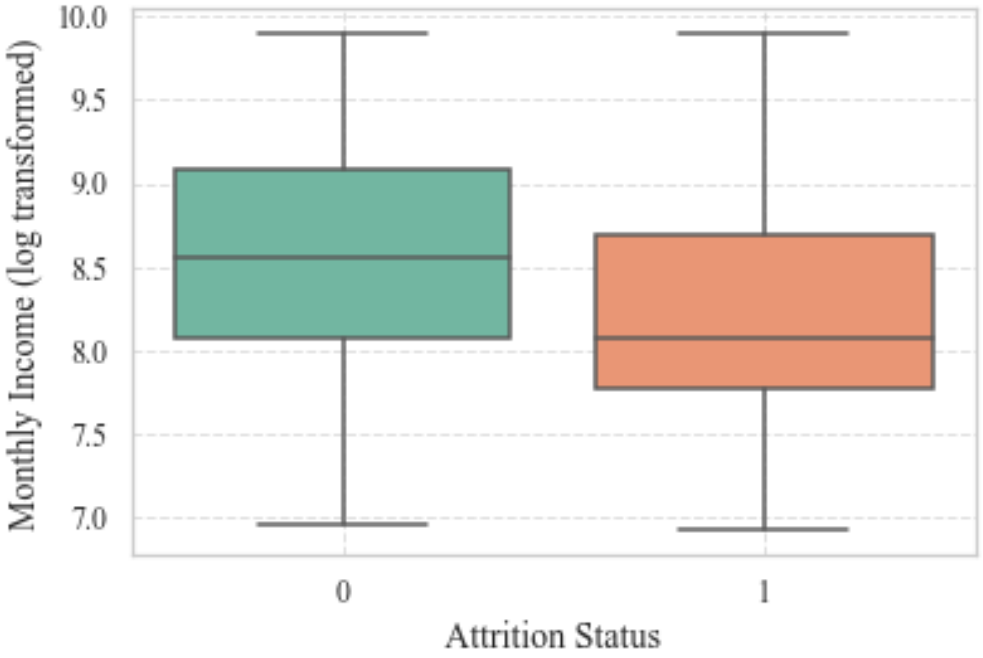

Figure 6: Salary Distribution and Attrition

Figure 7 shows the correlation heat map for the selected variables and the attrition variable. There is a moderate negative correlation (r = -0.1982) between the monthly income and attrition. This suggests that a higher monthly income is related to a lower likelihood of attrition, meaning that employees with higher salaries are less likely to leave. Furthermore, there are negative correlations in between total working years and attrition (r = -0.2309), as well as the years at the company and attrition (r = -0.1985), indicating that employees who have more years of work experience and stayed for longer with the company are less likely to quit. Moreover, there are variables with very high correlations within each other, such as job level and income (r = 0.9201) and the years at company and the years in the current role (r = 0.8363), which could lead to multicollinearity related issues. Since this possible issue wasn't addressed in Wang and Zhi (2021) the decision was made to retain these variables in this study to maintain consistency with their methodology.

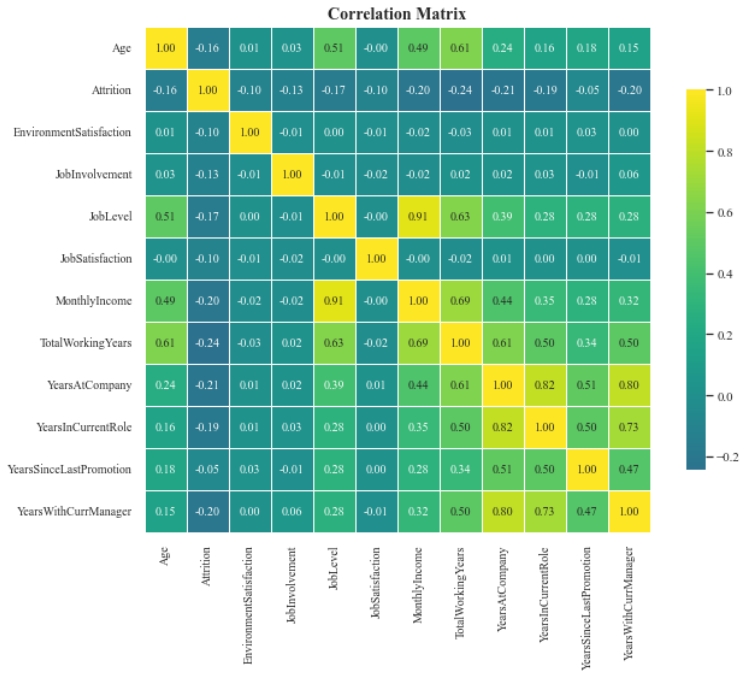

Figure 7: Correlation Matrix

Similarly to many attrition datasets, the IBM dataset is severely imbalanced (approx. 16% of employees are 'leavers'), as can be seen in figure 8. This can be explained through the nature of the matter, as employees leaving the company is a rather rare and non-standard phenomenon. Nevertheless, this problem affects many of the various analysed studies. After being acknowledged, this imbalance of attrition datasets is addressed in different ways by the various studies analysed. Wang and Zhi (2021) as well as Zhao et al. (2019), for example, point out that sole reliance on accuracy is unsuitable for the evaluation of a Machine Learning model in this case and that they therefore also consider evaluation measures such as precision and recall, the F1 score and the ROC-AUC curve. Saradhi and Palshikar (2011) on the other hand address the imbalance of their dataset by adjusting class penalties to their algorithms used. And Gurler et al. (2023), conversely, use an oversampling technique of the minority class to counteract the imbalance of their dataset.

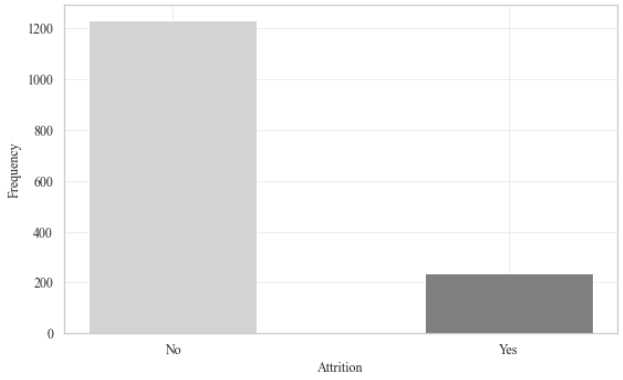

Figure 8: Imbalance of Target Variable

Altogether, it should be noted though that only a minority of the studies analysed recognised the class imbalance problem, addressed it and used appropriate solutions.

4.2.2 Model Evaluation

In this work, the aim was to evaluate and compare the performance of the base models presented by Wang and Zhi (2021) when applied to a reduced version of the dataset.

In particular, the models evaluated in Wang and Zhi's work included Random Forest Classifier (RF), Extra Trees Classifier (ET), Decision Tree Classifier (DT), Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGB), CatBoost Classifier (CB), Gradient Boosting Classifier (GBC), Ada Boost Classifier (ADA), K Nearest Neighbors Classifier (KNN), Quadratic Discriminant Analysis (QDA), Naïve Bayes (NB), Logistic Regression (LR), Ridge Classifier (Ridge), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM). The metrics used to evaluate these models included accuracy, recall, precision, and the F1 score. The results of the very same models used for this work were then compared against the results reported by (Wang & Zhi, 2021)

Table 6 shows the overall performance of the 15 models for both studies.

When comparing the results of table 6, it can be noted that, overall, the models with the reduced IBM dataset outperform the models from Wang and Zhi (2021) in terms of recall, precision, and the F1 score. It should be noted though that the exact parameters used for modelling in the study of Wang and Zhi (2021) were not available even upon request to the authors, making an accurate comparison challenging. This circumstance was further aggravated by not clarifying if and if yes, how many and which features were removed, as the

Table 6: Model Performance: Comparison of Results

Model Performance (Wang & Zhi, 2021)

Model	Accuracy	Recall	Precision	F1
LR	0.8755	0.4062	0.7314	0.514
LDA	0.8658	0.3827	0.6663	0.4765
ADA	0.8561	0.3611	0.6507	0.4568
NB	0.6645	0.7405	0.3137	0.4365
XGB	0.8570	0.3222	0.6963	0.4325
GBC	0.8502	0.2935	0.6477	0.3989
SVM	0.8278	0.3611	0.4830	0.3987
LGBM	0.8512	0.2820	0.6629	0.3851
DT	0.7587	0.3680	0.3177	0.3368
CB	0.8580	0.2124	0.8139	0.3277
Ridge	0.8570	0.1886	0.8683	0.3000
ET	0.8424	0.1670	0.5683	0.2533
RF	0.8473	0.1216	0.8083	0.2037
QDA	0.6643	0.2578	0.2252	0.1910
KNN	0.8366	0.1141	0.5733	0.1885

Model Performance Reduced IBM Dataset

Model	Accuracy	Recall	Precision	F1
LR	0.8735	0.6587	0.8180	0.6947
LDA	0.8694	0.6612	0.8013	0.6955
ADA	0.8680	0.6591	0.7883	0.6924
NB	0.8143	0.6868	0.6739	0.6766
XGB	0.8551	0.6494	0.7518	0.6747
GBC	0.8646	0.6396	0.7912	0.6725
SVM	0.8388	0.5000	0.4194	0.4562
LGBM	0.8646	0.6548	0.7922	0.6853
DT	0.7844	0.6137	0.6138	0.6116
CB	0.8646	0.6347	0.7930	0.6684
Ridge	0.8592	0.5785	0.8626	0.5958
ET	0.8578	0.6186	0.7722	0.6479
RF	0.8585	0.6157	0.7855	0.6456
QDA	0.8272	0.6602	0.6802	0.6673
KNN	0.8381	0.5507	0.6746	0.5530

authors only state that low correlated features were removed. This underlines the previously discussed black box approach problematic in HR Analytics.

Nevertheless, it can be assumed that the feature selection process in this work was stricter and with greater attention to detail, as it was based on literature as well as two feature selection methods.

No parameter tuning and only simple pre-processing steps (refer to the methods section) were employed for this work to adapt the working steps as closely as possible to the initial study of Wang and Zhi (2021). Therefore, the improvements can be attributed to thorough feature selection and the detection of the most important variables for predicting employee turnover.

In terms of accuracy, both the original work from Wang and Zhi (2021) as well as this study perform similarly and range, except for a few exceptions such as the Decision Tree model, Naive Bayes and Quadratic Discriminant Analysis, in the area of 0.8143 to 0.8755. The lower performance of the Decision Tree can be attributed to its tendency for overfitting, which may lead to poor generalization on new data. Additionally, Decision Trees are often ineffective in handling complex patterns, which is why Random Forests are preferred over Decision Trees in certain cases. The slightly poorer performance of Naive Bayes and Quadratic Discriminant Analysis in the study of Wang and Zhi (2021) may be explained to the violation of their underlying assumptions such as feature independence and normality.

