جامعة عجمان
AJMAN UNIVERSITY

College: Engineering & Information Technology

Department: Information Technology

Major: Data Analytics

Busi. Intell.&Data Warehousing - INS402

Employee Attrition

Team Members :

Haniyah Alzaben – 202110616

Leena Alsalhi – 202110613

Layan Ahmad – 202110844

Alya Alabdouli – 202111004

Maryam Fadhel – 202110910

Supervised by:

Dr. Ghazi AlNaymat

Due Date: 6 Dec 2024

Academic Year 2024- 2025 – Fall

Contents

Abstract

Most of the companies do know that employees are the basis for a successful company yet retaining them remains one of HR's toughest challenges. Understanding the factors that cause attrition and precisely predicting the time when these things occur will empower organizations to anticipate-and thus sometimes save-the valuable resource before it walks out the door. Employee Attrition Prediction and Prevention System (EAPPS) is a solution designed to deal with the issue of employee turnover. By Using historical employee data through machine learning models, EAPPS predicts whether employees will stay within or leave the organization, providing HR teams useful information to support their decision. Main features include attrition prediction, factor analysis, employee segmentation, and scenario simulation. Dashboard interactivity coupled up with an alert system would ensure that such decisions are data driven with timely intervention. EAPPS will help the organizations in managing attrition risks, increasing employee satisfaction and building a steady, high performing workforce.

Introduction

Employee attrition is a serious challenge for organizations, and it causes a huge financial loss, decreased productivity, and diminished morale among the workplaces. The Employee Attrition Prediction and Prevention System (EAPPS) comes with a solution by forecasting through machine learning and data analytics the reasons behind employee turnover and provides effective retention strategies for organizations. Some key features of the system include an attrition prediction module which classifies employees into two classes known as "Active" or "Not Active" indicating that and employee will stay or leave respectively, analysis of key factors that result in attrition, employee segmentation, retention recommendations by the EAPPS into risk groups, and interactive dashboards to monitor trends. The feature also includes a scenario simulator which allows one to play around with the intervention by EAPPS and provides alert notifications for high-risk employees to the manager. Because of the provision of proactive strategies for retention instead of the old reactivity model, organizations can improve employee satisfaction and save costs associated with employee turnover while having a workforce that is stable and productive.

Overview of the Business Scenario

Employee turnover creates a major issue for modern organizations since it negatively affects productivity, incurs recruitment costs, and generally lowers the morale of what is remaining in a team after most quits. Most prominent causes for high turnover are low job satisfaction, no opportunity for further growth, or pay disparity. Such typical reactions to attrition adopt a modal shift from replacement to prevention.

In other words, The Employee Attrition Prediction and Prevention System (EAPPS) tackles this difficult concept by offering a proactive, data-driven solution. Essentially, the EAPPS employs its analytical models based on historical employee data concerning turnover patterns, signals, and trends. Then, it classifies employees under different attrition risks with machine learning-enabled modules from which the HR teams can learn which workforce to emphasize. EAPPS thus predicts attrition but also signifies the strong predictors, such as low job satisfaction or fewer promotions.

Segment employees into risk groups and prescribe interventions involving HR teams in testing impact predictions of possible solutions through scenario simulations, thus utilizing EAPPS as a strategic edge. Using interactive dashboards and visual representations in the attrition trend monitoring process, the recovery from emerging risks is also made effective through the alert system for early warning intervention.

In today's competitive talent environment in which organizations are strategizing and planning on ways for EAPPS to help retain talents within organizations by reducing turnover costs, creating an encouraging work atmosphere, this guarantees a successful outcome into the future.

Business Objectives

The general business goal of this project is to create an Employee Attrition Prediction System, which will predict whether employees will leave the organization using machine learning methods. The problem it solves is one related to employee turnover and incorporates key determinants in any form of attrition.

Here are some of the business objectives:

1. **Predict Employee Attrition**: The main objective of this system is to accurately predict which employees are more likely to leave the company. The system will analyze past employee data to predict employee attrition risk by using complex machine-learning algorithms, which will help in early identification of at-risk employees will enable HR teams to reduce turnover by addressing the concerns of those employees.

2. **Interactive Dashboards for Enhanced Decision Making**: Our second main objective of this system is to create interactive dashboards, which will enable the HR team to monitor and interpret complex information which will help stakeholders make better decisions.

3. **Develop Tailored Retention Strategies:** This system is designed to develop focused retention programs with valid insights based on the profile of each employee. For instance, leadership propositional assignments can be offered to the employees who want to grow in their career and assignments with flexible working hours or reduced workload will be assigned to those employees who are suffering from burnout.

4. **Understand Key Drivers of Attrition:** Employee Attrition Prediction System deals with most of the business objectives which are very much required to improve employee stability, reduce costs, and improve employee satisfaction. Using the system will allow organizations to predict attrition in advance and understand its root causes for strategic interventions.

