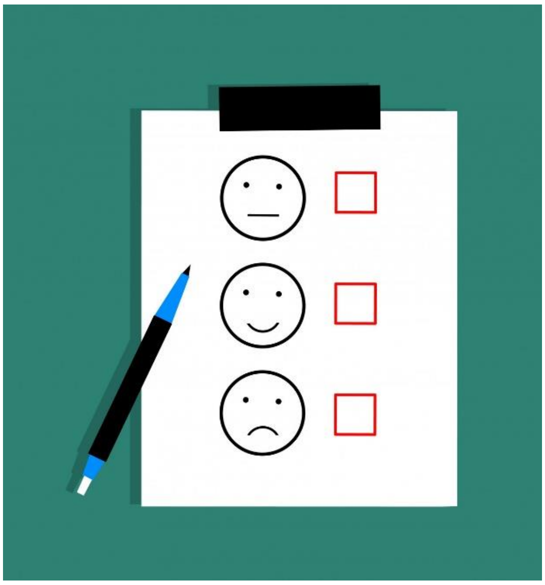
University of Sussex



Master Dissertation

“ How to predict employee turnover in technology companies using Machine Learning ”

Paulina Traub
MSc Human and Social Data Science 2023
Supervisor: Julie Weeds
University of Sussex

Contents

Abstract

In the present day, businesses have come to recognize the necessity of adjusting their approaches and enhancing the efficiency of each operation through digital means. This adjustment is crucial to maintain a leading position in the market and avoid lagging behind competitors. The quantity of technology firms is rapidly expanding, leading to an increasing demand for technology-oriented roles. This demand is exacerbated by the ongoing requirement for digital transformation, which commenced years ago.

In light of intense rivalry and the recent industrial revolution, employee turnover presents a significant issue and obstacle for both Human Resources and the company as a whole. This is why this study aims to examine the influence of employee turnover on productivity and the expenses associated with locating an optimal replacement. The research will also delve into potential methods of predicting turnover within the technology workforce and explore the advantages of proactive employment of artificial intelligence measures, particularly from a human resources standpoint.

1-. Introduction

Employee turnover within the Information Technology sector is influenced by various factors, with workforce competition being the foremost. UK IT companies are vying for top talent, leading to salary hikes and enhanced benefits for IT professionals (Sallaba, s.f.). To tackle talent retention challenges, UK companies are employing strategies such as professional development initiatives, continuous learning opportunities, fostering inclusive and flexible work environments, and offering competitive compensation packages (Frick et al., 2021).

Predicting employee turnover is intricate, encompassing individual motivations, job satisfaction, career prospects, and external market dynamics (Nwokocha & Iheriohanma, 2012). The authors said that technological employees possess unique skill sets and high market demand, rendering their turnover patterns intricate to accurately forecast. Furthermore, turnover prediction and selecting appropriate machine learning algorithms and feature engineering techniques. Balancing model accuracy and interpretability adds complexity to this research field (Rencheng, et al., 2022).

The onset of the COVID-19 pandemic in December 2019 has spurred significant changes across various domains, including the ICT industry (Yang, 2022). Government responses and increased adoption of non-face-to-face ICT technologies have reshaped the industry, fostering extensive online measures and video conferencing (Yang, 2022). These shifts have also impacted global ICT goods business and raised concerns about a country's ICT industry structure, its role in global value chains, and the transformation it is undergoing (Yang, 2022). According to the OECD, digital roles are growing even more critical, continuing to reshape daily life and work regardless of the crisis's evolution (OECD, 2020).

The development of 5G and the Internet of Things (IoT) is anticipated to amplify data generation, intensifying debates on data management, privacy, and security. As firms contemplate automation for bolstering resilience, data exchange could become paramount (OECD, 2020). This scenario appears to contribute to the increasing turnover rate in digital jobs.

To delve into the realm of human resources, it's imperative to define key concepts, including "turnover." As Atef, Elzanfaly, and Ouf (2022) affirm, employee turnover signifies the rate at which employees depart and new ones are hired. High turnover rates entail costs for organizations, encompassing separation, vacancy, recruitment, training, and replacement expenses (Atef et al., 2022). Furthermore, high turnover can lead to understaffing and reduced productivity, impacting organizational growth. Hence, predicting turnover is vital for sustainable organizations to avert such circumstances.

In the context of human resource management, artificial intelligence (AI) is integral. AI's speed and accuracy make it increasingly attractive in HR management (Atef et al., 2022). Machine learning, a subset of AI, enables machines to learn from data and experiences without explicit programming. AI's integration into HR, particularly in predicting turnover, can automate routine tasks, optimizing HR efficiency (Atef et al., 2022).

Specifically in The UK, The British workforce is currently undergoing a noticeable transformation. Despite the UK's historical struggles with productivity, the aftermath of the pandemic has introduced new terms like 'burnout,' 'quiet quitting,' and a 'Great Fatigue' (Kipping & Da Costa, 2022). This situation raises a pertinent question: Are we encountering a global crisis in labor motivation?

Within the United Kingdom, the Information Technology (IT) sector faces challenges in retaining talent. Despite the sector's vibrancy and prosperity, there's an unceasing demand for highly skilled IT professionals, creating intense competition for such talent (Blake, 2021). Talent retention has emerged as a significant concern in this field. Recent data from Dealroom for the Digital Economy Council reveals that the UK's tech sector is poised to finish as Europe's leading ecosystem (Scully & Department for Digital, Culture, Media & Sport, 2022). This assertion by Scully and the Department emphasizes the UK's competitiveness against global economic challenges and its positioning vis-à-vis the US and China.

Indeed, 2022 has seen the rapid expansion of tech companies in the UK, with record-breaking fundraising reaching £24 billion, surpassing the combined funds raised by France and Germany.

Over the past five years, UK tech companies have secured nearly £100 billion in funding (Scully & Department for Digital, Culture, Media & Sport, 2022).

2-. Literature Review

According to Park and Shaw (2013), the field of organizational-level turnover is less developed than its individual-level counterpart, despite increasing interest in the former. This points to a pressing need for more in-depth research and understanding in organizational turnover studies.

The same authors have examined the relationship between turnover rates and organizational performance through three unique lenses. First, some scholars posit that turnover negatively impacts performance across the board. Second, another school of thought suggests that turnover is particularly detrimental at low to moderate levels, but these adverse effects diminish as turnover escalates. Third, there exists an argument that low to moderate levels of turnover can enhance organizational performance, only becoming problematic when the rates are excessively high.

While numerous studies have indicated that high turnover rates are negatively correlated with various organizational outcomes such as profits, customer service, and return on assets, Park and Shaw (2013) note that the findings are not universally consistent. Some research has failed to establish any negative correlation, and there are instances where positive relationships have been reported.

To explore into the domain of human resources, our initial step will encompass elucidating a range of concepts, among them the notion of "turnover." According to the findings of Atef, Elzanfaly, and Ouf (2022), **employee turnover** pertains to the frequency with which employees depart an organization and are subsequently replaced by new hires. The consequence of elevated turnover rates can lead to substantial financial burdens for organizations, encompassing costs tied to separation, vacant positions, recruitment, training, and replacement (Atef et al., 2022). The authors further contend that heightened turnover can culminate in understaffing and decreased productivity, hampering the overall growth of the organization. Therefore, it becomes indispensable for sustainable organizations to anticipate employee turnover and implement preventive strategies against such eventualities.

While turnover generally refers to the rate of employee departures, it's crucial to distinguish between two categories of turnover as proposed by Mamun and Hasan (2017): **voluntary and involuntary**. **Voluntary turnover** unfolds when an employee voluntarily opts to leave the organization, often motivated by factors like job dissatisfaction, stress, or the allure of better opportunities elsewhere. On the other hand, **involuntary turnover** transpires when the employer initiates the termination of an employee's tenure, whether due to retirement, death, dismissal, or other extraneous circumstances beyond the employee's control, such as caring for an ailing family member or accompanying a spouse to a remote location. These authors accentuate the significance of comprehending the triggers behind voluntary turnover and devising effective management strategies. As such, this study will concentrate solely on voluntary turnover, although it's worth noting that involuntary turnover can also be managed to mitigate detrimental repercussions on both the organization and its employees.

Beyond these two turnover categories, Mamun and Hasan (2017) introduce the distinction between **Avoidable and Unavoidable turnover. Avoidable turnover** pertains to instances where organizations can avert employee departures by enhancing their recruitment practices, assessment methodologies, and strategies for motivating employees. In this context, it's crucial for businesses to effectively address voluntary turnover and implement measures to bolster employee retention. Conversely, **unavoidable turnover** emerges from circumstances beyond the employer's direct control, such as an employee's personal decision to relocate or a spouse's job-related transfer. However, nearly 80% of turnover can be attributed to avoidable recruitment missteps (Mamun & Hasan, 2017). Thus, organizations should adopt measures to meticulously select and assess prospective employees.

Regarding turnover variables, Belete (2018) asserts that multiple factors can influence turnover, and these variables may vary across different organizations. Belete contends that attributing turnover intentions to a single factor is insufficient, advocating for a comprehensive approach to understand the myriad factors shaping employees' decisions to leave an organization. Our analysis will incorporate factors deemed most pertinent based on the research of Mamun and Hasan, in addition to insights from Belete. Furthermore, **we will avoid taking demographic factors into account, such as gender, age, and marital status within the company.**

Demographics entail statistical information that delineates the attributes of populations and, analysis involves examining a population's composition through variables like age, ethnicity, and gender (Hayes, 2022). Hayes stated that demographic data pertains to socio-economic details conveyed in numerical form, encompassing aspects such as employment, education, earnings, rates of marriage, birth and mortality, and other relevant factors.

Governments, businesses, and non-governmental entities utilize demographic insights to gain deeper insights into a population's distinctive features, serving various objectives including policy formulation and economic market investigation (Hayes, 2022). To illustrate, a company specializing in luxury RVs might aim to target individuals approaching retirement age and assess the proportion of those with the financial means to afford their products (Hayes, 2022).

Hayes (2022) said that demographics encompass statistical information that delineates the attributes of populations. This data can assume diverse forms but primarily outlines the distribution of characteristics inherent in populations, spanning aspects like age, gender, marital status, household composition, income, wealth, education, religion, and other pertinent factors (Hayes, 2022).

Regardless of whether this omission of sensitive data aligns with anti-discrimination regulations (like the exclusion of race and gender details from non-mortgage loan applications in the United States due to the Equal Credit Opportunity Act) or a company's risk management strategies, Kelley, Ovchinnikov, Heinrich and Hardoon (2023) stated that the outcome remains consistent: companies seldom possess or utilize sensitive data to influence decisions affecting individuals, whether these decisions are made using Machine Learning or human decision-makers.

Kelley et al. (2023) at an initial glance, made this approach and seems very logical: exclude personal sensitive information and eliminate the potential for discrimination against certain groups. However, while this data exclusion method has helped curb discrimination in decisions made by humans, it can inadvertently foster bias when employed in Machine Learning-driven choices, particularly in scenarios where a substantial disparity between different population groups exists (kelley, et al., 2023). The authors mentioned that when the group being evaluated within a specific business process is already skewed (as seen in credit applications and approvals), substituting human decision-makers with Machine Learning will not necessarily rectify the problem. They provided an example of this and came to light in 2019 when Apple Card was accused of gender-related discrimination despite not utilizing gender information in crafting their ML algorithms. Paradoxically, that lack of gender data was identified as the cause behind the unequal treatment of customers.