Regarding the accuracy, it must be noted that especially in imbalanced datasets, like the IBM dataset, accuracy can be a misleading measure. An explanation for this fact is, that

models which simply predict the majority class most of the time can achieve high accuracy, but may perform poorly on the minority class, which in this case is the class of interest —namely, whether an employee left the organization. It is therefore important to consider the F1 score, which is the harmonic mean of precision and recall and is an effective measure for imbalanced datasets.

When looking at the F1 scores, the models in this work outperformed the models of Wang and Zhi (2021) in every instance, with most F1 scores, except for the SVM, ranging between 0.5530 and 0.6955. Regarding the best-performing models, in the work of Wang and Zhi (2021), the top five models were Logistic Regression, Linear Discriminant Analysis, Ada Boost Classifier, Naive Bayes, and Extreme Gradient Boosting. In this work, the five best-performing models were similar but not exactly the same: Linear Discriminant Analysis, followed by Logistic Regression, Ada Boost, Light Gradient Boosting Machine, and Naive Bayes performed best.

This consistency in top-performing models indicates that these are generally well-suited algorithms for this classification problem, are robust and effective across different datasets and contexts. The difference between LGBM and XGB in the comparison of the five models across studies can be considered as little significant, as XGB is the sixth best-performing model in our work and LGBM is the eight in the study of Wang and Zhi (2021).
An interesting difference, however, is the SVM, which is in seventh place in Wang and Zhi's experiment but scores worst in our study. Most likely, the poor performance of the SVM can be attributed to default parameter settings not being optimal for the selected features. The parameters 'C' and 'gamma', as well as the selected kernel, play an important role in the performance of a Support Vector Machine. As no tuning was performed to replicate Wang and Zhi's work in the most accurate way possible and no mention of tuning is made in their work, the performance was quite poor. This explanation is further supported by the fact that in a later model where hyperparameter tuning was performed, the Support Vector Machine performed much better (refer to results in table 7).

In general, it can be noted that feature selection and the reduction of the dataset to the most important features can significantly enhance the performance of models, as shown by this study. However, the unavailability of the complete process of pre-processing and parameter tuning in the original study by Wang and Zhi (2021) makes it difficult to make

definite assumptions, thus highlighting the need for an open science approach in scientific literature in general, and specifically in research on HR Analytics.

Model Tuning on Reduced Dataset

In the first step of this work, a basic version of the models was employed without parameter tuning or balancing of classes to provide a direct comparison with the models presented in the work of Wang and Zhi (2021).

Many of the analysed studies in this review adopted a "black box" approach, where the focus is primarily on deploying models without giving too much attention to the optimization of these processes. The aforementioned limitation will be addressed by treating the class imbalance of the dataset and optimising the parameters of the individual models using a parameter grid search. For the grid search, different grid sizes were tried, and for balancing computational costs with optimal results, the final grid consisted of 2-3 parameters per model. The grid search was cross-validated to ensure robust performance. To address the imbalance of the dataset, the commonly used SMOTE technique was employed. Also, scikit-learn's class weighting function was successfully used for performance metric calculations.

In table 7 the results of the tuned model can be found.

Table 7: Enhanced Model Performance

Model	Accuracy	Precision	Recall	F1
LR	0.7701	0.8126	0.7701	0.7868
LDA	0.7701	0.8105	0.7701	0.7862
ADA	0.7966	0.8307	0.7966	0.8093
NB	0.7034	0.8162	0.7034	0.7383
XGB	0.8170	0.8161	0.8170	0.8162
GBC	0.8156	0.8146	0.8156	0.8145
SVM	0.7728	0.8128	0.7728	0.7885
LGBM	0.8299	0.8210	0.8299	0.8246
DT	0.7469	0.7829	0.7469	0.7618
CB	0.8122	0.8167	0.8122	0.8140
Ridge	0.7701	0.8105	0.7701	0.7862
ET	0.8095	0.8100	0.8095	0.8095
RF	0.8150	0.8058	0.8150	0.8093
QDA	0.7789	0.8133	0.7789	0.7927
KNN	0.6986	0.7801	0.6986	0.7293

When comparing the results of the initial models with the tuned models, it is evident that accuracy worsened in almost every case. This deterioration may be attributed to the imbalance of the dataset, where simply predicting the majority class was able to achieve good accuracy results.

Overall, precision, recall, and F1 scores significantly improved in many instances. Precision generally improved after tuning, particularly in models like NB (0.6739 to 0.8162), ADA (0.7883 to 0.8307), and KNN (0.6746 to 0.7801). Also, F1-scores improved overall after tuning and balancing. Here, major improvements can be seen in models like XGB (0.6747 to 0.8162), ADA (0.6924 to 0.8093), and LGBM (0.6853 to 0.8246). At this point it was noticed that accuracy and recall consist of the same values, which can be explained through the calculation of the weighted recall (refer to Sklearn library's documentation). Even though recall did not increase to the same extent as precision and F1-score, it still improved in all instances, which is especially important in the context of attrition prediction. The detection of as many potential leavers as possible (recall), allows the employer to engage in proactive interventions to reduce attrition. High precision is also desirable; however, the cost of false negatives (missed employees who leave) usually outweighs the cost of false positives (unnecessary interventions) and thus it is important to minimize false negatives (increase recall).

In general, the tuning and balancing increased the performance of all our models, making them helpful tools for accurate prediction purposes.

Figures 9 and 10 show the ROC curve and confusion matrix for the LGBM model, which had the highest improvement, comparing the untuned and tuned versions. The AUC score for the untuned model is higher, which indicates better overall performance. However, this higher AUC score could again be attributed to the untuned model's tendency to overfit the training data and by predicting the majority class.

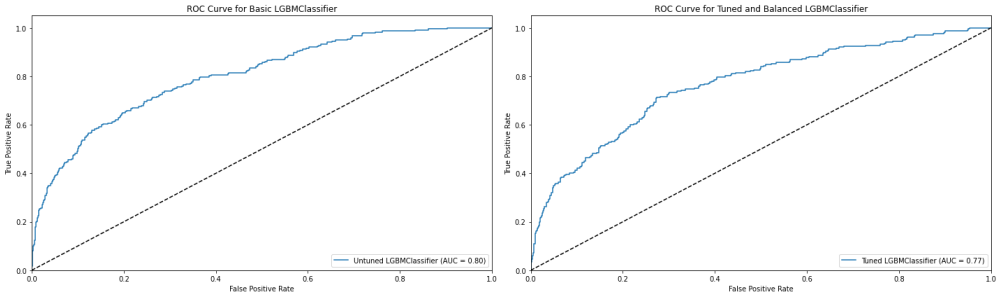

Figure 9: ROC Curves for LGBM

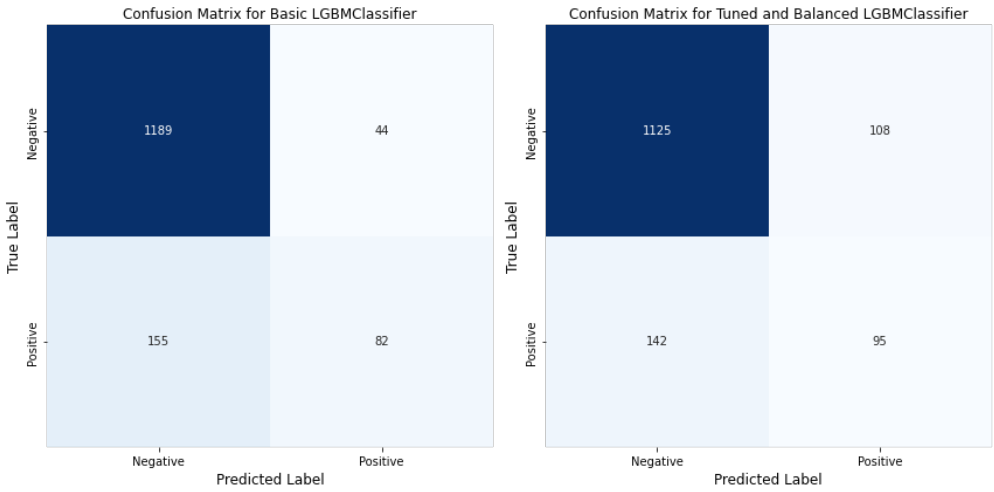

Figure 10: Confusion Matrices for LGBM

When looking at the confusion matrix, it can be seen that the untuned model has a high number of true negatives, which suggests it performs well at identifying employees who are not at risk of leaving. However, the 155 false negatives indicate that 'leavers' were not correctly identified. On the contrary, the tuned model has a slightly lower number of true negatives and more false positives. Crucially, it has fewer false negatives, which results in a higher recall. This is, as discussed previously, especially important in employee turnover prediction, suggesting that tuning the model and addressing class imbalance improved its performance on important metrics for attrition prediction.

The IBM Watson dataset does not specify a particular time frame that it is intended to represent. For the purposes of illustrating the costs of false negatives and false positives, it is assumed below that the numbers in the confusion matrix can be understood as annual results. As mentioned above, false negatives refer to employees who are at risk of leaving the organisation but who are incorrectly identified by the model as not at risk. This can be seen as critical, as the company incurs unanticipated costs in recruiting and training a replacement, as well as lost productivity due to absence. Had the potential departures been correctly identified, intervention could have taken place. Assuming an annual time horizon for the model, the untuned model does not identify 155 employees who will leave the company each year and therefore cannot implement retention strategies. With the tuned model, this number is slightly reduced to 142, which also means that 152 or 142 potential talents are lost, which could also lead to a delay in projects or a general weakening of the company's

image. Using the same annual assumption for false positives, the untuned model incorrectly predicts 44 employees per year who are at risk of leaving but do not, which could lead to unnecessary interventions such as bonuses or promotions. The tuned model increases this number to 108 incorrectly predicted employees. These false positives can therefore lead to a misallocation of resources each year and have a significant impact on company budgets. It could also lead to dissatisfaction among employees who do not receive bonuses or promotions because they feel the actions are unfair. In conclusion, as discussed above, the cost of false negatives exceeds the cost of false positives unless huge sums of money are spent on retention strategies. In this respect, attention should be paid to keeping false negatives low, as is the case with the tuned model.

5 Discussion

General Findings: The objective of this thesis was to analyze data types used to predict employee turnover and to reduce a commonly used dataset to key features in order to then compare its performance with that of the original.

The analysis of the data types seeks to address a gap in HR research, specifically the lack of clarity regarding the data types used for prediction and the challenge of counteracting the "black-box" approach in HR Analytics. It is important to note here that, as assumed at the beginning of this thesis, most of the studies analysed in the literature review do not provide precise information on the data or data types used for prediction purposes. Often, only a few features that are utilized are mentioned by way of example or – in some cases - the industry from which the data originates is briefly described. In addition, probably due to the sensitivity of HR data, many studies rely on the same datasets - in particular two that are freely available to the public: The dataset from IBM Watson, which was also used for the reduction in the second part of this paper, and another one available on Kaggle.