5. **Reduce Attrition Costs:** high staff turnover rate results in high financial costs which causes problems for the companies. This project will aim at reducing these costs by identification and addressing of the attrition risks before the decision to leave is made.

6. **Enhance Strategic Workforce Planning:** firms may achieve long-term effectiveness through workforce planning. Through the different insights it would offer, the system will help HR teams

make sure that their initiatives are in line with company objectives, it will also help in Understanding the change in workforce dynamics and the attrition patterns which will enable the organizations to enhance decision making and planning for future talent requirements.

Overall, the system helps the company to achieve a number of business objectives like employee satisfaction, cost reduction, and stability of workforce; it also facilitates will help companies to act faster in managing personnel by not only predicting the attrition but also understanding the factors that may cause it; hence intervention can be strategically carried out.

List of Questions That Can Be Answered by the Analytical Model

1. Which employees are most likely to leave the organization in the next six months?
2. What are the leading causes contributing to employee attrition?
3. Does the work-life balance and job satisfaction have impact on employee retention?
4. Which roles/ departments have the highest/ lowest employee engagement?
5. What characteristics-cum-attributes-for example, experience in years, educational level, training-are truly indicative of high performing employees?
6. What type of training programs develops the biggest improvement in performance?
7. Are employees who complete training programs more likely to stay with the organization?
8. Which occupations have the highest levels of labor turnover?
9. Are newly hired employees more likely to quit within their first year than longer-tenured employees?
10. Does age impact the attrition rate, and if so, how?
11. Are there real differences in attrition rates between the genders?
12. Does level of education have a direct relation to the chances of leaving?

Data Understanding

i. Snapshot of the Used Data

In our project, we have utilized 4 datasets that contain important information about the employees that work in the company. These datasets were merged to provide a complete picture of employee information, recruitment, training and engagement.

1. Employee Data:

This dataset contains personal information about the employees of the company. This dataset serves as the core for analyzing the employee lifecycle patterns and attrition details.

EmpID	FirstName	LastName	StartDate	ExitDate	Title	Supervisor	ADEmail	BusinessUnit	EmployeeStatus	...	Division	DOB	State	JobFunctionDescription	GenderCode	LocationCod
1001	Susan	Exantus	29-Aug-19	NaN	Software Engineer	Angela Carlson	susan.exantus@bilearner.com	BPC	Active	...	Engineers	21-09-1957	MA	Engineer	Female	174
1002	Sandra	Martin	12-Dec-22	28-May-23	Software Engineer	Angela Hayes	sandra.martin@bilearner.com	NEL	Active	...	Catv	08-07-1950	MA	Foreman	Female	213
1003	Keyla	Del Bosque	08-Mar-23	15-Mar-23	Software Engineer	Christina Copeland	keyla.del bosque@bilearner.com	WBL	Active	...	Field Operations	23-11-1973	MA	Foreman	Female	217
1004	Andrew	Szabo	29-May-20	05-Mar-21	Software Engineer	Jennifer Cohen	andrew.szabo@bilearner.com	PYZ	Active	...	Project Management - Con	27-01-1957	MA	Coordinator	Male	214
1005	Luke	Patronick	16-Sep-22	NaN	Software Engineer	Mr. Jesus Richards	luke.patronick@bilearner.com	SVG	Active	...	Field Operations	28-07-1970	MA	Project Manager	Male	184
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
3996	Leonel	Combs	01-Sep-21	NaN	Database Administrator	Summer Carter	leonel.combs@bilearner.com	MSC	Active	...	Catv	29-11-1999	MA	Laborer	Female	8228
3997	Carmen	Frost	16-Jul-19	NaN	Database Administrator	Jordan George	carmen.frost@bilearner.com	EW	Active	...	Engineers	02-12-1946	TX	Project Manager	Male	7256
3998	Ace	Krause	24-Dec-21	27-Jul-22	Database Administrator	Brittany Rubio	ace.krause@bilearner.com	CCDR	Terminated for Cause	...	Yard (Material Handling)	22-01-1971	TX	Administrative	Female	8255

2. Training and Development Data:

The information in this dataset could be used to track the employees' participation in training programs, the results of these training programs and the costs. This could aid in determining the effectiveness of the training programs that the company offers as well as how they affect the work productivity of the employees and their retention.