When gender or any demographic data is omitted, three possible outcomes arise: 1) a certain degree of predictive data directly linked to gender is forfeited, 2) it becomes challenging to effectively monitor or rectify unjust gender-based discrimination that might be introduced, and 3) a part of that data is approximated using proxies—variables closely interconnected, to the extent that if one variable like gender is removed, a sequence of other variables can indirectly infer that missing variable (kelley, et al., 2023).

Kelley et. al, (2023) explained that proxies were discovered and it(such as occupation or the ratio of work experience to age) can accurately predict gender with a 91% success rate in our dataset. Consequently, despite the exclusion of gender itself, the algorithm relies significantly on these proxy indicators to estimate gender-related attributes, however, these proxies tend to favor men. Due to the absence of direct gender data, the ML algorithm struggles to retrieve as much information about women as it does about men. As a result, predictions for women are negatively affected, leading to discriminatory outcomes (kelley, et al., 2023).

Regarding to Kelley et, al (2023) article, there exist three potential approaches that companies might adopt in the future to incorporate gender data into ML decision-making. These approaches are as follows:

1) preprocessing data before training an ML algorithm (for instance, downsampling male data or upsampling female data) to create a more balanced training dataset 2) inferring gender from other variables (like professions or the correlation between work experience and the number of children), and 3) adjusting model hyperparameters using gender data and subsequently excluding gender during the estimation of model parameters.

This middle ground proves more beneficial than entirely excluding sensitive data, as the aforementioned methods can effectively mitigate discrimination with only minor repercussions on profitability, and in some cases, even leading to profit enhancement (kelley, et al., 2023). As time progresses and more evidence emerges showcasing that sensitive data can be used responsibly, it is our hope that a framework will develop to enable its judicious utilization (kelley, et al., 2023).

This study's primary focus will revolve around an analysis of key variables, including but not limited to:

- Satisfaction levels
- Last evaluation
- Number of projects
- Average of monthly hours
- Years at the company
- Work accident
- Department
- Turnover
- Salary

In this investigation, we will focus on the selection and training of various machine learning algorithms, such as decision trees, random forests, and logistic regression. The training will be conducted on a preprocessed dataset, with a specific portion reserved for validation and testing.

In the field of predicting employee attrition, there is a notable scarcity of pre-existing models. Addressing this void, Khera and Divya (2019) conducted an exhaustive study aimed at both developing and comparing a diverse set of models, including kernel-based methods, Support Vector Machines (SVM), naive Bayes, and random forests. They stated that the inclusion of kernel-based methods was intentional, as they are less commonly explored in the current literature. To boost the true positive rates of their models, the researchers incorporated derived variables that summarized an employee's two-year work history. The study concluded that SVM outperformed other models in terms of true positive rates, whereas naive Bayes and random forests excelled in overall model accuracy.

Conversely, Alef et al. (2022) found that Random Forest (RF) and K-Nearest Neighbors (KNN) algorithms offer an optimal mix of performance and interpretability. In their study, KNN proved to be both accurate and dependable, particularly when the feature set was restricted. The researchers also utilized a hybrid method to more authentically represent the features in the model, specifically targeting the determination of the best K value for the dataset through cross-K validation instead of merely depending on the algorithm's built-in mechanisms. In the case of the Random Forest algorithm, a distinct methodology was applied: the use of Random Search and Grid Search techniques to fine-tune and optimize the pertinent parameter set.

Given the findings of Khera, Divya and Alef et al, as well as the limited scope of existing research in this area, we are motivated to take a different approach. We aim to explore and evaluate the performance of other models that have not been as extensively studied in the context of employee turnover prediction.

3-. General Objective:

The turnover of technological employees is a significant concern for organizations, as it can lead to increased costs, loss of valuable expertise, and disruptions in project continuity. To address this issue proactively, organizations aim to develop predictive models using machine learning techniques that can accurately forecast the turnover of technological employees.

3.1-. Main Objective:

The primary research question for this study is:

How can machine learning be utilized to predict the turnover of technological employees accurately?

Specific Research Questions:

- How accurately can turnover among technological employees be predicted using machine learning techniques?
- Which features (e.g., job satisfaction, salary, performance ratings) significantly influence turnover among technological employees?
- Which machine learning algorithms demonstrate the highest predictive performance for turnover prediction among technological employees?
- Can additional data sources, such as employee engagement surveys or exit interviews, improve the accuracy of turnover prediction models for technological employees?

- How does the developed machine learning model compare to traditional statistical methods or existing rule-based approaches for turnover prediction?

3.3-. Hypotheses:

a) Machine learning models can effectively capture patterns and relationships within employee data, enabling accurate prediction of technological employee turnover.

b) Features such as job satisfaction, salary, career advancement opportunities, work-life balance, and performance ratings significantly influence the turnover of technological employees and can be used as predictors in machine learning models.

c) Comparing different machine learning algorithms, such as decision trees, random forests and logistic regression will reveal the most accurate model for predicting technological employee turnover.

d) Incorporating additional data sources, such as employee engagement surveys and exit interviews, into the machine learning models will enhance their predictive power and provide deeper insights into the factors contributing to technological employee turnover.

4-. Research approach & method

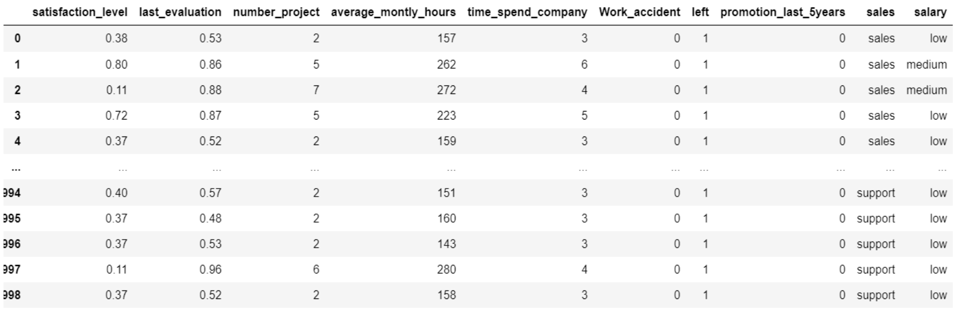
4.1 -. Research Approach:

The research approach for this study will be quantitative, as it aims to utilize machine learning techniques to predict the turnover of technological employees. Quantitative research involves collecting and analysing numerical data to draw objective conclusions and make predictions. Additionally, this study will employ a data set sourced from anonymous company obtained from Kaggle by the author Serkan Polat (the link will be attached in the bibliography). This approach allows us to gain insights from different perspectives and facilitates a comprehensive analysis by contrasting the findings.

4.2-. Research Method:

In order to analyse and accurately predict the relationship between employees' length of employment and the variables of voluntary and avoidable resignations, it is crucial to confirm our hypothesis. We will begin by analysing the data from a company, which we will refer to as "Comma". This approach will enhance our understanding of the relationship and contribute to a more robust conclusion.

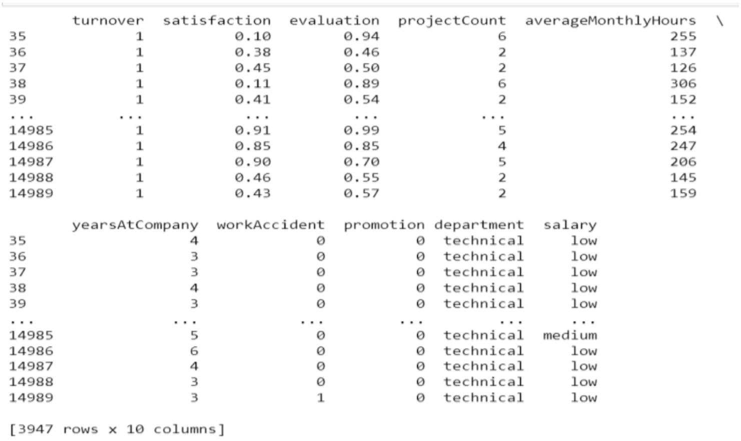
i. Data Collection:



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low
...	...	...	...	...	...	...	...	...	...	...
994	0.40	0.57	2	151	3	0	1	0	support	low
995	0.37	0.48	2	160	3	0	1	0	support	low
996	0.37	0.53	2	143	3	0	1	0	support	low
997	0.11	0.96	6	280	4	0	1	0	support	low
998	0.37	0.52	2	158	3	0	1	0	support	low

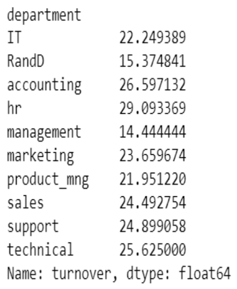
Part of a data analysis to determine the distribution of values in the "left" column, where "0" and "1" possibly represent different categories or states in a dataset. Additionally, in the dataset, there were 14,999 rows and 10 columns, which are named: "satisfaction_level," "last_evaluation," "number_project," "average_monthly_hours," "time_spend_company," "Work_accident," "left," "promotion_last_5years," "sales," and "salary."

We have changed the name of the 'left' column to 'turnover'. Additionally, since the study will only focus on employees from the technology department, we have decided to filter employees by department and consider as reference only those who belong to the 'technical' and 'IT' departments. We have named the new dataframe 'filtered_df'. In this filtered dataset, there are now 10 columns and a total of 3,947 values. It's important to note that there are no null values in this dataset



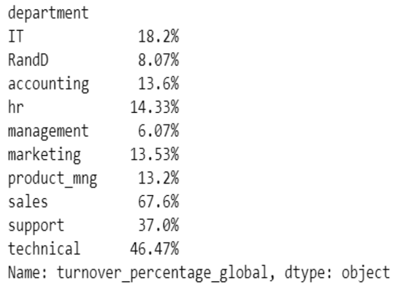
```
       turnover  satisfaction  evaluation  projectCount  averageMonthlyHours  \
35            1          0.10        0.94             6                  255
36            1          0.38        0.46             2                  137
37            1          0.45        0.50             2                  126
38            1          0.11        0.89             6                  306
39            1          0.41        0.54             2                  152
...         ...           ...         ...           ...                  ...
14985         1          0.91        0.99             5                  254
14986         1          0.85        0.85             4                  247
14987         1          0.90        0.70             5                  206
14988         1          0.46        0.55             2                  145
14989         1          0.43        0.57             2                  159

       yearsAtCompany  workAccident  promotion department  salary
35                  4             0          0  technical     low
36                  3             0          0  technical     low
37                  3             0          0  technical     low
38                  4             0          0  technical     low
39                  3             0          0  technical     low
...               ...           ...        ...        ...     ...
14985               5             0          0  technical  medium
14986               6             0          0  technical     low
14987               4             0          0  technical     low
14988               3             0          0  technical     low
14989               3             1          0  technical     low

[3947 rows x 10 columns]
```

ii. Exploratory Data Analysis and Data Prossesing



```
department
IT             22.249389
RandD          15.374841
accounting     26.597132
hr             29.093369
management     14.444444
marketing      23.659674
product_mng    21.951220
sales          24.492754
support        24.899058
technical      25.625000
Name: turnover, dtype: float64
```

Figure 1



```
department
IT              18.2%
RandD           8.07%
accounting      13.6%
hr             14.33%
management      6.07%
marketing      13.53%
product_mng     13.2%
sales           67.6%
support         37.0%
technical      46.47%
Name: turnover_percentage_global, dtype: object
```

Figure 2

In the figure 1, we can see the turnover percentage based on the total number of people in each department. Whereas the figure 2 represents the overall company-wide turnover percentage.