This reliance can be explained by the difficulty of obtaining HR data, as it is sensitive in nature and suffers from potential to misuse. The GDPR, which applies in Europe and to organizations outside Europe that process personal data of individuals within the EU, plays a significant role here. However, the difficulty in availability cannot solely be explained through the sensitivity of the data, especially since studies from countries which lack strict data protection rules also focus on the publicly available datasets. This may be attributed to the scientific benefits of using these datasets. Piwowar and Vision (2013), for example, have shown that research with publicly available datasets receive more citations and foster transparency and replicability.

A second observation was that the data used is similar across many studies. For example, data recurring in literature, such as salary, promotion opportunities, satisfaction, years with the company, distance from home or department.

A major goal of research in HR analytics is the automation of processes that were previously laborious to complete by hand, as well as the possibility of making predictions in order to reduce costs, optimise profits and retain employees in the long term. Specifically, the prediction of employee turnover, which is the focus of this work, should enable companies to

recognize potential departures at an early stage and implement countermeasures.

Not all reasons for turnover can be prevented, though. For example, an employee might seek a career change, wants to reorient themselves (Mozaffari et al., 2023) or retire. However, other reasons can be addressed by the company through the implementation of retention strategies. In order to do this, at-risk employees and the factors that motivate their departures must be analysed first. Where turnover can be prevented, the employee should be persuaded to stay with the company through incentives such as salary increases or reduced working hours. If retention is not feasible, internal company knowledge should be secured, for example by the departing employee passing it on to their replacement before they leave.

Many of the studies analysed make recommendations as to what could lead to effective employee retention. Wang and Zhi (2021), for example, see the possibility of applying the framework they have developed to other decision-making processes within different areas of the company. The corporate sector also seems to play a role in the area of turnover prediction. Zhu et al. (2019) found that the IT sector was most affected by turnover, whereas the government sector was less affected. Yuan also identifies different turnover rates by sector. AbdElminaam et al. (2023) concluded that the prediction of turnover helps to start timely retention measurements and thus reduce the costs associated with turnover. Gao et al. (2019) argued that the most effective way to reduce turnover is to increase salaries. In addition, predictions make it possible for responsible HR people to find timely solutions in the form of retention or rapid replacement (Gao et al., 2019). El-Rayes et al. (2020) also sees salary as one of the most important factors. In addition El-Rayes et al. (2020) also recognises that higher Glassdoor ratings contribute to higher retention. However, it is questionable whether it is the good ratings themselves that contribute to retention, or whether the same factors that lead to good ratings also encourage employees to stay with their companies.

In particular, the department variable, which has been shown to be important by various studies (Abdullah et al., 2023; Yuan, 2021; Zhu et al., 2019), should make companies realise that retention strategies may not be necessary company-wide, but rather target certain departments. This targeted approach is also supported by Goswami and Jha (2012), who differentiated between different management levels with distinct retention needs.

This thesis demonstrated that meaningful feature selection and hyper-parameter tuning can be employed to enhance predictive performance. This approach serves not only to en-

hance the reliability of predictions but also to play a critical role in counteracting the "black box" phenomenon. As previously observed by Kuhn and Johnson (2019) and Yahia et al. (2021), the careful selection of features is of great importance. Moreover, it was demonstrated that utilising a mixed-methods approach to feature selection ensures the incorporation of the most significant variables, thereby enhancing interpretability. Furthermore, hyper-parameter tuning optimises the performance of the model further. It is important to note, however, that in a practical context, the application of careful feature selection and tuning—while enhancing reliability and accuracy—can present certain challenges. From one perspective, the financial implications are considerable and the process is frequently impractical. Furthermore, managers and employees frequently lack the requisite skills and knowledge to fully comprehend the models, which adds another layer of complexity to the selection process. It is therefore essential that domain experts are involved when utilising such models. It is only when these issues are addressed that the challenges posed by opaque models, particularly in relation to careful feature selection, can be effectively addressed.

Improving prediction accuracy through, for example, meaningful selection of features and hyper-parameter tuning, as in this study, cannot only help to counteract the generally prevailing black box approach in HR Analytics (Kuhn, Johnson, et al., 2013), but also, when considering the costs of employee attrition, reduce the temporal, monetary and human resource costs of companies caused by attrition (Chung et al., 2023). Similarly, Goswami and Jha (2012) sees high attrition, which can be counteracted by timely prediction, as a systematic cycle. Attrition would lead to low productivity and thus to an increase in the workload of employees who stayed with the company, which in turn leads to more attrition.

In general, attrition, as described in more detail at the beginning of this thesis, is associated with increased costs and reduced productivity. Costs because talented employees may leave the company and take important knowledge with them. Additionally, there is often a delay before the position is filled again. Employee prediction therefore offers the opportunity to recognise the intention of employees to leave the company in good time and to counteract this. Countermeasures can include passing on knowledge to new employees so that no knowledge is lost or offering incentives which may convince employees to stay with the company. Machine prediction methods are particularly suitable for this, as the intention to leave the company can be predicted easily and favourably. Without automatic

machine prediction, HR managers and bosses would have to hold regular feedback meetings, which are time-consuming and cost-intensive and may not lead to the desired results due to social desirability. Also, the assessment of turnover intentions by managers is a traditional method, but this approach is often costly and not always accurate. In order to persuade employees to remain in their posts once the intention has been identified through traditional means or by means of a machine learning model, it is necessary to implement appropriate retention measures. The analysis of the literature and the developed model indicates that factors such as salary or related promotion and various satisfaction variables appear to be significant. Consequently, for companies to retain their employees, it is essential that they offer competitive salaries which are adjusted on a regular basis. Hansen (2007) also identifies a correlation between competitive salaries and an increase in the number of talented employees. Conversely, talented employees are attracted by good salaries. With regard to the proposition that satisfaction is a key factor in employee retention, a number of studies can be consulted. Skelton et al. (2020), for example, finds that satisfied and committed employees are less likely to leave their job, while Lee et al. (2018) finds that in the 21st century there has been a generational shift and that nowadays factors such as purpose development and satisfaction are more likely to make people stay in a job.

Employee Perspective: The role of employees is also central to the discussion about predicting employee turnover. In general, employees initially exhibit a certain degree of apprehension regarding the adoption of HR analytics. They tend to be hesitant to utilise it themselves and perceive themselves as lacking the requisite proficiency to effectively engage with it (Saxena et al., 2022). However, this hesitation does not mean that they oppose the adoption of HR analytics per se. A closer look of the impact of HR analytics, particularly in the context of employee turnover, reveals a number of potential risks for employees (Tursunbayeva et al., 2022). First, there is the potential for bias or discrimination. For instance, the case of the Amazon recruiting tool, briefly referenced in the theoretical part of this thesis, is a well-documented example of a tool that exhibited a bias, specifically a preference for male applicants. To overcome this issue, human oversight of datasets is essential to catch potential biases that machines might miss. Another option could be, to ask for active feedback from employees when it comes to their data to identify and address potential biases or

discriminatory practices. It must be acknowledged that this option requires a certain degree of proficiency with datasets, which limits its applicability.

Moreover, the analytical measurement of, for instance, employee performance, which is a commonly utilised metric for employee turnover, reduces the employee to a mere number (Tursunbayeva et al., 2022). In addition, in the case of automated prediction models, measuring employee performance solely on the basis of one metric can neglect important aspects that are difficult to measure or neglect context (O'Neil, 2016; Tursunbayeva et al., 2022). Furthermore, employees might be concerned about how their data is being used. The feeling of being monitored can lead to anxiety and dissatisfaction. For example, constant logging of performance data and supervisors' access to it can create a sense of surveillance, which can have a negative impact on job satisfaction and mental health (Booth, 2019; Tursunbayeva et al., 2022). As job satisfaction is a significant predictor of employee turnover, monitoring would, in fact, have a negative impact on turnover rates overall.
It is therefore crucial to provide employees with a clear and comprehensive explanation of how their data is collected and used, in order to break down barriers. Also, employees need to be empowered with control over their personal data, such as being given the option to opt out of data collection. Moreover, the establishment of official channels would provide employees with a platform to express their concerns, which could lead to their resolution.

Adaptive Retention Strategies and HR Analytics: Evaluating and refining retention strategies allows organisations to proactively address employee turnover, while at the same time remaining flexible to change. This raises the question of the potential of HR analytics for continuous evaluation and adjustment of employee retention strategies. As previously discussed in the theoretical part of this thesis, the retention of talented employees is of importance to the success of companies, particularly in terms of the costs associated with the recruitment of new employees and the potential loss of knowledge. However, the algorithms used to predict employee turnover usually rely on static data, which makes it difficult to capture the dynamic nature of employee behaviour. This may result in a challenge for predicting turnover accurately, as a model based on static data fails to account for the dynamic changes of human behaviour.
In the area of retention strategies, it is especially important to continually adapt and see

which strategies are most effective. One example of a retention strategy that could benefit from continuous evaluation and adaptation is the Personal Development Plan (PDP). These plans involve planning the employee's career aspirations and future through action plans and linking them to the needs of the organisation. In contrast to compensation-based retention strategies, which alone are insufficient for long-term retention, PDPs can foster job embeddedness and subsequently lead to effective retention (Mitchell et al., 2001). Since PDPs are more sensitive to changes in employee sentiment, such as dissatisfaction and frustration than, for example, compensation-based factors, they are well-suited candidates for adaptive retention strategies. For example, they could be updated dynamically based on real-time feedback and performance metrics. If an employee expresses dissatisfaction or his or her performance falls below a certain level, a personalised career development and retention plan could be deployed.

However, this would require the development of systems that continuously adapt to real-time data. Nevertheless, many predictive models are based on historical and static data, such as salary or demographic data, which can change rapidly. In contrast, dynamic data sources - such as real-time employee feedback, performance trends and even external factors such as industry trends - could enable organisations to predict attrition more accurately. Anyhow, the development of such adaptive systems is limited by current technical and data collection challenges. One major challenge is that not all critical features can be collected automatically. Key predictors, such as employee satisfaction and organisational climate, are difficult to assess in real-time, which hinders updates and real-time predictions. Therefore, even if real-time data systems were developed and implemented, certain features would remain difficult to be measure timely. Therefore, to advance continuous and adaptive retention strategies, more research in automated data collection is required as well as a focus on how to employ data in a timely manner.

Accuracy of Models: This work focused on the accuracy of employee turnover prediction models, particularly in relation to the use of a reduced data set with feature selection. Comparing the results of this study with those of Wang and Zhi (2021) reveals significant improvements, particularly in terms of recall, precision, and F1 score. The accuracy on the other hand, remained relatively similar. As noted earlier in this thesis, accuracy can be

misleading, especially for unbalanced datasets like this one. A model that simply predicts the majority class while neglecting the minority class can still achieve high accuracy. However, this does not indicate the model's ability to identify actual leavers, which is crucial for turnover prediction.

This underlines the need for using additional metrics for performance measuring. The F1 score, for example, provides a balanced representation between recall and precision and therefore offers a more comprehensive model assessment. Additionally, model performance is often highly dependent on the algorithm used and even with optimised parameters through grid search, the results remain sensitive to the chosen parameters.

Another factor influencing model performance is the imbalance of the dataset. The target variable, which determines how many people have left the organisation, consists of about 16% leavers, which makes sense, as turnover rates are usually low, can influence model performance. In the second run, where models were tuned, an attempt was made to counteract this problem using the SMOTE technique.

This resulted in a significant increase in recall for nearly all models. This is particularly important for the prediction of employee turnover, recall measures the model's ability to identify employees who are likely to leave. For example, Logistic Regression and Linear Discriminant Analysis not only achieved some of the highest accuracies in the reduced model but also scored high on F1. Naive Bayes (NB) achieved the highest recall from the untuned models, but relatively low precision compared to other models. This can be attributed to the characteristics of NB, as its probabilistic approach tends to overestimate the minority class.

Logistic regression, on the other hand, performed well in the tuned and untuned models and showed robustness against overfitting. Linear discriminant analysis also worked effectively in these cases, due to the clear separation of classes and the normally distributed features, which were ensured during the data preprocessing phase. After parameter tuning using grid search and the application of the SMOTE technique, improvements in F1 scores were observed for most algorithms, especially for NB, AdaBoost and LightGBM.

Despite these improvements, it is important to note that the accuracy decreased for most models after tuning. However, this underscores the argument that in turnover prediction, recall should be considered the most important metric. Prioritizing recall ensures that the

model is better at identifying those employees who are at risk of leaving, which is crucial in real-world applications. However, it is important to notice that the most reliable metric depends largely on the specific use case.

Lastly, the real-world applicability of these models must be considered. As previously mentioned, the results are heavily dependent on data preprocessing, the specific dataset, the individual algorithms and their parameters. Thus, their transferability to real world scenarios needs to be seen with caution. While these models may work well in a controlled setting,heir effectiveness in practical application cold be affected by noise, biases or changes that were not captured. Therefore, careful validation by experts is crucial, as well as awareness of their limitations and the ongoing fine-tuning when applied in real business environments.

Challenges in Automated Turnover Prediction: In practice, though, the automated prediction of turnover, as desired by many companies, encounters several challenges. On the one hand, after analysing the data types used to predict turnover, it becomes evident that some of the data types used are not automatically accessible to companies. On the other hand, data protection regulations limit the usage of available data. First, many of the variables with a high predictive value are not automatically available. For example, the important variables remaining in the reduced IBM Watson, such as job involvement, relationship satisfaction and environment satisfaction, are of great value, but must first be painstakingly collected by the employer or other responsible bodies via surveys. Surveys are cost-intensive, often ineffective and are inherently opposed to automated prediction. Second, the GDPR might limit the implementation of predictive models. The GDPR consists of seven principles: lawfulness, fairness, and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality and accountability.
Different variables found in the studies can be affected to varying degrees by these principles. In general, it can be stated that some theoretical points outlined by the GDPR principles could lead to difficulties in their practical implementation. For example, data minimisation involves only collecting and processing the data necessary for the intended purpose. In practice, this raises the question of which data can be considered necessary. Ethnicity, for instance, is processed by Zhao et al. (2019) in their prediction task and can therefore be considered necessary. However, it is not utilized in other studies, so the question arises as

to whether a general necessity can be argued.

For variables such as salary, promotion in last year, the principles of lawfulness and data minimisation are generally fulfilled. In order to also fulfill the principle of transparency, employees should be informed that and how their data will be used for predictions. Whether data such as mission cost, which indicates the remuneration for a mission, complies with the principle of data minimisation can again be difficult to assess and must be considered in the respective context.

Variables from the demographic data category are generally more problematic in their consideration. Variables such as age, gender, marital status, ethnicity, whether children are present or not as well as an employee's physical condition could conflict with the principle of fairness, as they could potentially lead to discrimination or misuse. Therefore, these data may only be collected if they have a clear purpose, the data subjects have been informed and have given consent and the data are deleted after use. The principle of storage limitation also applies to data related to job tenure. According to the GDPR, data such as years at company should only be stored for as long as necessary (e.g. for salary and pension statements) and then deleted. Data such as last contract duration could in turn lead to discrimination if used incorrectly.

Variables from the education and experience category could affect the principle of accuracy in particular, as these variables can change, and it is crucial to ensure that they are up to date. Especially for variables in the last two categories - satisfaction and organisational climate - it is important that they are treated confidentially and that employees do not fear any disadvantages from disclosing them.

In conclusion, despite its importance, adhering to all GDPR principles when predicting turnover presents several challenges. The complexity of implementing GDPR compliance in practice, especially with regard to data minimization, purpose limitation and fairness, can limit the effectiveness of turnover prediction models or make them cost intensive. As a result, some companies may choose to avoid such models due to the potential legal and ethical consequences and considerations, despite the benefits they could provide in employee retention.

Third, the development of predictive models is not yet sufficiently advanced. Fernandez and Gallardo-Gallardo (2021), for example, found that HR analytics in general is still at

an early stage and that existing data is sometimes insufficient for its purposes. Meddeb et al. (2022) is also more critical of the pure use of machine learning models. According to them, models and HR managers often come to different conclusions, which is why many HR managers do not trust machine learning models and do not use them. If they are to be used, then their usage should be integrated with human knowledge.

In summary, predicting employee turnover with the help of machine algorithms has great potential, but also some pitfalls. So, if predictive models are being employed, special attention must be paid to protecting the personal data of employees, using the models together with the help of employee managers and using them for targeted areas or departments. Only if they are used in targeted departments only instead of companywide can shortcomings such as the cost of collecting satisfaction variables and the help of HR managers in analysing them make their use cost-effective and the implementation work.
With regard to HR analytics research, it can be stated that research in this area is still in its infancy, but that more research is currently being carried out in this area, so that the black box approach will decline in the near future.

Limitations and Future Research

The first limitation of this thesis is the partial lack of data description of the analysed studies. Even though this thesis was actually concerned with counteracting the black box approach in HR and conducting a review of studies that describe their datasets, some of the studies analysed did provide their dataset in full but did not fully describe the data types. This makes it difficult to make precise statements about the data used in predictive modelling and it also means that this work itself could not always provide fully transparent data. Furthermore, this problem also came to bear in the modelling part of the work. Here, too, more detailed information from the original study was unfortunately missing, which made a comparison with this study difficult in some cases.

Given the limitations identified, future research should aim to improve the transparency and comprehensiveness of data descriptions in HR analytics. Specifically, there is a need for datasets in published research to be fully specified and clearly described. This would not only enable a clearer understanding of the variables used in predictive models and counteract the challenges of the black box approach, but also improve the reproducibility of the results

of study results.

In HR, this lack of transparency is problematic as decisions based on opaque models can affect employee trust and lead to ethical and legal issues, especially when it comes to personal data. Future research should therefore aim to ensure that ML models and their associated datasets are fully explained and specified. A further step would be to make AI models transparent and explainable. A promising approach, for example, would be Explainable AI (XAI). XAI techniques aim to explain the influencing factors behind AI so that users can understand which data was used and which factors led to a decision. Techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) could be used to show the importance of different features and provide insights into how individual predictions are made. These techniques would be especially beneficial for HR applications like employee turnover prediction, where transparency and interpretability are essential.

Another way to increase transparency would involve documenting and logging the data. This approach would not only enhance clarity about which data has been utilised but also establish a framework for data transparency. In addition, instead of simply stating which characteristics were used (e.g. 'age', 'job satisfaction'), future research should provide justifications for each variable included in predictive models. This could avoid variables being used solely on the basis of their availability, rather than their predictive value.

In terms of GDPR compliance, several possibilities should be considered. As many datasets are owned by private companies, research organisations could collaborate more closely with these to facilitate access to diverse datasets. In practice, private companies could share their datasets in aggregated and anonymised forms in exchange for financial compensation, so that individual identities are protected, while still enabling the use of different datasets for research purposes.

In terms of ethical policies, employees need to be clearly informed of their rights and familiarised with the GDPR. Transparent, clear and mandatory consent processes should be implemented across all companies and organisations to inform employees about how their data is being used and providing them the opportunity to opt out. This would not only alleviate employees' scepticism regarding new data-driven approaches but also provide protection of personal data. Furthermore, an establishment of clear ethical guidelines is

needed for both legal compliance but also to guarantee fairness and transparency.

To make HR analytics more attractive and practical for companies, it is essential to address both the legal and ethical concerns, as well as the challenges posed by the black box nature of AI models. Only when these conditions are met, can predictive analytics thrive and contribute to the future success of HR practices.

6 Conclusion

The objective of this study was to analyse the data utilised to predict employee turnover, addressing the so-called 'black box' problem of AI in human resources. Furthermore, the study investigated the impact of reducing a well-known and frequently utilised dataset in this field on the performance of machine learning algorithms. The analysis yielded several findings regarding the data used for predicting employee turnover. Firstly, a significant number of the studies analysed were found to rely on the same datasets. Moreover, even when different datasets were used, many of the variables in the different datasets were identical. For example, of the 148 studies subjected to full-text analysis, 82 employed either the IBM Watson dataset or another publicly accessible Kaggle dataset. The 500 variables identified in 30 studies were reduced to 286 distinct variables after the removal of duplicates. However, variables that were conceptually identical or similar were not yet excluded. Moreover, the variables could be classified into categories that were conceptually related. It was not always feasible to assign a variable to a single category, particularly given that many variables lacked explicit definitions in the associated studies and had to be inferred. The automated prediction of employee turnover, as a goal of HR analytics, is further hindered by the fact that many of the analysed variables cannot be collected automatically. Variables pertaining to employee satisfaction or organisational climate, for instance, may serve as significant predictors; however, they are not readily accessible to employers and frequently necessitate manual collection through surveys or interviews. This raises concerns about the cost-effectiveness and practicality of using such variables for automated turnover prediction. The second part of the thesis examined the impact of reducing the dataset using a mixed-methods approach to feature selection on the predictive power of machine learning algorithms. The analysis revealed that meaningful feature selection led to enhanced performance. Precision, recall, and

F1 scores demonstrated improvement across the majority of algorithms when compared to the initial study. Further performance gains were achieved by tuning the algorithms, which, despite a slight decline in accuracy, resulted in enhanced precision, recall, and F1 scores. In the context of employee turnover prediction, correctly identifying potential leavers is of paramount importance for the implementation of preventive measures, making a high recall rate particularly crucial. In conclusion, this study provided a comprehensive analysis of the data used to predict employee turnover. However, further work is required to enhance transparency in studies and mitigate the black box problem in HR. Additionally, the selection of significant features from a dataset could significantly enhance the performance of machine learning algorithms.