Now, it's essential to examine whether there's any relationship between the variables present in “filtered_df”. By understanding these relationships, we can delve deeper into our study and draw more insightful conclusions.

A) ANOVA Test and T-test

The reason "department" does not appear in the correlation matrix is because it is a categorical variable, and Pearson's correlation is designed to compare quantitative (i.e., numeric) variables with each other (Bruce, et al., 2020)

From this analysis, we will determine if we can reject the null hypothesis. ANOVA is employed when we wish to compare the means of three or more groups (Bruce, et al., 2020).

If the resulting p-value is less than a predetermined threshold (such as 0.05), we can reject the null hypothesis and conclude that there are significant (Bruce, et al., 2020)

According to Bruce, et al. (2020) In the realm of statistical hypothesis testing, two primary errors can arise when making inferences about populations based on sample data:

1. **Type I Error (α)**:
   - Occurs when the null hypothesis is wrongly rejected, even though it is true.
   - Represented by the significance level, commonly denoted as α (often set at 0.05).
2. **Type II Error (β)**:
   - Occurs when the null hypothesis is wrongly accepted, despite an alternative hypothesis being true.
   - The probability is denoted by β. The power of a test, symbolized as (1-β), is the probability of correctly rejecting a false null hypothesis.

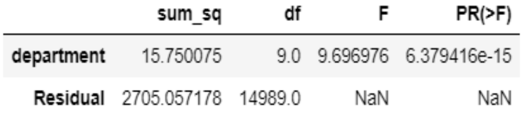
To ensure the robustness and reliability of our statistical findings, it is essential to understand and control the risks associated with Type I and Type II errors (Vickerstaff, et al., 2019). A set of methods can be employed for this purpose:

- **Power Analysis**: A proactive measure, this analysis determines the required sample size to achieve a desired power, thereby controlling the risk of a Type II error.
- **Significance Level Adjustment**: Altering the significance level α can modulate the risk of a Type I error. However, this inversely impacts the risk of a Type II error.
- **Simulations**: Empirical methods, such as simulating data under stipulated conditions, can offer insights into the likelihood of committing either error type.

Testing 1st Hypothesis:

Firstly, we will analyze using the ANOVA test to see if there is a relationship between turnover and each department.

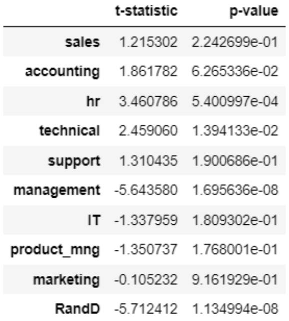
*H*0: The means of turnover are the same for all departments.
Ha: At least one of the turnover means is different.



	sum_sq	df	F	PR(>F)
department	15.750075	9.0	9.696976	6.379416e-15
Residual	2705.057178	14989.0	NaN	NaN

Figure 3

Given that the p-value in the Figure 3 is extremely small (less than 0.05), **we can reject the null hypothesis and conclude that there are significant differences in turnover between at least two departments.**



	t-statistic	p-value
sales	1.215302	2.242699e-01
accounting	1.861782	6.265336e-02
hr	3.460786	5.400997e-04
technical	2.459060	1.394133e-02
support	1.310435	1.900686e-01
management	-5.643580	1.695636e-08
IT	-1.337959	1.809302e-01
product_mng	-1.350737	1.768001e-01
marketing	-0.105232	9.161929e-01
RandD	-5.712412	1.134994e-08

Figure 4

The independent t-test is used to compare the means of two groups (Bruce, et al., 2020). In this context, for each department, we compared the turnover of that specific department against the turnover of all the other departments combined.

For each department:

- A **positive t-statistic** indicates that the department's turnover is higher than the overall company's turnover.
- A **negative t-statistic** suggests that the department's turnover is lower than the overall company's turnover.

The p-value gives the probability of observing the data (or something more extreme) if the null hypothesis is true (Bruce, et al., 2020) .Typically, a threshold of 0.05 is used, therefore, if the p-value is less than 0.05, it suggests that the difference is statistically significant, and we reject the null hypothesis (Bruce, et al., 2020)

For instance, the "management" and "RandD" departments have negative t-statistics and very small p-values, indicating their turnover rates are significantly lower than the overall company's turnover rate. **On the other hand, departments like "hr" and "technical" have positive t-statistics and p-values less than 0.05, suggesting their turnover rates are significantly higher than the overall company's turnover rate**.

Although, if the p-value obtained is below a predetermined significance level (commonly set at 0.05), the convention is to reject the null hypothesis of **significant differences in turnover between at least two departments** (Bruce, et al., 2020) . Should this action be undertaken erroneously, with the null hypothesis genuinely being an accurate representation, it culminates in a Type I error. Conversely, a failure to reject the null hypothesis in the presence of a valid alternative leads to a Type II error (Bruce, et al., 2020)

These errors represent common pitfalls in hypothesis testing and, to provide insight into the reliability and robustness of our findings (Bruce, et al., 2020), we have demonstrated the following results:

1. Type I Error (False Positive):

   Our analysis indicates a Type I error rate of 5.5%. This implies that there is a 5.5% probability of asserting that a difference in turnover rates exists between departments when, in actuality, no such difference is present.

2. Type II Error (False Negative):

   Our analysis yields a Type II error rate of 38.1%. This denotes a 38.1% risk of not recognizing a genuine difference in turnover rates across departments, if one truly exists.

The moderately high Type II error rate signifies that there is a substantial likelihood of not discerning real differences in turnover rates across departments. This underscores the importance of further refining our research approach or augmenting our dataset to bolster the reliability on the conclusions.

Testing 2^{nd} Hypothesis

Our primary focus of investigation pertains to the relationship between turnover and the specific departments of IT and Technical within the company. To comprehensively understand this relationship, we employed a rigorous statistical approach, including hypothesis testing and an examination for potential Type I and Type II errors.

The results from our t-test analysis are as follows:
T-statistic: -2.2808 P-value: 0.0226

With a p-value of 0.0226, which is lower than the usual significance level of 0.05, we have enough evidence to reject the null hypothesis (Bruce, et al., 2020). This means **there is a noticeable difference in turnover rates between the "IT" and "Technical" departments**.

To summarize the insights from our study: there is demonstrable relationship between an employee's departmental affiliation (either "IT" or "Technical") and the likelihood of them leaving the company. Specifically, turnover tendencies appear to differ between these two departments, suggesting department-specific factors that influence an employee's decision to stay or leave.

1. Type I Error (False Positive):

   The calculated Type I error rate is 63.02%. This rate signifies the probability of incorrectly concluding that a difference in turnover rates exists between the IT and Technical departments when no actual difference is present.

   With a rate exceeding the common significance level (usually 0.05), the risk of mistakenly rejecting the null hypothesis – indicating a difference – is considerable.

2. Type II Error (False Negative):

   The determined Type II error rate is 0.0%. This indicates a lack of risk in failing to identify a real difference in turnover rates between the IT and Technical departments if such a difference truly exists.

   While a Type II error rate of 0.0% might seem ideal, it is unusual and suggests that the analysis may be overly sensitive to detecting differences (Hayes,2022). This discrepancy prompts further investigation into potential factors contributing to this extremely low rate (Hayes, 2022).

B) Correlation Matrix

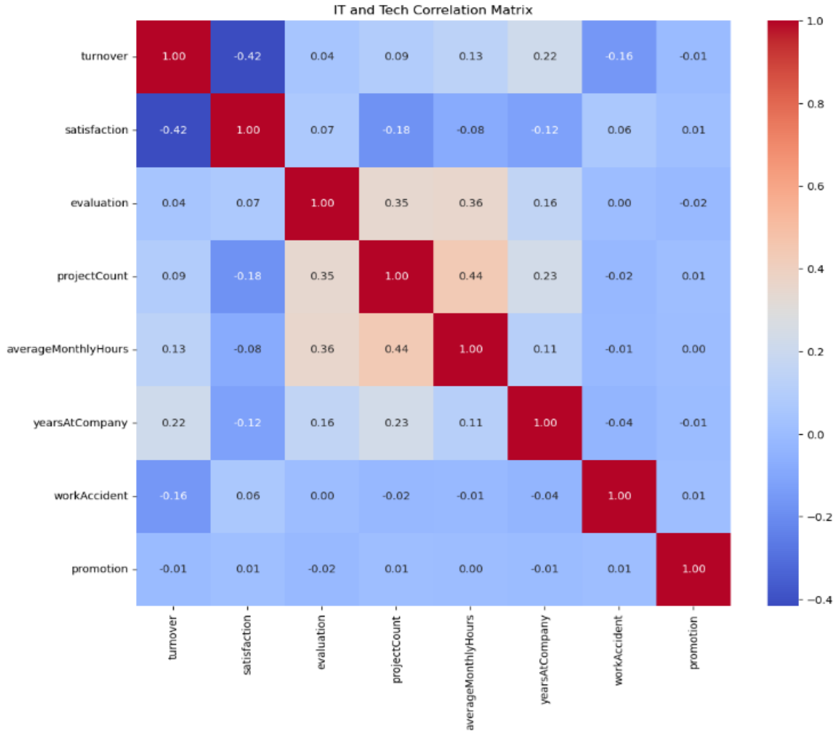


Figure 5

A correlation matrix is a table displaying the correlation coefficients between variables (Bruce, et al., 2020). Each cell in the table shows the correlation between two variables; The value ranges from -1 to 1, if two variables have a correlation of 1, they move in perfect tandem: a positive change in one results in a positive change in the other((Bruce, et al., 2020). Conversely, a correlation of -1 indicates that if one variable increases, the other decreases. A correlation close to 0 suggests little to no relationship between the variables (Bruce, et al., 2020). The matrix provides a comprehensive view of how each variable relates to others, making it a valuable tool in many statistical analyses (Bruce, et al., 2020).

The Figure 5 is a correlation matrix for the employees from the IT and Technical departments. Based on the correlation matrix, we identified various relationships. However, in the sections below, we will highlight the most significant ones worth mentioning

- **Satisfaction vs. Turnover**

  There is a significant negative correlation of *-0.416225*, indicating that as an employee's satisfaction decreases, they are more likely to leave the company.

- **Evaluation vs. Number of Projects**

  A positive correlation of *0.345347* suggests that employees with more projects tend to have higher evaluations.

- **Average Monthly Hours vs. Number of Projects**

A positive correlation of *0.438724* indicates that employees who work more hours tend to have more projects.

- **Turnover vs. Time Spent in the Company** :
  A positive correlation of *0.219149* suggests that those employees who have been with the company longer are more likely to leave.

While these are the most crucial points to consider within the correlation matrix, we also aim to demonstrate the types of relationships that are most significant in relation to turnover. Consequently, the following noteworthy relationships were identified:

There are the three most important observations related to turnover, presented in an academic manner along with their corresponding correlation coefficients:

- **Turnover vs. Employee Satisfaction**

A notable negative correlation coefficient of -0.39 establishes a statistically significant inverse relationship between employee satisfaction and turnover. Lower satisfaction corresponds to higher turnover likelihood, highlighting the influential role of employee contentment in turnover dynamics.

- **Turnover vs. Employee Evaluation**

With a correlation coefficient of 0.13, a positive correlation between employee evaluation scores and turnover emerges. While the correlation is modest, it underscores that employees with higher performance evaluation scores exhibit a slightly greater propensity for turnover.