References

Abbasi, S. M., & Hollman, K. W. (2000). Turnover: The real bottom line. *Public personnel management*, *29*(3), 333–342.

AbdElminaam, D. S., Maged, M., Mousa, M. K., Younis, A. O., Abdelsalam, M. S., Hisham, Y., & Talaat, T. (2023). Empturnoverml: An efficient model for employee turnover and customer churn prediction using machine learning algorithms. *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 1–8.

Abdullah, A. S., Kailash, P. J. S., Ramesh, D., & Guntha, P. (2023). Evaluating employee attrition and its factors using machine learning approaches. *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, 1–11.

Aggarwal, S., Singh, M., Chauhan, S., Sharma, M., & Jain, D. (2022). Employee attrition prediction using machine learning comparative study. *Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2021*, 453–466.

Al-Amin, M., Alam, S., & Hossain, S. (2020). Analysis of the present condition of garment workers' rights and its protection under domestic and international legal framework: Bangladesh perspective. *Br. J. Arts Humanit*, *2*(6), 101–112.

Alamsyah, A., & Salma, N. (2018). A comparative study of employee churn prediction model. *2018 4th international conference on science and technology (ICST)*, 1–4.

Allen, N. J., & Meyer, J. P. (1993). Organizational commitment: Evidence of career stage effects? *Journal of business research*, *26*(1), 49–61.

Alshehhi, K., Zawbaa, S. B., Abonamah, A. A., & Tariq, M. U. (2021). Employee retention prediction in corporate organizations using machine learning methods. *Academy of Entrepreneurship Journal*, *27*, 1–23.

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., et al. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46.

Amabile, T. M. (1996). *Creativity and innovation in organizations* (Vol. 5). Harvard Business School Boston.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. *Handbook of reading research*, *1*, 255–291.

Argote, L., & Ingram, P. (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational behavior and human decision processes*, *82*(1), 150–169.

Augusto, D. A., Bernardino, H. S., & Barbosa, H. J. (2013). Predicting the performance of job applicants by means of genetic programming. *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*, 98–103.

Backhaus, K., & Tikoo, S. (2004). Conceptualizing and researching employer branding. *Career development international*, *9*(5), 501–517.

Bandyopadhyay, N., & Jadhav, A. (2021). Churn prediction of employees using machine learning techniques. *Tehnički glasnik*, *15*(1), 51–59.

Bentum, E. v. (2023). *Kennzahlengestütztes hr-risikomanagement*. Erich Schmidt Verlag GmbH & Co. KG.

Bondarouk, T., Ruël, H., & van der Heijden, B. (2009). E-hrm effectiveness in a public sector organization: A multi-stakeholder perspective. *The International Journal of Human Resource Management*, *20*(3), 578–590.

Booth, R. (2019, April 7). *Uk businesses using artificial intelligence to monitor staff activity* [Accessed: 2021-09-17]. https://www.theguardian.com/technology/2019/apr/07/uk-businesses-using-artificial-intelligence-to-monitor-staff-activity

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic reviews*, *6*, 1–12.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101.

Budhwar, P., Malik, A., De Silva, M. T., & Thevisuthan, P. (2022). Artificial intelligence–challenges and opportunities for international hrm: A review and research agenda. *The InTernaTIonal Journal of human resource managemenT*, *33*(6), 1065–1097.

Chakraborty, R., Mridha, K., Shaw, R. N., & Ghosh, A. (2021). Study and prediction analysis of the employee turnover using machine learning approaches. *2021 IEEE 4th*

International Conference on Computing, Power and Communication Technologies (GUCON), 1–6.

Chambers, E. G., Foulon, M., Handfield-Jones, H., Hankin, S. M., & Michaels III, E. G. (1998). The war for talent. *The McKinsey Quarterly*, (3), 44.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & electrical engineering*, *40*(1), 16–28.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*, 171–209.

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through ai capability framework. *Human Resource Management Review*, *33*(1), 100899.

Chung, D., Yun, J., Lee, J., & Jeon, Y. (2023). Predictive model of employee attrition based on stacking ensemble learning. *Expert Systems with Applications*, *215*, 119364.

Das, B. L., & Baruah, M. (2013). Employee retention: A review of literature. *Journal of business and management*, *14*(2), 8–16.

Dastin, J. (2022). Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296–299). Auerbach Publications.

Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. *Harvard business review*, *88*(10), 52–58.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business Press.

Dobbs, R., Madgavkar, A., Barton, D., Labaye, E., Manyika, J., Roxburgh, C., Lund, S., & Madhav, S. (2012). *The world at work: Jobs, pay, and skills for 3.5 billion people* (Vol. 28). McKinsey Global Institute Washington.

Dolatabadi, S. H., & Keynia, F. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor. *2017 2nd international conference on computer and communication systems (ICCCS)*, 74–77.

Duan, Y. (2022). Statistical analysis and prediction of employee turnover propensity based on data mining. *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, 235–238.

Elnaga, A., & Imran, A. (2013). The effect of training on employee performance. *European journal of Business and Management*, *5*(4), 137–147.

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. *International Journal of Organizational Analysis*, *28*(6), 1273–1291.

Esmaieeli Sikaroudi, A. M., Ghousi, R., & Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of industrial and systems engineering*, *8*(4), 106–121.

Faliagka, E., Ramantas, K., Tsakalidis, A., & Tzimas, G. (2012). Application of machine learning algorithms to an online recruitment system. *Proc. International Conference on Internet and Web Applications and Services*, 215–220.

Falletta, S. V., & Combs, W. L. (2020). The hr analytics cycle: A seven-step process for building evidence-based and ethical hr analytics capabilities. *Journal of Work-Applied Management*, *13*(1), 51–68.

Fernandez, V., & Gallardo-Gallardo, E. (2021). Tackling the hr digitalization challenge: Key factors and barriers to hr analytics adoption. *Competitiveness Review: An International Business Journal*, *31*(1), 162–187.

Fitz-enz, J. (1984). *How to measure human resources management*. McGraw-Hill.

Foster, S. P., & Dye, K. (2005). Building continuity into strategy. *Journal of corporate real estate*, *7*(2), 105–119.

Gao, X., Wen, J., Zhang, C., et al. (2019). An improved random forest algorithm for predicting employee turnover. *Mathematical Problems in Engineering*, *2019*, 1–12.

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2022). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), 1590–1610.

Gelens, J., Dries, N., Hofmans, J., & Pepermans, R. (2013). The role of perceived organizational justice in shaping the outcomes of talent management: A research agenda. *Human Resource Management Review*, *23*(4), 341–353.

Goswami, B. K., & Jha, S. (2012). Attrition issues and retention challenges of employees. *International Journal of Scientific & Engineering Research*, *3*(4), 1–6.

Greenleaf, G. (2013). Sheherezade and the 101 data privacy laws: Origins, significance and global trajectories. *SSRN*.

Groß, M. (2021). Yes, ai can: The artificial intelligence gold rush between optimistic hr software providers, skeptical hr managers, and corporate ethical virtues. In *Ai for the good: Artificial intelligence and ethics* (pp. 191–225). Springer.

Gurler, K., Pak, B. K., & Gungor, V. C. (2023). Deep learning based employee attrition prediction. *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 57–68.

Hackman, J. R., & Oldham, G. R. (1974). The job diagnostic survey: An instrument for the diagnosis of jobs and the evaluation of job redesign projects. *Affective Behavior*, *4*, 87.

Hall, D. T., & Moss, J. E. (1998). The new protean career contract: Helping organizations and employees adapt. *Organizational dynamics*, *26*(3), 22–37.

Hana, U., & Lucie, L. (2011). Staff turnover as a possible threat to knowledge loss. *Journal of competitiveness*, *3*(3), 84–98.

Hansen, F. (2007). Currents in compensation and benefits. *Compensation & Benefits Review*, *39*(3), 5–27.

Hennig-Thurau, T. (2004). Customer orientation of service employees: Its impact on customer satisfaction, commitment, and retention. *International journal of service industry management*, *15*(5), 460–478.

Herzberg, F. I. (1966). Work and the nature of man.

Huselid, M. A., & Day, N. E. (1991). Organizational commitment, job involvement, and turnover: A substantive and methodological analysis. *Journal of Applied psychology*, *76*(3), 380.

Ibidunni, S., Osibanjo, O., Adeniji, A., Salau, O. P., & Falola, H. (2016). Talent retention and organizational performance: A competitive positioning in nigerian banking sector. *Periodica Polytechnica Social and Management Sciences*, *24*(1), 1–13.

Iqbal, S., Guohao, L., & Akhtar, S. (2017). Effects of job organizational culture, benefits, salary on job satisfaction ultimately affecting employee retention. *Review of Public Administration and Management*, *5*(3), 1–7.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in hrm using c4. 5 classification algorithm. *International Journal on Computer Science and Engineering*, *2*(8), 2526–2534.

Johnson, J. T., Griffeth, R. W., & Griffin, M. (2000). Factors discriminating functional and dysfunctional salesforce turnover. *Journal of business & industrial marketing*, *15*(6), 399–415.

Johnson, R., & Kuhn, K. (2024). Information technology and human resource management: Revisiting the past to inform the future. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, 6300–6309.

Johnson, R. D., Lukaszewski, K. M., & Stone, D. L. (2016). The evolution of the field of human resource information systems: Co-evolution of technology and hr processes. *Communications of the Association for Information Systems*, *38*(1), 28.

Katz, D., & Kahn, R. (2015). The social psychology of organizations. In *Organizational behavior 2* (pp. 152–168). Routledge.

Khanzode, K. C. A., & Sarode, R. D. (2020). Advantages and disadvantages of artificial intelligence and machine learning: A literature review. *International Journal of Library & Information Science (IJLIS)*, *9*(1), 3.

Kim, H., & Park, J. (2017). The effects of longer commutes, unsolicited job offers, and working in the seoul metropolitan area on the turnover intentions of korean employees. *International Journal of Manpower*, *38*(4), 594–613.

Krämer, A., & Mauer, R. (2023). *Datenschutz für entscheider in marketing und vertrieb: Die dsgvo vom spielverderber zum wettbewerbsvorteil* (1st ed.). Springer Gabler.

Krausz, M., Yaakobovitz, N., Bizman, A., & Caspi, T. (1999). Evaluation of coworker turnover outcomes and its impact on the intention to leave of the remaining employees. *Journal of Business and Psychology*, *14*, 95–107.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman; Hall/CRC.

Kwon, K., & Jang, S. (2022). There is no good war for talent: A critical review of the literature on talent management. *Employee Relations: The International Journal*, *44*(1), 94–120.

Lawler, E., Levenson, A., & Boudreau, J. (2004). Hr metrics and analytics - uses and impacts. *Human Resource Planning Journal*, *27*(4), 27–35.

Lee, T. W., Hom, P., Eberly, M., & Li, J. J. (2018). Managing employee retention and turnover with 21st century ideas. *Organizational dynamics*, *47*(2), 88–98.

Loebbecke, C., Van Fenema, P. C., & Powell, P. (2016). Managing inter-organizational knowledge sharing. *The Journal of Strategic Information Systems*, *25*(1), 4–14.

Madanchian, M., Taherdoost, H., & Mohamed, N. (2023). Ai-based human resource management tools and techniques; a systematic literature review. *Procedia Computer Science*, *229*, 367–377.

Madden, L., Mathias, B. D., & Madden, T. M. (2015). In good company: The impact of perceived organizational support and positive relationships at work on turnover intentions. *Management Research Review*, *38*(3), 242–263.

Mallick, A. (2021). Application of machine learning (ml) in human resource management. *New Business Models in the Course of Global Crises in South Asia: Lessons from COVID-19 and Beyond*, 209–220.

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of hr analytics. *The International Journal of Human Resource Management*, *28*(1), 3–26.

Marvin, G., Jackson, M., & Alam, M. G. R. (2021). A machine learning approach for employee retention prediction. *2021 IEEE Region 10 Symposium (TENSYMP)*, 1–8.

McDonnell, A. (2011). Still fighting the "war for talent"? bridging the science versus practice gap. *Journal of business and psychology*, *26*, 169–173.

Meddeb, E., Bowers, C., & Nichol, L. (2022). Comparing machine learning correlations to domain experts' causal knowledge: Employee turnover use case. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 343–361.

Memon, M. A., Salleh, R., & Baharom, M. N. R. (2016). The link between training satisfaction, work engagement and turnover intention. *European Journal of Training and Development*, *40*(6), 407–429.

Mitchell, T. R., Holtom, B. C., & Lee, T. W. (2001). How to keep your best employees: Developing an effective retention policy. *Academy of Management Perspectives*, *15*(4), 96–108.

Mittal, S., et al. (2023). Employee attrition prediction using machine learning algorithms. *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 1–6.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological bulletin*, *86*(3), 493.

Mohammed, A. Q. (2019). Hr analytics: A modern tool in hr for predictive decision making. *Journal of Management*, *6*(3).

Morrow, P. C., McElroy, J. C., Laczniak, K. S., & Fenton, J. B. (1999). Using absenteeism and performance to predict employee turnover: Early detection through company records. *Journal of vocational behavior*, *55*(3), 358–374.

Mozaffari, F., Rahimi, M., Yazdani, H., & Sohrabi, B. (2023). Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. *Benchmarking: An International Journal*, *30*(10), 4140–4173.

Nahar, L., Tasnim, F., Sultana, Z., & Tuli, F. A. (2022). Employee turnover prediction model for garments organizations of bangladesh using machine learning technique. *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 1–5.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, *16*(1).

O'Connell, M., & Kung, M.-C. (2007). The cost of employee turnover. *Industrial Management*, *49*(1).

O'Neil, C. (2016, September 21). *Rogue algorithms and the dark side of big data* [Accessed: 2024-09-17]. https://knowledge.wharton.upenn.edu/article/rogue-algorithms-darkside-big-data/

Ongori, H. (2007). A review of the literature on employee turnover. *African Journal of Business Management*, 049–054.

Ore, O., & Sposato, M. (2022). Opportunities and risks of artificial intelligence in recruitment and selection. *International Journal of Organizational Analysis*, *30*(6), 1771–1782.

Pan, Y., & Froese, F. J. (2023). An interdisciplinary review of ai and hrm: Challenges and future directions. *Human Resource Management Review*, *33*(1), 100924.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.

Porter, L. W., & Steers, R. M. (1973). Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological bulletin*, *80*(2), 151.

Pratibha, G., & Hegde, N. P. (2022). Hr analytics: Early prediction of employee attrition using kpca and adaptive k-means based logistic regression. *2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS)*, 11–16.

PriceWaterhouseCoopers. (2008, September). *The 11th annual global ceo survey* (tech. rep.) (PriceWaterhouseCoopers, 2008a). PwC. New York.

Punnoose, R., & Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence*, *5*(9), 22–26.

Qualtrics. (2022, May). *Employee lifecycle – definition & 7-phasen-modell — qualtrics*. Retrieved June 28, 2024, from https://www.qualtrics.com/de/erlebnismanagement/mitarbeiter/employee-lifecycle/

Ramalho Luz, C. M. D., Luiz de Paula, S., & de Oliveira, L. M. B. (2018). Organizational commitment, job satisfaction and their possible influences on intent to turnover. *Revista de Gestão*, *25*(1), 84–101.

Ranjan, S., & Yadav, R. S. (2018). Uncovering the role of internal csr on organizational attractiveness and turnover intention: The effect of procedural justice and extraversion. *Asian Social Science*, *14*(12), 76–85.

Romer, P. M. (1992). Two strategies for economic development: Using ideas and producing ideas. *The World Bank Economic Review*, *6*(suppl_1), 63–91.

Samad, S. (2006). The contribution of demographic variables: Job characteristics and job satisfaction on turnover intentions. *Journal of International Management Studies*, *1*(1).

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006.

Sato, K., Oka, M., & Kato, K. (2019). Early turnover prediction of new restaurant employees from their attendance records and attributes. *Database and Expert Systems*

Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30, 277–286.

Saxena, M., Bagga, T., Gupta, S., & Mittal, A. (2022). Employees' experiences of accepting and adopting hr analytics: A phenomenology study. *The Open Psychology Journal*, *15*(1).

Schlechter, A. F., Syce, C., & Bussin, M. (2016). Predicting voluntary turnover in employees using demographic characteristics: A south african case study. *Acta Commercii*, *16*(1), 1–10.

Schweyer, A. (2004). *Talent management systems: Best practices in technology solutions for recruitment, retention and workforce planning*. John Wiley & Sons.

Sharma, M., Singh, D., Tyagi, M., Saini, A., Dhiman, N., & Garg, R. (2022). Employee retention and attrition analysis: A novel approach on attrition prediction using fuzzy inference and ensemble machine learning. *Webology*, *19*(2).

Shaw, J. D., Delery, J. E., Jenkins Jr, G. D., & Gupta, N. (1998). An organization-level analysis of voluntary and involuntary turnover. *Academy of management journal*, *41*(5), 511–525.

Sibiya, M., Buitendach, J. H., Kanengoni, H., & Bobat, S. (2014). The prediction of turnover intention by means of employee engagement and demographic variables in a telecommunications organisation. *Journal of Psychology in Africa*, *24*(2), 131–143.

Singh, A. K., & Thakral, P. (2023). Analysis of employee attrition using statistical and machine learning approaches. *International Conference on Artificial Intelligence of Things*, 269–279.

Skelton, A. R., Nattress, D., & Dwyer, R. J. (2020). Predicting manufacturing employee turnover intentions. *Journal of Economics, Finance and Administrative Science*, *25*(49), 101–117.

Srivastava, P. R., & Eachempati, P. (2021). Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach. *Journal of Global Information Management (JGIM)*, *29*(6), 1–29.

Stone, D. L., Deadrick, D. L., Lukaszewski, K. M., & Johnson, R. (2015). The influence of technology on the future of human resource management. *Human resource management review*, *25*(2), 216–231.

Storey, J., Ulrich, D., & Wright, P. M. (2019). *Strategic human resource management: A research overview*. Routledge.

Stovel, M., & Bontis, N. (2002). Voluntary turnover: Knowledge management–friend or foe? *Journal of intellectual Capital*, *3*(3), 303–322.

Sutherland, J. (2002). Job-to-job turnover and job-to-non-employment movement: A case study investigation. *Personnel Review*, *31*(6), 710–721.

Swailes, S. (2013). The ethics of talent management. *Business Ethics: A European Review*, *22*(1), 32–46.

Taeger, J., & Gabel, D. (2022). *Dsgvo-bdsg-ttdsg*. Fachmedien Recht und Wirtschaft.

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15–42.

Thalmann, S., & Ilvonen, I. (2020). Why should we investigate knowledge risks incidents?: Lessons from four cases. *Hawaii International Conference on System Sciences*.

Tursunbayeva, A., Pagliari, C., Di Lauro, S., & Antonelli, G. (2022). The ethics of people analytics: Risks, opportunities and recommendations. *Personnel Review*, *51*(3), 900–921.

Tziner, A., & Birati, A. (1996). Assessing employee turnover costs: A revised approach. *Human Resource Management Review*, *6*(2), 113–122.

Van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of hr analytics: A study into the future application, value, structure, and system support. *Journal of Organizational Effectiveness: People and Performance*, *4*(2), 157–178.

Vecchio, R. P., & Norris, W. R. (1996). Predicting employee turnover from performance, satisfaction, and leader-member exchange. *Journal of Business and Psychology*, *11*, 113–125.

Venkat, M. V. V., Khan, S. R. K., Gorkhe, M. D., Reddy, M. K. S., & Rao, S. P. (2023). Fostering talent stability: A study on evaluating the influence of competency management on employee retention in the automotive industry. *Remittances Review*, *8*(4).

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC medical research methodology*, *19*, 1–9.

Wang, X., & Zhi, J. (2021). A machine learning-based analytical framework for employee turnover prediction. *Journal of Management Analytics*, *8*(3), 351–370.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii. Retrieved March 28, 2024, from http://www.jstor.org/stable/4132319

Wei, Y.-C. (2015). Do employees high in general human capital tend to have higher turnover intention? the moderating role of high-performance hr practices and po fit. *Personnel Review*, *44*(5), 739–756.

Wójcik, P. (2017). Shortage of talents–a challenge for modern organizations. *International Journal of Synergy and Research*, *6*.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381.

Yadav, S., Jain, A., & Singh, D. (2018). Early prediction of employee attrition using data mining techniques. *2018 IEEE 8th international advance computing conference (IACC)*, 349–354.

Yahia, N. B., Hlel, J., & Colomo-Palacios, R. (2021). From big data to deep data to support people analytics for employee attrition prediction. *Ieee Access*, *9*, 60447–60458.

Yamin, M. (2020). Examining the role of transformational leadership and entrepreneurial orientation on employee retention with moderating role of competitive advantage. *Management Science Letters*, *10*(2), 313–326.

Yuan, J. (2021). Research on employee turnover prediction based on machine learning algorithms. *2021 4th international conference on artificial intelligence and big data (icaibd)*, 114–120.

Zhang, H., Xu, L., Cheng, X., Chao, K., & Zhao, X. (2018). Analysis and prediction of employee turnover characteristics based on machine learning. *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, 371–376.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2019). Employee turnover prediction with machine learning: A reliable approach. *Intelligent Systems and Ap-*

plications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2, 737–758.