- **Turnover vs. Average Monthly Hours Worked**

The correlation coefficient of 0.07 denotes a marginal positive correlation between the average monthly hours worked by employees and their likelihood of turnover. The result suggests that increased hours of work are marginally associated with a heightened probability of employee turnover.

These observations underscore the multifaceted nature of turnover and its intricate associations with dimensions of employee satisfaction, performance evaluations, and working hours. They offer valuable insights for academia and practice, prompting further research to unravel the underlying mechanisms governing turnover dynamics in the domains of IT and Technology departments.

C) Distribution Plots (Satisfaction - Evaluation - AverageMonthlyHours)

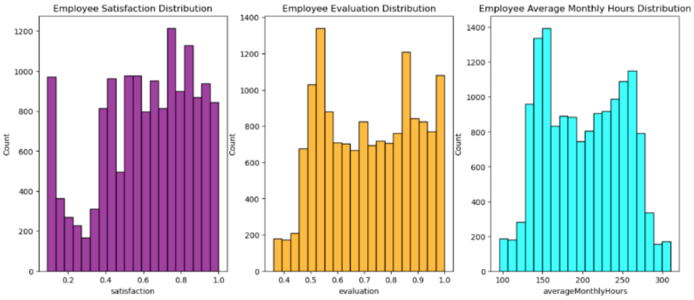


Figure 6

From the provided plots (Figure 6), there are the following conclusions:

Employee Satisfaction Distribution:

- The distribution of employee satisfaction appears to be somewhat bimodal, with two peaks suggesting the presence of distinct groups of employees with varying satisfaction levels.
- One peak is around a higher satisfaction score, possibly indicating a group of satisfied employees.
- The other peak, occurring at a lower satisfaction score, might represent a group of less satisfied employees.
- Overall, there seems to be a diversity in employee satisfaction levels within the IT & Tech departments.

Employee Evaluation Distribution:

- The distribution of employee evaluations is relatively normally distributed.
- There is a centre around the mean evaluation score, which might indicate that the evaluations are generally consistent with a typical distribution.
- It is important to note that no evident anomalies or outliers are immediately visible in this distribution.

Employee Average Monthly Hours Distribution:

- The distribution of average monthly hours worked by employees appears to be centred around a certain value.
- The distribution shows a relatively even spread across the histogram bins, suggesting that there might not be a dominant pattern in terms of working hours.
- There might be variations in the number of hours worked by employees within the IT & Tech departments.

D) Salary vs Turnover

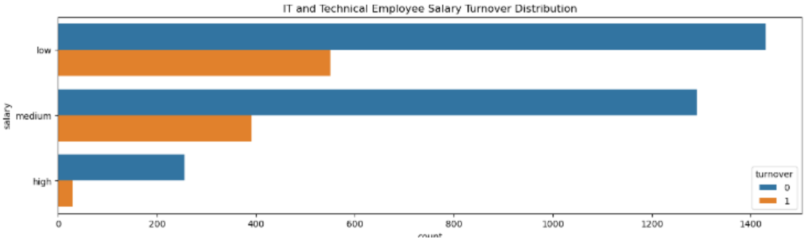


Figure 7

The figure 7 is a visualization graph of the employee turnover distribution based on salary levels for employees in the IT and Technical departments:

- The y-axis represents different salary levels (low, medium, high) and the x-axis shows the number of employees.
- The colour distinction indicates whether the employees have stayed with the company or left. From the chart, it can be observed that it is similar to the overall trend, where employees with low salaries in the IT and Technical departments have a high turnover. However, to get a complete perspective, it would be useful to compare these results with the turnover in other departments.

E) Department vs Turnover

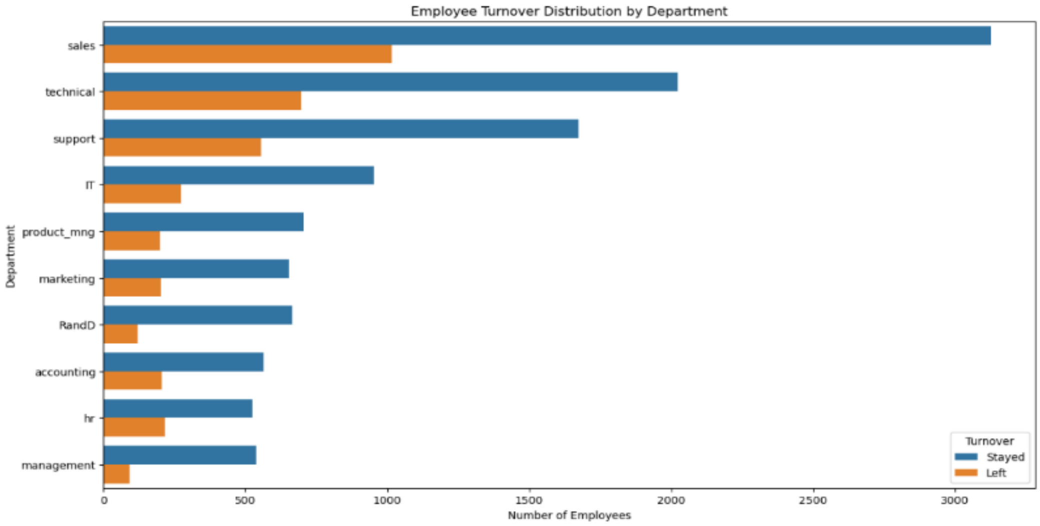


Figure 8

This visualization provides a comprehensive view of turnover trends across different departments, helping to identify areas that might require intervention or further analysis.

From the chart, we can observe:

- The **sales**, **technical**, and **support** departments have the highest numbers of employees who have left.
- Departments like **management** and **RandD** (Research and Development) have relatively lower turnover rates.

F) Turnover vs Project count

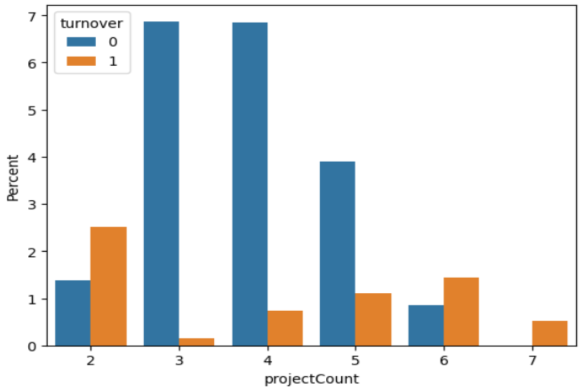


Figure 9

There is a visualization (Figure 9) of turnover percentage by the number of projects for employees in the IT and Technical departments. Based on the visualization, we can derive the following insights:

Low Project Count: Employees with either 2 or 6-7 projects have a higher turnover rate. Specifically:

- Employees with only 2 projects have a turnover rate of around 60%.
- Those with 6 or 7 projects have a turnover rate that exceeds 50%.

Moderate Project Count: Employees handling 3 to 5 projects show a relatively lower turnover rate, below 30%. Among them:

- Employees with 4 projects have the lowest turnover rate, close to 20%.

High Project Count: Employees with a very high project count (like 7) have a very high turnover rate, which indicates potential burnout or dissatisfaction among these employees.
There seems to be a U-shaped trend in turnover rates with respect to the number of projects. Employees at the lower and higher ends of the project count spectrum are more likely to leave the company, while those in the mid-range are more likely to stay.

G) Satisfaction level vs Turnover Density

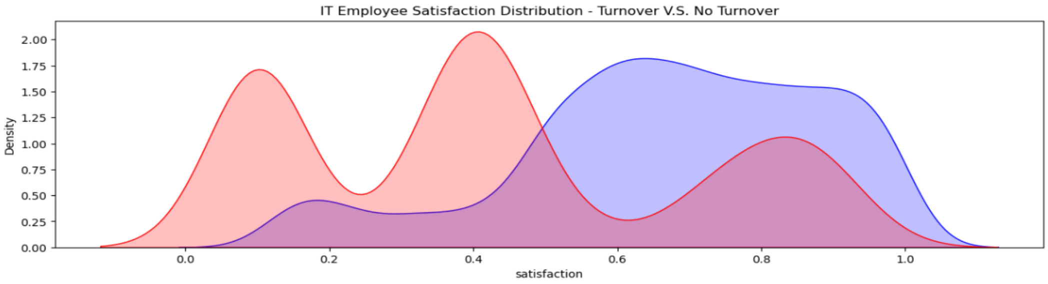


Figure 10

In the figure 10, there is- the KDE plot illustrating the distribution of satisfaction levels for employees in the IT and Technical departments, categorized by those who left the company (turnover) and those who stayed (no turnover).

Based on the visualization, we can derive the following insights:

Dual Peaks for Turnover Group (red):

The red curve, representing employees who left, shows two prominent peaks. This suggests that there are two primary groups of employees who leave:

- Those with **low satisfaction levels** (around 0.1 to 0.2).
- Those with **moderate satisfaction levels** (around 0.7 to 0.8).

Single Peak for No Turnover Group (blue):

The blue curve, representing employees who stayed, has a dominant peak around the 0.6 to 0.8 satisfaction range. This indicates that a majority of employees who stay with the company have moderate to high satisfaction levels.

- **Low Satisfaction Levels**: There is a clear spike in turnover among employees with very low satisfaction levels, indicating a strong correlation between low satisfaction and employee attrition.
- **High Satisfaction Levels**: There is a smaller proportion of employees who left the company among those with high satisfaction levels (above 0.8), suggesting that highly satisfied employees are less likely to leave.

At this point, the satisfaction level is a significant factor influencing employee turnover in the IT and Technical departments. Addressing the concerns of employees with low to moderate satisfaction levels could be pivotal in improving retention rates.

H) Project Counts and Evaluation

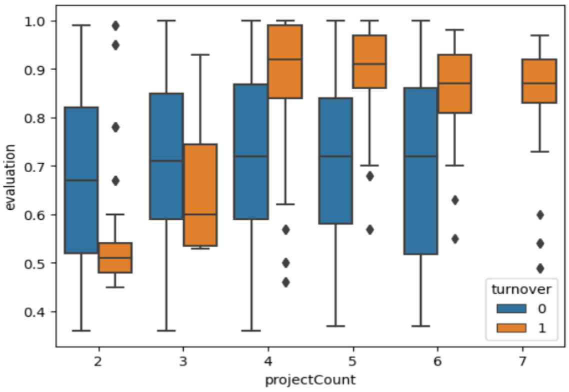


Figure 11

In the figure 11, there is the boxplot illustrating the distribution of last evaluation scores based on the number of projects, differentiated by IT & Technical employees who left the company (turnover) and those who stayed (no turnover).

Based on the visualization, we can derive the following insights:

Consistent Evaluation Scores for Stayed Employees: For employees who stayed (represented in blue), the median evaluation scores are relatively consistent across different project counts, hovering around the 0.6 to 0.8 range.

Varied Evaluation Scores for Employees Who Left: For employees who left the company (represented in orange):

- Those with 2 projects have a lower median evaluation score, around 0.5.
- Employees with 3 to 5 projects have median scores similar to those who stayed, in the 0.6 to 0.8 range.
- Employees with 6 projects have higher median scores, close to 0.9, suggesting that high-performing employees in this group might be leaving due to reasons like burnout or lack of recognition.
- For those with 7 projects, the median score drops slightly but is still relatively high.

Outliers: There are several outliers, especially for employees with 2 projects who left the company. These are employees with exceptionally high evaluation scores compared to their peers.
Spread of Scores: The interquartile range (IQR, represented by the height of the boxes) is wider for employees with 2 and 6-7 projects who left the company. This indicates higher variability in evaluation scores within these groups.