Zhu, Q., Shang, J., Cai, X., Jiang, L., Liu, F., & Qiang, B. (2019). Coxrf: Employee turnover prediction based on survival analysis. *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1123–1130.

Zieglmeier, V., Gierlich-Joas, M., & Pretschner, A. (2022). Increasing employees' willingness to share: Introducing appeal strategies for people analytics. *International Conference on Software Business*, 213–226.

Appendix

Summary of IBM HR Analytics Dataset

Variable	Type	Description
Age	Numeric	Age of the employee
Attrition	Categorical	Whether the employee has left the company
BusinessTravel	Categorical	Frequency of business travel
DailyRate	Numeric	Daily rate of pay
Department	Categorical	Department in the company
DistanceFromHome	Numeric	Distance from home to workplace
Education	Categorical	Level of education
EducationField	Categorical	Field of education
EmployeeCount	Numeric	Count of employees
EmployeeNumber	Numeric	Unique identifier for employees
EnvironmentSatisfaction	Numeric	Satisfaction with work environment
Gender	Categorical	Gender of the employee
HourlyRate	Numeric	Hourly rate of pay
JobInvolvement	Numeric	Level of job involvement
JobLevel	Numeric	Level of job position
JobRole	Categorical	Role in the company
JobSatisfaction	Numeric	Satisfaction with job
MaritalStatus	Categorical	Marital status of the employee
MonthlyIncome	Numeric	Monthly income
MonthlyRate	Numeric	Monthly rate of pay
NumCompaniesWorked	Numeric	Number of companies worked for
Over18	Categorical	Whether the employee is over 18
OverTime	Categorical	Whether the employee works overtime
PercentSalaryHike	Numeric	Percentage salary hike
PerformanceRating	Categorical	Performance rating
RelationshipSatisfaction	Categorical	Satisfaction with work relationships
StandardHours	Numeric	Standard working hours
StockOptionLevel	Categorical	Level of stock options
TotalWorkingYears	Numeric	Total years of working
TrainingTimesLastYear	Numeric	Number of training sessions last year
WorkLifeBalance	Categorical	Balance between work and personal life
YearsAtCompany	Numeric	Years spent at the company
YearsInCurrentRole	Numeric	Years in the current role
YearsSinceLastPromotion	Numeric	Years since last promotion
YearsWithCurrManager	Numeric	Years with current manager

Most Important Features According to Feature Importance Rankings

Features	
Age	Average Working Hours
Business Travel	Children Cost
City of Residence	Competence Center
Department	Department Type
Direct Manager	Distance from Home
Education	Education Status
Environment Satisfaction	First-Grade Experience
Gender	Grade
Graduated Major	Income Base
Job Involvement	Job Level
Job Performance	Job Satisfaction
Last Contract Duration	Location
Manager	Title
Marital Status	Mission Cost
Monthly Income	Office Base
Overtime	Percent Salary Increase
Position Category	Previous Experience
Record by Day	Record by Month
Rewards	Tenure
Total Experience	Total Working Years
Training	Training Times Last Year
Units	University Graduated
Years at Company	Years in Current Role
Years since last Promotion	Years with Current Manager

Note: Based on Feature Importance Rankings of (Gao et al., 2019; Gurler et al., 2023; Mozaffari et al., 2023; Yahia et al., 2021)

Analysed Studies of Literature Review

Paper	Authors	Features
Early Turnover Prediction of New Restaurant Employees from Their Attendance Records and Attributes	Koya Sato, Mizuki Oka, Kazuhiko Kato	Age, DayoftheWeek, EmploymentMonth, FinishingTime, Gender, HiringChannel, PreviousWorkingPeriods, RoleinRestaurant, StartingTime, WorkingHours, WorkingType
Explaining and predicting employees' attrition: a machine learning approach	Praphula Kumar Jain, Madhur Jain, Rajendra Pamula	AverageMonthlyHour, Department, LastEvaluation, NumberofProjects, PromotionLast5Years, Salary, SatisfactionLevel, TimeSpentCompany, WorkAccident
A machine learning-based analytical framework for employee turnover prediction	Xinlei Wang, Jianing Zhi	Age, AverageMonthlyHours, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, LastEvaluation, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, NumberProject, Over18, OverTime, PercentSalaryHike, PerformanceRating, PromotionLast5Years, RelationshipSatisfaction, Salary, Sales, SatisfactionLevel, StandardHours, StockOptionLevel, TimeSpendCompany, TotalWorkingYears, TrainingTimesLastYear, WorkAccident, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
Employee Turnover Prediction with Machine Learning: A Reliable Approach	Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, Xiaoyu Zhu	Age, Compensation, Department, EducationBackground, Ethnicity, Gender, HighestEducation, IsClientFacingRole, JobTenure, LastPayRaise, ManagementLevel, ServiceLine, SpecializedArea, Team, Title, 2015Performance, 2016Performance
CoxRF: Employee Turnover Prediction based on Survival Analysis	Qianwen Zhu, Jiaxing Shang, Xinjun Cai, Linli Jiang, Feiyi Liu, Baohua Qiang	ColumnNumber, CommentNumber, EndYear, GDP, Gender, ImpressionTagNumber, IndustryType, InfluenceNumber, LengthWorkingTime, Likes, MaxDegree, MaxSchoolType, NumberInteractions, NumberPosts, NumberTurnovers, PercentageInfluenceonOthers, PercentageInformationPerfection, PositionLevel, ProfessionTagNumber, RecentFeeds, StandpointNumbers, StartYear, TimeWorked, Views
Employee Turnover Prediction Model for Garments Organizations of Bangladesh Using Machine Learning Technique	Lutfun Nahar, Zinnia Sultana, Farzana Tasnim, Farjana Akter Tuli	Age, AnnualRefreshmentFacility, Departments, DistanceFromHome(km), EducationField, EducationQualification, EmployeePromotionStatus, Gender, HumanResourcePolicy, JobEnvironmentSatisfaction, JobSatisfaction, MaritalStatus, On-TimeSalary, OvertimeBillAllowance, SafetySatisfaction, SalarySatisfaction, StressLevel, TransportFacility, YearlySalaryIncrement
EmpTurnoverML: An Efficient Model for Employee Turnover and Customer Churn Prediction Using Machine Learning Algorithms	Diaa Salama AbdElminaam, Mariam Maged, Mariam Khaled Mousa, AbdurRahman Ossama Younis, Mostafa Saleh Abdelsalam, Yara Hisham, Tarek Talaat	AverageMonthlyHours, Department, LastEvaluation, NumberProjects, Promotionlast5years, Salary, SatisfactionLevel, TimeSpendCompany, WorkAccident
An Improved Random Forest Algorithm for Predicting Employee Turnover	Xiang Gao, Junhao Wen, Cheng Zhang	Age, AvgWorkHours, DepartmentType, DistanceFromHome, Education, EducationField, EmployeeNumber, EmploymentNature, EnvironmentSatisfaction, Gender, HaveChildren, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, NativePlace, NumberCompaniesWorked, OverTime, PercentSalaryIncrease, PerformanceRatingLastYear, PhysicalCondition, RelationshipSatisfaction, TotalWorkingYears, TrainingTimesLastYear, WinningCount, WorkLifeBalance, YearsatCompany, YearsinCurrentRole, YearsSinceLastPromotion, YearswithCurrentManager
Research on Employee Turnover Prediction Based on Machine Learning Algorithms	Jia Yuan	AffectiveCommitment, Age, BusinessTravel, CareerGrowthOpportunities, ContinuousCommitment, DistanceFromHome, DistributiveJustice, Education, EnvironmentSatisfaction, Gender, GeneralTraining, HaveChildren, JobAutonomy, JobInvolvement, JobLevel, JobRole, JobSatisfaction, JobSearchBehavior, MainFinancialSupport, MaritalStatus, MonthlyIncome, NegativeAffectivity, NormativeCommitment, NumberofCompaniesWorked, Opportunity, OrganizationalCommitment, OverTime, Payment, PercentSalaryHike, PerformanceRating, PerformanceScore, PositiveAffectivity, PromotionalChance, RelationshipSatisfaction, TotalWorkingYears, TransferCost, TurnoverIntention, WorkLifeBalance, Workload, YearsSinceLastPromotion, YearsatCompany
Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data	Fatemeh Mozaffari, Marzieh Rahimi, Hamidreza Yazdani, Babak Sohrabi	Age, ChildrenCost, Department, DirectManager, Education, FirstGradeExperience, IncomeBase, LastContractDuration, Manager, MissionCost, OfficeBase, PositionCategory, RecordbyDay, RecordbyMonth, Title, Units
Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction	Praveen Ranjan Srivastava, Prajwal Eachempati	AppraisalRating, CTCLevel, NumberofProjects, Promotion, SafetyMeasure, Satisfaction, TimeSpentonProjects
Predicting employee attrition with a more effective use of historical events	Abdel-Rahmen Korichi, Hamamache Kheddouci, Daniel J West	Age, EffectiveDate, Gender, Grade, MedianSalary, MonthofObservation, MonthofStartingDate, MonthsBeforeTerminationDate, MonthsSinceNoGradeIncrease, MonthsSinceNoSalaryIncrease, Salary, StartDate, Tenure, TerminationDate, TimeSinceLastGradeIncrease, TimeSinceLastSalaryIncrease
From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction	Nesrine Ben Yahia, Jihen Hlel, Ricardo Colomo-Palacios	Age, BusinessTravel, EnvironmentSatisfaction, Grade, JobInvolvement, JobPerformance, JobSatisfaction, MaritalStatus, Rewards, Tenure, Training
Predicting employee attrition using tree-based models	Nesreen El-Rayes, Ming Fang, Michael Smith, Stephen M. Taylor	AverageSalary(NewJob), AverageSalary(OriginalJob), EndDate, Industry, JobTitle, NumberEmployees(NewJob), NumberEmployees(OriginalJob), OverallEmployeeRating(NewJob), OverallEmployeeRating(OriginalJob), StartDate, StartingYear(NewJob), StartingYear(OriginalJob), Tenure, YearCompanyFounded(NewJob), YearCompanyFounded(OriginalJob)
Employee Attrition Prediction using Artificial Neural Networks	Akansha Chaurasia, Shreyas Kadam, Kalyani Bhagat, Shreenath Gauda, Priyanka Shingane	Age, EnvironmentSatisfaction, Gender, JobInvolvement, JobSatisfaction, NumberCompaniesWorked, OverTime, Salary, TotalWorkingYears, WorkAccident, WorkLifeBalance, YearsSinceLastPromotion, YearsinCurrentRole
Churn Prediction of Employees Using Machine Learning Techniques	Nilasha Bandyopadhyay, Anil Jadhav	Age, ChallengingWork, Discrimination, Gender, PeersLeaving, PromotionLastYear, SalaryLevel, Satisfaction, WorkRecognition, YearsofExperience
Designing of Customer and Employee Churn Prediction Model Based on Data Mining Method and Neural Predictor	Sepideh Hassankhani Dolatabadi, Farshid Keynia	Age, AreaofExpertise, AverageNumberofServicesGivenperYear, AverageTimeAssignedtoEachService, DuplicateServiceorNot, Gender, LevelofEducation, MaritalStatus, MissCallServiceorNot, NumberofDuplicateService, ReferringtotheSoftwareDevelopmentTeamorNot, RelatingCertificatetoEmployeeService, RequestChannel, ResolvedServiceontheFirstCallingorNot, ResolvedServiceontheFirstDayorNot, ScopeofService, ServiceClass, ServiceTime, ServiceType, WorkExperience, Year
Employee churn prediction	V. Vijaya Saradhi, Girish Keshav Palshikar	Age, Billed, Department, Designation, DesignationClientOrganisation, EmployeeLocation, ExperienceClientOrganisation, ExperienceParentOrganisation, Gender, On-site, PastExperienceYears, Qualification
A Comparative Study of Employee Churn Prediction Model	Andry Alamsyah, Nisrina Salma	Age, BaseSalary, Division, FinalSalary, Gender, ID, JobVacancy, LevelofPosition, Location, NumberYearsCompany, Position
Employee Attrition Prediction Using Machine Learning Comparative Study	Shobhit Aggarwal, Manik Singh, Shivam Chauhan, Mugdha Sharma, Deepti Jain	Age, Attrition, Behaviour, BusinessTravel, CommunicationSkills, Department, DistanceHome, Education, EducationField, EmployeeNumber, EnvironmentSatisfaction, Gender, ID, JobInvolvement, JobSatisfaction, MaritalStatus, MonthlyIncome, NumberCompaniesWorkedat, Overtime, PercentageSalaryHike, PerformanceRating, StockOptionsLevel, TotalWorkingYears, TrainingsLastYear, YearsCurrentManager, YearsSinceLastPromotion, YearsatCompany, YearsinCurrentRole
Evaluating employee attrition and its factors using machine learning approaches	A. Sheik Abdullah, P.Je Sai Kailash, Deepthi Ramesh, Prithviraj Guntha	Age, Branch, Department, Gender, Grade, GrossSalary, Tenure
A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)	Amir Mohammad Esmaieeli Sikaroudi, Rouzbeh Ghousi, Ali Esmaieeli Sikaroudi	Age, AverageWorkingExperienceOtherCompanies, CompatibilityExperiencewithJob, CompatibilityofBodywithJob, DurationRelevantWorkExperience, Education, KnowledgeAboutWorkingConditionsandLaws, MaritalStatus, NumberJobChanging, PerseveranceandInteresttoWork, SocialInteractionSkills, TechnicalSkills, VeteranStatus, WorkingExperienceCurrentCompany
Exploratory Data Analysis and Classification of Employee Retention based on Logistic Regression Model	Anupkumar M Bongale, Deepak Dharrao, Siddhaling Urolagin	Age, CityOfficeWherePosted, EducationLevel, EverBenchedorNot, ExperienceintheCurrentDomain, Gender, PaymentTier, YearofJoiningCompany
Employee retention prediction in corporate organizations using machine learning methods	Abdullah A. Abonamah Ab, Khaled Alshehhi, Safeya Bin Zawbaa, Muhammad Usman Tariq	ActionDate (attrition), ContractType, DateofBirth, DegreeLevel, DegreeMajor, Employee'sDepartment, EmployeeID, EmployeeName, EmploymentStatus, EmploymentTitle, EngagementDate, Gender, Grade, LastActionTypeTakenAgainstanEmployeebytheEmployerwithRespecttoTheirEmployment, LeaveEntitlement, MaritalStatus, TrainingCount
A Machine Learning Approach for Employee Retention Prediction	Ggaliwango Marvin, Majwega Jackson, Md. Golam Rabiul Alam	City, CityDevelopmentIndex, CompanySize, CompanyType, EducationLevel, EnrolledUniversity, Experience, Gender, ID, LastNewJob, MajorDiscipline, RelevantExperience, TrainingHours
Analysis of Employee Attrition using Statistical and Machine Learning Approaches	Akash Kumar Singh, Prateek Thakra	Age, BusinessTravel, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeID, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, NumCompaniesWorked, Over18, PercentSalaryHike, PerformanceRating, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager
An Efficient Employee Retention Prediction Model for Manufacturing Industries Using Machine Learning Approach	S. Radhika, S. Umamaheswari, R. Ranjith, A. Chandrasekar	Age, DepartmentofWorking, Designation, EducationQualification, Encouragement, Experience, FlexibleWorkingHours, Gender, Goodwill, Income, ManagerSupport, MaritalStatus, Promotion, Satisfaction, Training, WorkEnvironment
Deep Learning Based Employee Attrition Prediction	Kerem Gurler, Burcu Kuleli, Vehbi Cagri Gungor	Age, City of residence, Competence center, Department, Education status, Experience current company, Gender, Graduated major, Location, Marital status, Previous experience, Total experience, University graduated
Predicting Bus Operator Retention Based on Employee Characteristics and Work History	Eric Lind and Joel Huting	Age, Assigned Garage, CDL, Cumulative absences, Cumulative incidents, Cumulative responsible accidents, Gender, Highest education, Probation, Separation type, Training, Yearhired
Employee Retention And Attrition Analysis: A Novel Approach On Attrition Prediction Using Fuzzy Inference And Ensemble Machine Learning	M. K. Sharma, Dhanpal Singh, Manaswita Tyagi, Ayushi Saini, Nitesh Dhiman, Riddhi Garg	CompanySize, CompanyType, Education Level, Gender, LastNewJob, MajorDiscipline, Relevant Experience, University enrolled in

Overview of all Distinct Variables

Features n = 286			
actiondate	employmentnature	missioncost	salary
age	employmenttitle	monthlyincome	salarylevel
affectivecommitment	encouragement	monthlyrate	salarysatisfaction
annualrefreshmentfacility	enddate	monthofobservation	sales
appraisalrating	endyear	monthofstartingdate	satisfaction
areaofexpertise	engagementdate	monthsbeforeterminationdate	satisfactionlevel
assigned garage	enrolleduniversity	monthssincenogradeincrease	scopeofservice
averagemonthlyhours	environmentsatisfaction	monthssincenosalaryincrease	separation type
averagenumberofservicesgivenperyear	ethnicity	nativeplace	serviceclass
averagesalary(originaljob/newjob)	everbenchedornot	negativeaffectivity	serviceline
averagetimeassignedtoeachservice	experience	normativecommitment	servicetime
averageworkingexperienceothercompanies	experienceclientorgansisation	numbercompaniesworked	servicetype
avgworkhours	experienceparentorganisation	numberemployees(newjob)	socialinteractionskills
basesalary	experience current company	numberemployees(originaljob)	specializedarea
behaviour	experienceinthecurrentdomain	numberinteractions	standardhours
billded/notbilled	finalsalary	numberjobchanging	standpointnumbers
branch	finishingtime	numberofduplicateservice	startdate
businesstravel	firstgradeexperience	numberofprojects	startingtime
careergrowthopportunities	flexibleworkinghours	numberposts	startingyear(newjob)
cdl	gdp	numberturnovers	startingyear(originaljob)
challengingwork	gender	numberyearscompany	startyear
childrencost	generaltraining	officebase	stockoptionlevel
city	goodwill	on site/off site	stresslevel
city of residence	grade	on-timesalary	team
citydevelopmentindex	graduated major	opportunity	technicalskills
cityofficewhereposted	grosssalary	organizationalcommitment	tenure
columnnumber	havechildren	over18	terminationdate
commentnumber	highesteducation	overallemployeerating(newjob)	timesincelastgradeincrease
communicationskills	hiringchannel	overallemployeerating(originaljob)	timesincelastsalaryincrease
companysize	hourlyrate	overtime	timespentcompany
companytype	humanresourcepolicy	overtimebillallowance	timespentonprojects
compatibilityexperiencewithjob	id	payment	timeworked
compatibilityofbodywithjob	impressiontagnumber	paymenttier	title
compensation	income	peersleaving	totalexperience
competence center	incomebase	percentageinfluenceonothers	totalworkingyears
continuouscommitment	industry	percentageinformationperfection	training
contracttype	industrytype	percentagesalaryhike	trainingcount
ctclevel	influencenumber	percentsalaryhike	traininghours
cumulative absences	isclientfacingrole	percentsalaryincrease	trainingslastyear
cumulative incidents	jobautonomy	performancerating	trainingtimeslastyear
cumulative responsible accidents	jobenvironmentsatisfaction	performanceratinglastyear	transfercost
dailyrate	jobinvolvement	performancescore	transportfacility
dateofbirth	joblevel	perseveranceandinteresttowork	turnover intenion
dayoftheweek	jobperformance	physicalcondition	units
degreelevel	jobrole	position	universityenrolledin
degreemajor	jobsatisfaction	positioncategory	university graduated
department	jobsearchbehavior	positionlevel	veteranstatus
departmentofworking	jobtenure	positiveaffectivity	views
departmenttype	jobtitle	previous experience	winningcount
designation	jobvacancy	previousworkingperiods	workaccident
designationclientorganisation	knowledgeaboutworking conditionsandlaws	probation	workenvironment
directmanager	lastactiontypetakenagainst anemployeebytheemployerwith respecttotheiremployment	professiontagnumber	workexperience
discrimination	lastcontractduration	promotion	workingexperience currentcompany
distancefromhome	lastevaluation	promotionalchance	workinghours
distributivejustice	lastnewjob	promotionlast5years	workingtype
division	lastpayraise	promotionlastyear	worklifebalance
duplicateserviceornot	leaveentitlement	qualification	workload
durationrelevantworkexperience	lengthworkingtime	recentfeeds	workrecognition
education	levelofeducation	recordbyday	year
education status	levelofposition	recordbymonth	yearcompanyfounded(newjob)
educationbackground	likes	referringtothesoftware developmentteamornot	yearcompanyfounded(originaljob)
educationfield	location	relatingcertificatetoemployeeservice	yearhired
educationlevel	mainfinancialsupport	relationshipsatisfaction	yearlysalaryincrement
educationqualification	majordiscipline	relevantexperience	yearofjoiningcompany
effectivedate	managementlevel	requestchannel	yearsatcompany
employeecount	manager	resolvedserviceonthefirstcallingornot	yearsincurrentrole
employeeid	managersupport	resolvedserviceonthefirstdayornot	yearsofexperience
employeename	maritalstatus	rewards	yearssincelastpromotion
employeenumber	maxdegree	roleinrestaurant	yearswithcurrentmanager
employeepromotionstatus	maxschooltype	safetymeasure	performance(2015,2026)
employee'sdepartment	mediansalary	safetysatisfaction	employee number
employmentmonth	misscallserviceornot		