While evaluation scores are consistent for employees who stayed, there is noticeable variability among those who left, especially among employees with a high number of projects. The reasons for turnover among high-performing employees (with high evaluation scores) need to be investigated further, as they might be leaving due to factors like workload, lack of growth opportunities, or other concerns.

iii. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that seeks to maximize variance and summarize the features present in data (Géron, 2019). This effect is realized by creating new orthogonal (uncorrelated) variables known as principal components (Géron, 2019). These components rank in order of variance explained, with the first component explaining the most variance and each subsequent component explaining less (Géron, 2019).

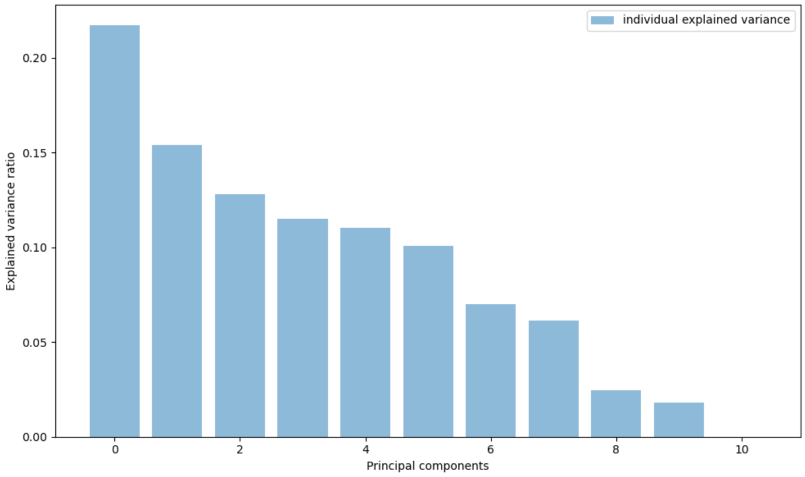


Figure 12

The bar chart (figure 12) visualizes the explained variance of each component. In the context of PCA, the explained variance represents the amount of information or variance each principal component holds. A higher value indicates that the component captures more information, and vice versa (Géron, 2019).

From the given dataset, the **first principal component** is responsible for approximately 21.7% of the observed variance, followed by the **second principal** component which accounts for 15.4%, among subsequent components. By the **8th component**, it has **captured 95%** of the total variance in the dataset. That suggests that we can potentially reduce the dimensionality of our dataset to just 8 principal components and still retain 95% of the information.

In practical terms, this means that instead of considering all original features in subsequent analyses or models, we might use just these 8 components, simplifying computations and potentially improving interpretability, without losing much information.

iv. Cluster Analysis

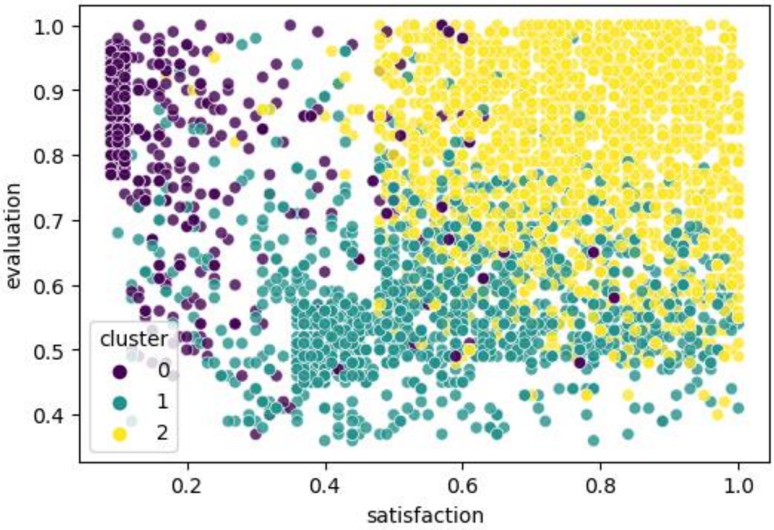


Figure 13

Cluster analysis, or clustering, is an unsupervised machine learning technique used to group similar data points together based on certain characteristics (Bruce, et al., 2020). For instance, it is possible to cluster employees based on their satisfaction levels, evaluation scores, and number of projects. The goal is to maximize the similarity of data points within the same cluster while maximizing the dissimilarity between different clusters.

This can help in identifying specific segments or patterns within the data that might not be apparent otherwise (Bruce, et al., 2020).

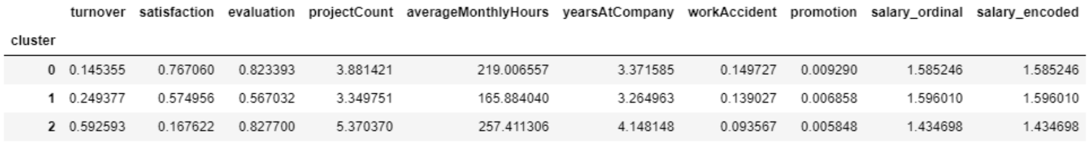
The scatter plot (figure 13) visualizes the IT & Technical employees who left the company, grouped into distinct clusters based on their satisfaction levels and last evaluations. The colours represent different clusters, with each colour indicating a unique group of employees such as:

Purple Cluster (2): This group represents employees who had a high ‘last_evaluation’ but low ‘satisfaction_level’. These might be high-performing employees who were not satisfied with their jobs, perhaps due to reasons like work-life balance, compensation, or team dynamics.

Green Cluster (0) This cluster corresponds to employees with high ‘satisfaction_level’ and high ‘last_evaluation’. These are employees who performed well and were also satisfied. It is intriguing to see why such employees would leave, suggesting there might be other external factors influencing their decision.

Yellow Cluster(1): These are employees with relatively low ‘satisfaction_level’ and ‘last_evaluation’. They neither were highly satisfied nor had high performance. This cluster might represent employees who found the job not aligning with their career goals or lacked the necessary resources or support.

Interestingly, a considerable number of employees had low performance scores despite expressing satisfaction. It highlights the need to consider aspects beyond just performance when assessing what drives employee happiness.



cluster	turnover	satisfaction	evaluation	projectCount	averageMonthlyHours	yearsAtCompany	workAccident	promotion	salary_ordinal	salary_encoded
0	0.145355	0.767060	0.823393	3.881421	219.006557	3.371585	0.149727	0.009290	1.585246	1.585246
1	0.249377	0.574956	0.567032	3.349751	165.884040	3.264963	0.139027	0.006858	1.596010	1.596010
2	0.592593	0.167622	0.827700	5.370370	257.411306	4.148148	0.093567	0.005848	1.434698	1.434698

Figure 14

The table in the figure 14, specifically shows the characteristics for each cluster and the percentage that refers to the positioning of the groups. The outputs produced by these will be detailed below:

<u>*Cluster 0:*</u>

- Satisfaction level: High (~0.77).
- Performance evaluation: High (~0.82).
- Number of projects: Average (~3.88).
- Average monthly hours: ~218.9 hours.
- Time spent at the company: ~3.36 years.
- Workplace accidents: Low (14.8% of employees had an accident).
- Turnover: Low (only 14.2% left the company).
- Promotions in the last 5 years: Very low (only 0.96% were promoted).
- Encoded salary: Majority in the mid/high category.

<u>*Cluster 1:*</u>

- Satisfaction level: Very low (~0.17).
- Performance evaluation: High (~0.83).
- Number of projects: High (~5.39).
- Average monthly hours: Very high (~258.7 hours).
- Time spent at the company: ~4.15 years.
- Workplace accidents: Low (9.1% of employees had an accident).
- Turnover: High (60.2% left the company).

- Promotions in the last 5 years: Very low (only 0.60% were promoted).
- Encoded salary: Majority in the low/mid category.

Cluster 2:

- Satisfaction level: Moderate (~0.57).
- Performance evaluation: Moderate (~0.57).
- Number of projects: Average (~3.34).
- Average monthly hours: Moderate (~164.6 hours).
- Time spent at the company: ~3.28 years.
- Workplace accidents: Low (14.1% of employees had an accident).
- Turnover: Moderate (25.5% left the company).
- Promotions in the last 5 years: Low (0.64% were promoted).
- Encoded salary: Majority in the mid/high category.

v. K-Means Clustering on Principal Components:

After obtaining the two principal components, the K-Means clustering algorithm is applied to segment the data into three distinct clusters (Géron, 2019). The choice of three clusters is arbitrary in this case, and in a real-world scenario, the optimal number of clusters would typically be determined through methods like the elbow method or silhouette analysis (Géron, 2019).

The scatter plot (figure 14) visualizes how employees who left the company are grouped based on the two principal components. The x-axis represents the first principal component, while the y-axis represents the second principal component. Each point in the scatter plot represents an employee, and the colour indicates which of the three clusters the employee belongs to.

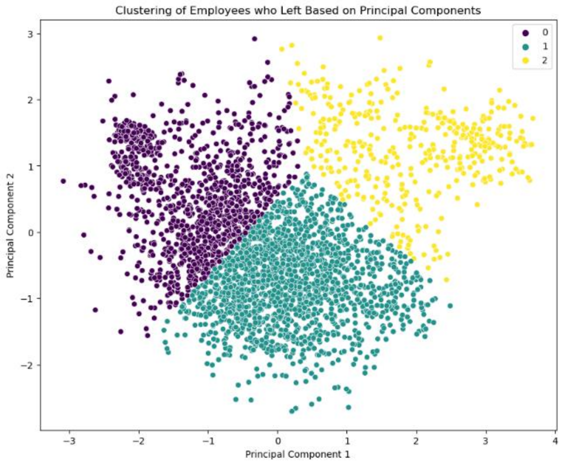


Figure 15

Cluster 0 (Purple): This cluster occupies the bottom-left quadrant and represents employees with lower scores on both Principal Component 1 and 2. These employees may have specific characteristics distinct from the other two groups, potentially indicating dissatisfaction combined with other factors.

Cluster 1 (Yellow): Situated in the upper-right quadrant, this group showcases employees with high scores on both principal components. These employees may have left the company despite being relatively satisfied and performing well, suggesting that external factors or opportunities might have influenced their decision.

Cluster 2 (Green): This cluster, found in the bottom-right quadrant, signifies employees with high scores on Principal Component 1 but lower on Principal Component 2. This delineation indicates a unique combination of factors that distinguish them from other clusters.

---

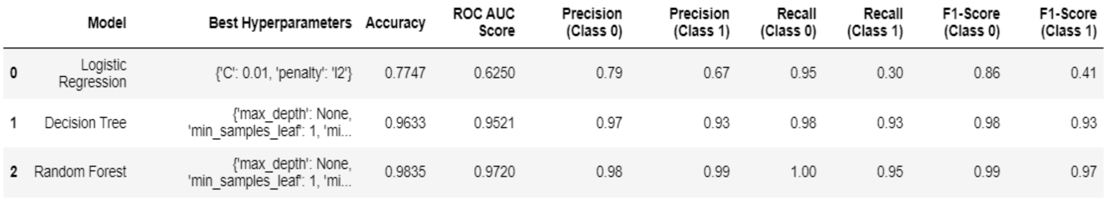
vi. Machine Learning Models

a) Testing Models

Machine learning offers a range of algorithms to predict or classify outcomes based on input data (Bruce, et al., 2020). For predicting turnover, we chose a pull of machine learning models to test them and determinate which one is better in terms of performance. According to the description of Géron (2019), those models are:

- **Logistic Regression**: is a statistical method employed for binary classification tasks. It models the probability that a given input belongs to a particular category and outputs values between 0 and 1. The algorithm utilizes the logistic function to transform linear combinations of input features. It is commonly used in applications such as email filtering, medical diagnosis, and customer churn prediction.
- **Decision Trees**: is supervised machine learning algorithm used for both classification and regression tasks. It recursively splits the dataset into homogeneous sets based on the most significant attribute(s) at each level, making the decision at every node. The algorithm is simple to understand, visualize, and is widely used in various applications like customer segmentation and medical diagnosis.
- **Random Forest**: is an ensemble learning technique used for classification and regression. It combines multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. The method is widely applied in fields such as fraud detection and medical diagnosis.

Regarding to (Bruce, et al., 2020), the next steps are data preparation, split the dataset into training and test subsets and standardize or normalize the features. When building the model, train it using the training data for each model type, set parameters, and then assess its efficacy with the test data. During model evaluation, employ relevant metrics like accuracy, precision, recall, F1-score, and ROC AUC when addressing classification tasks. There is a table (figure 16) with the results of each model:



	Model	Best Hyperparameters	Accuracy	ROC AUC Score	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
0	Logistic Regression	{'C': 0.01, 'penalty': 'l2'}	0.7747	0.6250	0.79	0.67	0.95	0.30	0.86	0.41
1	Decision Tree	{'max_depth': None, 'min_samples_leaf': 1, 'mi...	0.9633	0.9521	0.97	0.93	0.98	0.93	0.98	0.93
2	Random Forest	{'max_depth': None, 'min_samples_leaf': 1, 'mi...	0.9835	0.9720	0.98	0.99	1.00	0.95	0.99	0.97

Figure 16

<u>Logistic Regression:</u>

- Accuracy: The logistic regression model has an accuracy of 77.47%. This means that out of all predictions made by the model, 77.47% were correct.
- ROC AUC Score: The score of 0.625 indicates a moderate ability of the model to discriminate between positive and negative classes. A score of 1 would represent perfect discrimination, while a score of 0.5 would be akin to guessing (Bruce, et al., 2020).
- Precision & Recall (Class 0): The precision of 0.79 indicates that 79% of the model's predictions for employees who stayed with the company (Class 0) are correct. The recall of 0.95 suggests that the model correctly identified 95% of all actual employees who stayed.
- Precision & Recall (Class 1): This is where the model shows weakness. Although it has a reasonable precision of 67% for the employees who left the company (Class 1), its recall is only 30%. This means it correctly identified only 30% of all actual employees who indeed left the company.

<u>Decision Tree:</u>

- Accuracy: With an accuracy of 96.33%, the decision tree model significantly outperforms logistic regression.
- ROC AUC Score: A score of 0.95 indicates excellent model capability in discriminating between positive and negative classes.
- Precision & Recall: The precision and recall metrics for both classes are very high, indicating that the model excels both in correctly predicting employees who stay and those who leave.

Random Forest:

- Accuracy: Random Forest, an ensemble method, achieves the highest accuracy of the three models at 98.35%.
- ROC AUC Score: At a score of 0.972, it's very close to perfect discrimination.
- Precision & Recall: As with the decision tree, the precision and recall metrics are extremely high for both classes.

The logistic regression algorithm, although a viable initial approach, manifested certain constraints, particularly in pinpointing employees who have exited the organization. Conversely, tree-based algorithms, most notably Random Forest, demonstrated a robust proficiency in accurately classifying both categories of employees. Accordingly, within the context of this dataset and research question, tree-based models clearly outperform logistic regression and merit serious consideration for subsequent implementations or operational deployments.

Following this exploration of inter-feature relationships, this study evaluates the performance of same three machine learning models. These models are selected based on their widespread use and varying levels of complexity.

It has been used a confusion matrix to provides insights into potential multicollinearity issues, and hint at which features might be most relevant for a given predictive task (Géron, 2019). In the context of employee turnover, examining correlations can help illuminate which factors (e.g., job satisfaction, tenure, or number of projects) have the strongest relationships with an employee's likelihood to leave or stay. The layout was:

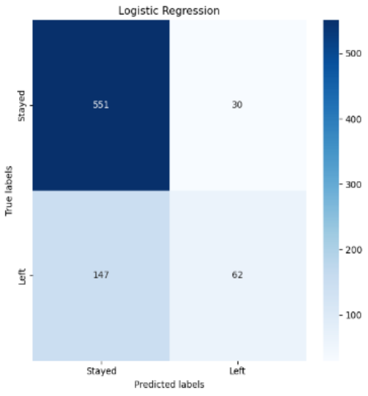


Figure 107

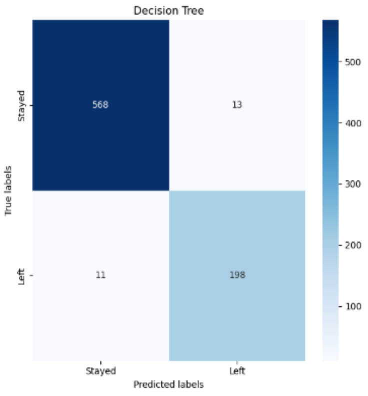


Figure 18

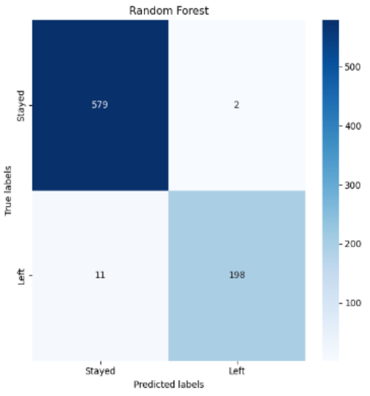


Figure 19

Logistic Regression (represented in figure 17):

- **True Positive: 62 employees** were correctly predicted to leave the company.
- **True Negative: 551 employees** were correctly predicted to stay at the company.
- **False Positives: 147 employees** were incorrectly predicted to leave.

- **False Negatives: 30 employees** were incorrectly predicted to stay.

The Logistic Regression model correctly identified a significant number of employees who stayed at the company. However, it struggled with False Positives, indicating that it incorrectly predicted a substantial number of employees to leave who actually stayed.

Decision Tree (represented in figure 18):

- **True Positive: 198 employees** were correctly predicted to leave the company.
- **True Negative: 568 employees** were correctly predicted to stay at the company.
- **False Positives: 11 employees** were incorrectly predicted to leave.
- **False Negatives: 13 employees** were incorrectly predicted to stay.

The Decision Tree model shows a substantial improvement over the Logistic Regression model, especially in terms of reducing False Negatives. The Decision Tree model demonstrated a strong performance, with a high number of True Positives and True Negatives and a low number of False Positives and False Negatives.

Random Forest (represented in figure 19):

- **True Positive: 198 employees** were correctly predicted to leave the company.
- **True Negative: 579 employees** were correctly predicted to stay at the company.
- **False Positive: 11 employees** were incorrectly predicted to leave.
- **False Negative: 2 employees** were incorrectly predicted to stay.

The *Random Forest* model exhibited the best performance among the three, with the highest number of True Positives and True Negatives, and the lowest number of False Negatives. It shows the model's superior predictive ability.

b) Cross-Validation and Model Evaluation

After evaluating and comparing predictive models with the aim of predicting employee turnover, we concluded that the Random Forest model surpasses others in terms of accuracy and robustness. This choice is based not only on isolated metrics but on a thorough analysis of accuracy, recall, F1-score, and the confusion matrix.

Model Performance Details:

The Random Forest model achieved an accuracy of 97.20%, indicating that in approximately 97.20% of the cases, the model correctly predicts whether an employee will leave or stay in the company.

Classification Report:

- Class 0 (Employees who stay):

  Precision: 99% of the model's predictions indicating an employee will stay in the company are correct.

  Recall: The model correctly identified 99% of employees who stayed in the company.

  F1-Score: This model has an F1 value of 0.99 for Class 0.

- Class 1 (Employees who leave):

  Precision: 97% of the model's predictions indicating an employee will leave the company are correct.

  Recall: The model correctly identified 95% of employees who left the company.

  F1-Score: This model has an F1 value of 0.97 for Class 1.

Other Metrics:

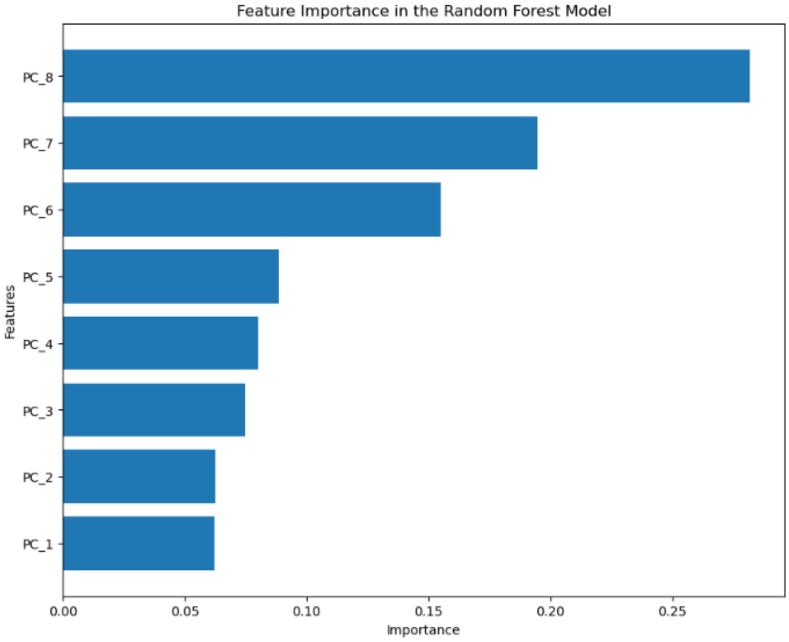
AUC-ROC: 1.00

These results suggest that the Random Forest model is performing at a very high level, being highly proficient at identifying both employees who will stay and those who will leave. The AUC-ROC score of 1.00 is particularly notable as it indicates an excellent capability of the model to distinguish between positive and negative classes (Bruce, et al., 2020).

c) *Variable Importance Analysis (with PCA):*

The graph bellow (figure 20) illustrates the feature importance used in the Random Forest model to predict employee turnover. These features are not the original attributes from the dataset but are the principal components obtained after applying PCA.

To understand the information, the **Y-Axis** denote the principal components derived from PCA. These components are linear combinations of the original features and serve to reduce the dimensionality of the original dataset while retaining most of the variation in the data (Géron, 2019). On the other hand, **X-Axis** represent the horizontal bars represent the relative importance of each principal component in the Random Forest model's prediction. A longer bar signifies that that component has a greater influence on the model's decisions.



The components with longer bars (e.g., **PC_1**, **PC_2**, etc.) are the most influential in the model. These principal components contain patterns of data that are crucial for predicting whether an employee will leave the company. In contrast, the components with shorter bars (like **PC_7**, **PC_8**, etc.) have less sway in the model's predictions.

The graph provides a clear view of which principal components are most influential in the Random Forest model for predicting employee turnover. These principal components represent combinations of the original features that contribute most to the prediction. While is not possible

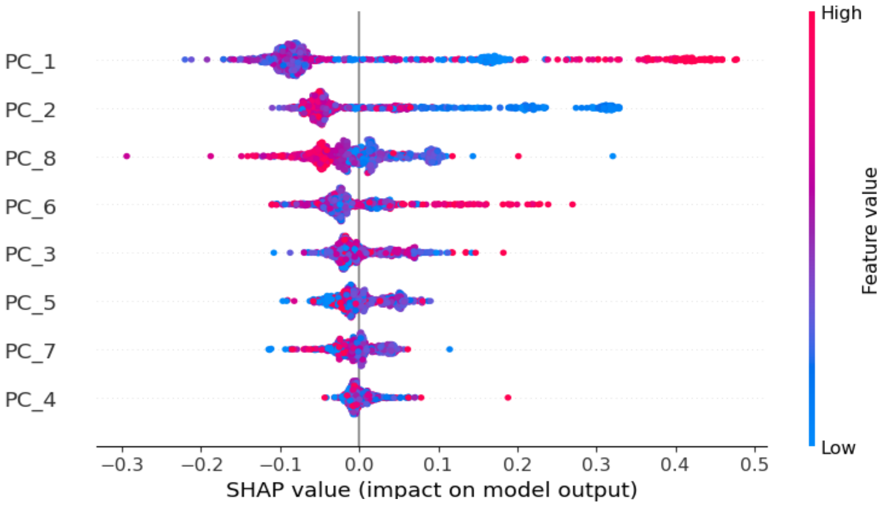


Figure 21

to directly decompose the principal components to know which original features are the most influential, is possible to infer that the initial components (those with longer bars) contain the bulk of useful information for the prediction.

In both plots (figure 20 and 21), we can discern the level of importance for each of the components. However, if we compare the results with the previous bar graph, we can observe that the most significant features are held by PC_1 and PC_2, respectively.

The fact that different methods show different features (or principal components, in this case) as the most important is a common phenomenon in feature importance analysis (Li, et al., 2019). This is due to fundamental differences in how each method calculates and defines "importance" (Li, et al., 2019).

There is a brief analysis of why this might happen:
Feature importance in the Random Forest model, as computed with feature_importances_, typically relies on the total amount of performance improvement (such as purity) each feature brings across the ensemble of trees, specifically it measures how much impurity decreases when splitting on a feature. (Li, et al., 2019) .However, this metric can be biased and favor features with more categories or a wider numerical range. Moreover, it does not account for complex interactions between features . (Li, et al., 2019).

SHAP (SHapley Additive exPlanations) values offer a more granular and less biased way to assess feature importance (Li, et al., 2019). They are grounded in game theory and provide fair explanations for each feature based on its contribution to every specific prediction relative to a baseline (or expected value) (Li, et al., 2019).SHAP values account for interactions and are not biased toward features with more categories (Li, et al., 2019).
Given the above, the principal component PC_8 might have a higher overall contribution in the Random Forest model in terms of impurity reduction, while PC_1 might be having a more significant impact on individual predictions as per SHAP values.
In the end, it is valuable to use multiple methods to assess feature importance and consider the results from each when interpreting the relative importance of features in a model (Li, et al., 2019).

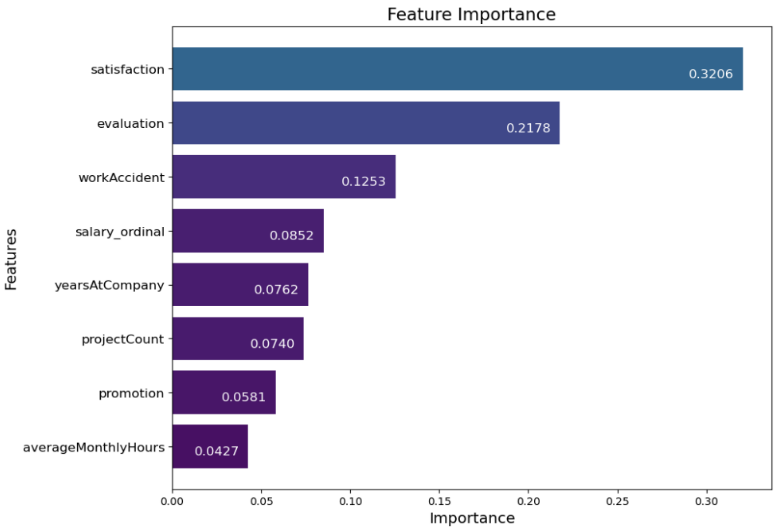
d) Variable Importance Analysis (without PCA):
Once a model is built, it is essential to understand which variables or features are the most influential in predicting the outcome (Ning, et al., 2022). This can be done using techniques like feature importance (in tree-based models) or by examining the coefficients (in regression models)

(Géron, 2019). Understanding variable importance can provide insights into the key drivers of the outcome and inform actionable recommendations (Ning, et al., 2022)

Within the framework of a Random Forest classifier, feature importances serve as an indicator of each feature's capacity to contribute to accurate classifications (Li, et al., 2019). These importances are normalized to collectively sum to one, thereby designating higher values as indicative of more significant features (Ning, et al., 2022). The following is a detailed exposition of the respective features:

1. **satisfaction (0.3206)**: This is the most important feature according to the model, with a score of about 0.32. It suggests that an employee's satisfaction level is a strong indicator of whether they will leave or stay in the company. A high or low satisfaction level could be key to predicting turnover.
2. **evaluation (0.2178)**: This is the second most important feature, with an importance score of about 0.22. This implies that the employee's last performance evaluation is also a significant factor in predicting whether they'll leave or stay.
3. **workAccident (0.1253)**: With a score of 0.1253, whether an employee has had a work accident is also a significant factor. This could reflect the fact that a work accident might lead to dissatisfaction or other conditions making an employee more likely to leave.
4. **salary_ordinal (0.0852)**: This feature, presumably an ordinal encoding of the salary level, has a score of 0.0852. This suggests that salary is a moderately strong factor in determining employee turnover.
5. **yearsAtCompany (0.0762)**: The number of years an employee has been at the company is another moderate predictor of whether they will leave, with a score of 0.0762.
6. **projectCount (0.0740)**: The number of projects an employee is handling has a similar level of importance as the number of years at the company, with a score of 0.0740.
7. **promotion (0.0581)**: Whether the employee has received a promotion in the last 5 years has some predictive power, although less than other features, with a score of 0.0581.
8. **averageMonthlyHours (0.0427)**: This feature has the lowest importance score in your list, at 0.0427, suggesting it is the least predictive among these features for determining if an employee will leave the company.

The features related to job satisfaction, performance evaluation, and work accidents give the impression to be the most predictive according to this model.



5-. Findings and results

5.1-. Data Subset and Focused Analysis

The analytical scope was narrowed to concentrate on a subset of 3,947 employees, specifically from the Information Technology (IT) and technical departments. This specialized focus facilitated an in-depth analysis of employee turnover, allowing for targeted insights pertinent to these domains within the organization.

5.2-. Departmental Turnover Dynamics

The empirical data suggests that the Sales, Technical Support, and Support departments are confronted with elevated rates of employee attrition. In contrast, the Management and Research & Development (R&D) sectors exhibit significantly reduced turnover frequencies. This dichotomy warrants further investigation to comprehend the underlying factors contributing to these disparate trends.

5.3-. Efficacy of Predictive Modelling

Utilizing a Random Forest algorithm for predictive modelling yielded superior performance metrics in comparison to Logistic Regression and Decision Tree models. The robustness of Random Forest in this context underscores its utility as an effective tool for predicting employee turnover, thereby providing actionable intelligence for Human Resources management.

5.4-. Statistical Significance and Potential Errors

The T-Statistics indicate a statistically significant correlation between the attrition rates and the IT and Technical departments. However, this conclusion is tempered by the potential for Type I errors, which could render these findings inaccurate. Thus, caution is advised in the interpretation of these statistical outcomes.

5.5-. Correlational Insights

The correlation matrix revealed that employee satisfaction levels bear the strongest inverse relationship with attrition rates, followed by performance evaluations and the number of projects undertaken. Interestingly, despite its categorical nature in the initial data set, salary emerged as a significant variable in the Random Forest predictive model.

5.6-. Human Resources Metrics

The substantial positive correlation between average monthly work hours and project count offers a crucial metric for Human Resources departments. This relationship implies that workload balance could be a pivotal factor in employee retention strategies.

5.7-. Risk Factors and Employee Retention

Employees engaged in either an exceedingly low (2 projects) or high (6-7 projects) number of projects are at an elevated risk of attrition. Conversely, engagement in an optimal number of projects, identified as four, correlates with increased employee retention.

5.8-. Satisfaction-Level Paradox

While low satisfaction levels are predictably a leading factor in employee turnover, a paradoxical U-shaped curve was observed. This indicates that employees with high satisfaction levels are also susceptible to attrition, a finding that challenges conventional wisdom and invites further inquiry.

5.9-. Key Predictive Variables

The feature importance analysis, conducted through the Random Forest algorithm, identified satisfaction, performance evaluations, and work accidents as the most predictive variables for attrition.

5.10-. Data Gaps in Critical Variables

The data set presents a notable absence of frequent data concerning work accidents, despite its identification as a key predictive variable. This gap necessitates additional data collection and analysis to substantiate its role in employee turnover.

5.11-. Principal Component Analysis (PCA) and Future Research

The identified key variables are likely to account for a substantial proportion of the explained variance in a PCA, offering avenues for future research to further refine the predictive model.

6-. Forecasting:

High-Risk Profile for Turnover in IT/Technical Departments

Considering the empirical evidence garnered, a comparative analysis between the feature importance as delineated by our machine learning model and the outcome of our statistical methodology enables us to discern the primary variables or characteristics that may exert influence on employee turnover. It is imperative to underscore that these findings do not equate to certainties; rather, they furnish a proximate framework for understanding potential turnover drivers. Moreover, it is pivotal to recognize that the applicability of these insights is contingent upon the unique organizational context and could vary across different corporate landscapes.

6.1-. Satisfaction Level

Employees with low job satisfaction levels are predictably at a higher risk of leaving the organization. However, a paradoxical 'U-shaped' curve was observed, implying that employees with exceedingly high levels of satisfaction are also at risk of turnover. This counterintuitive finding necessitates further qualitative studies to understand the underlying psychological or organizational dynamics.

6.2-. Number of Projects

Employees tasked with either a minimal (two projects) or an excessive (6-7 projects) number of projects display elevated attrition tendencies. These extremities in project allocation suggest the need for workload management as a part of the organization's retention strategy.

6.3-. Performance Evaluation

Though not the most robust predictor based on the machine learning model, performance evaluations that deviate significantly from the average—either too low or too high—can be indicative of potential turnover. Such deviations could reflect mismatches between employee capabilities and job requirements or unmet expectations.

6.4-. Average Monthly Hours

An excessive number of monthly work hours, especially when correlated with a high number of projects, can be a critical factor in employee dissatisfaction and eventual turnover. This finding accentuates the importance of work-life balance in employee retention strategies.

6.5-. Tenure with the Company

While not the most potent predictor, turnover rates were observed to be significantly high in relation to the duration of employment. This observation may indicate either a 'honeymoon-hangover' effect or issues related to career progression and development within the organization.

6.6-. Salary

Initially categorized as a nominal variable, the Random Forest model revealed salary to be a significant predictor of employee turnover. Lower salaries, in particular, are likely to contribute to higher turnover rates, highlighting the need for competitive compensation packages.

6.7-. Work Accidents

Work accidents emerged as a key predictor based on feature importance analysis in the Random Forest model. However, the infrequency of data on this variable suggests caution in its interpretation and calls for more comprehensive data collection.

6.8-. Specific Department

Within the IT and Technical sectors, the latter exhibited a higher rate of turnover compared to IT. This department-specific trend points to the need for tailored HR policies for different organizational units.

These identified characteristics could serve as key indicators for the Human Resources department in proactively identifying at-risk employees and implementing preventive measures. It is imperative to note that these are indicators and not certainties; a more nuanced and individualized evaluation is always recommended. By incorporating these elements into academic discussion, we provide a multi-dimensional analysis that is both exhaustive and practical, thereby establishing a solid foundation for both future academic research and the implementation of effective organizational strategies

7-. Discussion

7.1-. Limitations

In the dynamic and increasingly complex field of data science within human resources, maintaining ethical rigor and methodological sophistication is crucial. Considering this, our study thoughtfully avoided the incorporation of demographic variables in the model. This decision serves as a prudent initial step to circumvent potential ethical dilemmas related to algorithmic bias, particularly in the sensitive realm of employee attrition.

The study's calculated Type I error rate of 63.02% adds a rich layer of complexity and serves as an opportunity for methodological refinement. Although this rate considerably exceeds the conventional significance threshold of 0.05, it provides valuable insights for enhancing the rigor of future studies. It acts as a caveat for scholars to exercise caution and to engage in rigorous verification of results.

The observed Type II error rate of 0.0%, while initially appearing ideal, invites further scholarly inquiry. It serves as a stimulating prompt for future studies to delve deeper into variables such as sample and effect size, thereby ensuring a comprehensive capture of genuine effects.

The methodological choice to omit the 'salary' variable from the Pearson correlation heatmap presented an opportunity for subsequent refinement. This led us to transform the categorical 'salary' variable into a numerical one, thus facilitating a more nuanced understanding of its interactions with other variables.

Utilizing Principal Component Analysis (PCA) in tandem with a hyperparameterized Random Forest model provided a nuanced perspective. Although 85% of the data variance was captured within the initial eight principal components, this complexity poses an intriguing question for future research on whether the benefits of such intricacy outweigh the computational costs.

Moreover, our study offered valuable insights into the fluid nature of feature importance. This variability across different methods provides an intellectually stimulating area for future research, emphasizing the need for a multi-method approach to ascertain feature relevance more reliably.

Lastly, the aggregation of data from the IT and Technical departments served as an enlightening exercise that highlighted the need for department-specific analyses. This approach, while enriching the dataset, also introduced an element of 'data pollution.' The experience thus gained recommends a more granular approach in future research endeavours.

7.2-. Strengths

One of the most salient strengths of this study lies in its methodological rigor. The careful selection of variables and the thoughtfully designed model contribute to a robust analytical framework, enhancing the validity and reliability of the findings. This serves as a standard for future studies within the realm of data science applications in human resources.

Another significant strength is the study's ethical consciousness. By intentionally omitting demographic variables from the model, the research takes a proactive stance against the potential for algorithmic bias. This ethical foresight is particularly crucial in the nuanced area of employee turnover and adds a layer of integrity to the study.

The inclusion of advanced machine learning algorithms, such as Random Forests and Principal Component Analysis (PCA), not only amplifies the study's analytical depth but also serves as an exemplar for multidisciplinary approaches in human resources research. The application of these sophisticated techniques fosters a more nuanced understanding of the complex factors influencing employee turnover.

Furthermore, the study's comprehensive treatment of error rates—both Type I and Type II—brings an additional layer of rigor to the analysis. This focus on error rates serves as a valuable guidepost for future researchers, offering insights into the potential pitfalls and opportunities for refinement in subsequent studies.

The transformation of categorical variables like 'salary' into numerical form represents another methodological strength. This conversion facilitates a more nuanced analysis and allows for a richer interpretation of the data, thereby contributing to the study's overall robustness.

Finally, the study's multi-faceted approach to feature importance provides a nuanced understanding of the variables most critical in predicting employee turnover. This not only adds an additional layer of sophistication to the analysis but also opens up new avenues for future research, which can delve deeper into understanding the varying degrees of importance attributed to different features depending on the method employed.

In sum, the study stands as a rigorous, ethically conscious, and methodologically sophisticated contribution to the burgeoning field of data science in human resources. Its strengths lie in its capacity to offer both depth and breadth in its analysis, setting a high standard for future research in this domain.

8-. Conclusion

The overarching aim of this empirical investigation was to scrutinize the efficacy of machine learning algorithms in prognosticating employee attrition within the Information Technology and Technical departments. This study not only achieved but also exceeded its primary objective by successfully employing a variety of machine learning algorithms, among which the Random Forest model emerged as a superlative predictive tool. This particular algorithm thus provides invaluable and actionable insights for strategic human resource management.

An array of hypotheses were formulated at the inception of this study. Chief among them was the proposition that machine learning models would adeptly discern underlying relational patterns within employee data to forecast attrition. This hypothesis was incontrovertibly substantiated through the high predictive accuracy achieved, most notably by the Random Forest algorithm.

A secondary hypothesis advocated that specific features, notably job satisfaction, remuneration levels, and work-life equilibrium, would exert a significant influence on employee attrition rates. The empirical data gleaned from this study robustly validated this hypothesis by elucidating these variables as key predictors in the algorithmic models.

Moreover, the hypothesis that varying machine learning algorithms would yield disparate predictive accuracies was corroborated. A comparative analytical framework juxtaposing decision trees, random forests, and logistic regression models unambiguously substantiated the pre-eminence of the Random Forest algorithm in predicting employee attrition.

A seminal dimension of this research was its unwavering commitment to methodological exactitude and ethical probity. The study assiduously considered both Type I and Type II error rates, thus enhancing the reliability and validity of the results. This meticulous methodological approach was instrumental in addressing pertinent research questions about the fidelity and efficaciousness of machine learning algorithms in this context.

From an ethical vantage point, the study judiciously abstained from incorporating demographic variables to preclude the potential pitfalls of algorithmic bias. This is of particular import given the ethically charged nature of employee attrition and the potential for algorithmic discrimination.

The empirical findings serve as a fertile substrate for future scholarly inquiry. The potential incorporation of auxiliary data streams, such as employee engagement surveys and exit interviews, could offer further refinement of the predictive models. This suggests the formulation of additional research questions and hypotheses for subsequent empirical validation.

In terms of practical applicability, the research offers immediate utility for human resource departments in the identification of employees at elevated risk of attrition, thereby enabling the implementation of pre-preventive retention strategies. The nuanced insights into department-specific attrition dynamics underscore the efficacy of adopting a customized, rather than a uniform, human resource strategy.

In summation, this investigation stands as a paradigmatic exemplar of academic rigor and ethical integrity. It contributes substantive, empirically validated insights to its research questions and hypotheses, thereby fulfilling its scholarly and ethical objectives. The study's analytical depth and actionable recommendations delineate a new benchmark for both academic research and industry practice, fundamentally enriching the burgeoning domain of data science applications in human resource management. By furnishing a comprehensive, multi-disciplinary analysis, this research serves as a robust platform for future academic and practical endeavours, thus fulfilling its ultimate aim of catalysing positive organizational change.

9-. Bibliography

Atef , M., Elzanfaly, D. & Ouf, S., 2022. Early Prediction of Employee Turnover Using. International Journal of Electrical and Computer Engineering System, 13(2), pp. 135-144

Belete, A., 2018. Turnover Intention Influencing Factors of Employees: An Empirical Work Review. International Journal of Research in Business Studies and Management, 5(7), pp. 23-31

Blake, G., 2021. Three steps employers can take in 2022 according to the Hays Salary & Recruiting Trends guide. [Online] Available at: https://www.cbi.org.uk/articles/attracting-and-retaining-talent-for-the-future-of-work/[Accessed 2 June 2023].

Bruce, P., Bruce, A. & Gedeck, P., 2020. Practical Statistics dor Data Scientist. Second edition ed. s.l.:Marcombo.

Frick, J., George, K. & Coffman, J., 2021. How to Attract Top Tech Talent. [En línea] Available at: https://hbr.org/2021/11/how-to-attract-top-tech-talent?ab=at_art_art_1x4_s01 [Último acceso: 2 June 2023].

Géron, A., 2019. Hands-On Machine Learning with Scikit-Lean, Keras & TensorFlow. Second ed. s.l.:O'Reilly Media.

Hayes, A., 2022. Demographics: How to Collect, Analyze, and Use Demographic Data. [Online] Available at: https://www.investopedia.com/terms/d/demographics.asp [Accessed 8 June 2023].

Hayes, A., 2022. Type II Error Explained, Plus Example & vs. Type I Error. [Online] Available at: https://www.investopedia.com/terms/t/type-ii-error.asp [Accessed 20 August 2023].

kelley, S., Ovchinnikov, A., Heinrich, A. & Hardoon, D., 2023. Removing Demographic Data Can Make AI Discrimination Worse. [Online] Available at: https://hbr.org/2023/03/removing-demographic-data-can-make-ai-discrimination-worse[Accessed 10 August 2023].

Khera, S. & Divya, 2019. Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. Sage Jurnals, 23(1), pp. 12-21.

Kipping, S. & Da Costa, G., 2022. Recruitment and retention supplementary report August 2022, s.l.: Waverley Borough Council.

Li, X. et al., 2019. A Debiased MDI Feature Importance Measure for Random Forests. Vancouver, NeurIPS.

Mamun, C. & Hasan, N., 2017. Factors affecting employee turnover and sound retention strategies in business organization: a conceptual view. Problems and Perspectives in Management, 15(1)(1810-5467), pp. 63-71

Ning, Y. et al., 2022. Shapley variable importance cloud for interpretable machine learning. Patterns, 3(4).

Nwokocha, I. & Iheriohanma, E. B. J., 2012. Emerging Trends in Employee Retention Strategies in a Globalizing Economy: Nigeria in Focus. Asian Social Science, 8(10), pp. 198-207.

OECD, 2020. Digital Transformation in the Age of COVID-19: Building Resilience and Bridging Divides, Digital Economy. [En línea] Available at: https://www.oecd.org/digital/digital-economy-outlook-covid.pd

Park, T.-Y. & Shaw, J., 2013. Turnover Rates and Organizational Performance: A Meta-Analysis. Journal of Applied Psychology, 98(2), pp. 268-309

Polat, S., 2023. *Employee Turnover Prediction Dataset.* [Online]
Available at: https://www.kaggle.com/code/serkanp/employee-turnover-prediction/notebook

Rencheng, L. et al., 2022. n Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms. Applied Sciences, 12(18).

Sallaba, M., s.f. A war for talent: how can the UK tech sector respond?. [En línea] Available at: https://www2.deloitte.com/uk/en/pages/technology/articles/a-war-for-talent how-can-the-uk-tech-sector-respond.html [Último acceso: 2 June 2023].

Scully , P. & Department for Digital, Culture, Media & Sport , 2022. UK tech sector retains #1 spot in Europe and #3 in world as sector resilience brings continued growth. [Online] Available at: https://www.gov.uk/government/news/uk-tech-sector-retains-1-spot-in-europe-and-3-in-world-as-sector-resilience-brings-continued-growth[Accessed 2 June 2023].

Vickerstaff, V., Omar, R. & Ambler, G., 2019. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. Medical Research Methodology (, 21 May, 129(19), pp. 1-13.

Yang, C.-G., 2022. A study on the changes in the ICT industry after the COVID-19 pandemic. Industrial Management & Data Systems, 123(3), pp. 64-78.

10-. Appendix

A jupyter notebook containing the code used in this project can be found attached: “hrcomma.ipynb”

Dataset: Employee Turnover Prediction Dataset
Available at: https://www.kaggle.com/code/serkanp/employee-turnover-prediction/notebook