Employee ID	Training Date	Training Program Name	Training Type	Training Outcome	Location	Trainer	Training Duration(Days)	Training Cost
1001	21-Sep-22	Customer Service	Internal	Failed	Port Greg	Amanda Daniels	4	510.83
1002	19-Jul-23	Leadership Development	Internal	Failed	Brandonview	Brittany Chambers	2	582.37
1003	24-Feb-23	Technical Skills	Internal	Incomplete	Port Briannahaven	Mark Roberson	4	777.06
1004	12-Jan-23	Customer Service	Internal	Completed	Knightborough	Richard Fisher	2	824.30
1005	12-May-23	Communication Skills	External	Passed	Bruceshire	Heather Shaffer	4	145.99
...	...	...	...	...	...	...	...	...
3996	09-Jan-23	Customer Service	Internal	Failed	North Emily	Carmen Cortez	2	808.51
3997	19-Sep-22	Project Management	External	Failed	West Zachary	Katrina Parker	4	629.16
3998	26-Sep-22	Customer Service	External	Completed	Port Kyle	Andre Donaldson	1	994.09
3999	02-Jul-23	Leadership Development	Internal	Incomplete	Williamsland	Brian Obrien	5	477.78
4000	06-Oct-22	Leadership Development	Internal	Completed	Jamesfurt	Michael Mckenzie	4	951.27

3. Recruitment Data:

This dataset records information related to the recruitment process of the employees, It sheds light on the hiring process's effectiveness in addition to the characteristics of successful applicants.

Applicant ID	Application Date	First Name	Last Name	Gender	Date of Birth	Phone Number	Email	Address	City	State	Zip Code	Country	Education Level	Years of Experience	Desired Salary	Job Title	Status
1001	03-Jun-23	Scott	Sheppard	Male	31-08-1992	421-429-7655x39421	perezjanet@example.org	597 Smith Point	Hollandfort	NV	57588	Micronesia	High School	8	60103.21	Chief Technology Officer	Interviewing
1002	15-May-23	Stanley	Lewis	Male	29-04-1965	+1-451-574-5308x1681	grossmark@example.com	8116 Stuart Loop	Port Margaretfurt	TN	14726	Greenland	Bachelor's Degree	17	64575.84	Designer, furniture	Rejected
1003	04-Aug-23	Javier	Li	Female	10-03-1973	(858)901-5499	katiemaldonado@example.com	5940 Barr Villages Suite 075	Dianaland	TX	4699	China	PhD	20	39422.71	Sound technician, broadcasting/film/video	Rejected
1004	28-Jul-23	Christopher	Johnston	Other	04-04-2001	(853)681-1839x2010	sheila73@example.com	442 Lewis Mount	Youngfurt	GA	34455	Ghana	High School	8	51045.11	Air cabin crew	Rejected
1005	05-Jun-23	Melissa	Hicks	Other	17-06-1978	364-575-8478x67812	emilypatterson@example.org	95961 Taylor Circles Apt. 169	East Ashleyborough	IN	21014	Solomon Islands	Master's Degree	0	52792.86	Art therapist	Interviewing
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3996	18-May-23	Melinda	Butler	Male	25-11-1993	001-324-747-3549x4392	davisvickie@example.org	1307 Stephen Walk Apt. 998	West Jennifer	WY	17130	India	PhD	20	65428.69	Psychologist, forensic	Offered
3997	12-Jul-23	Latasha	Johnson	Male	29-01-1978	(870)454-9481x49987	gpeterson@example.net	2945 Morse Wells	New Lindsey	MH	81966	Turkey	Bachelor's Degree	0	37297.03	Operational investment banker	Applied
3998	19-Jun-23	Cheryl	Gill	Other	08-02-2000	001-847-873-3665	bschultz@example.net	36716 Kevin Locks Suite 422	Schmidtfurt	ME	59399	Maldives	High School	10	31093.70	Petroleum engineer	Rejected
3999	16-May-23	Danielle	Villegas	Female	08-08-1994	(385)467-6434x67311	alvarezstephen@example.net	0983 Jeremy Burgs	Michaelhaven	KY	37855	Ghana	Bachelor's Degree	14	59442.38	Housing manager/officer	Applied
4000	07-Jul-23	Charles	Hernandez	Female	01-06-1980	(772)767-2580	murrayallison@example.com	146 Cheryl Highway	Hallland	OR	8592	Netherlands	Bachelor's Degree	1	89853.85	Loss adjuster, chartered	Rejected

4. Employee Engagement Survey Data:

This dataset presents survey results that demonstrates levels of engagement, satisfaction, and work-life balance. This information can be beneficial in analysing the sentiment of the employees and to identify areas for improvement in the company.

	Employee ID	Survey Date	Engagement Score	Satisfaction Score	Work-Life Balance Score
0	1001	10-10-2022	2	5	5
1	1002	03-08-2023	4	5	3
2	1003	03-01-2023	2	5	2
3	1004	30-07-2023	3	5	3
4	1005	19-06-2023	2	4	5
...	...	...	...	...	...
2995	3996	14-04-2023	3	5	1
2996	3997	10-09-2022	2	4	1
2997	3998	22-02-2023	5	5	2
2998	3999	02-10-2022	5	4	2
2999	4000	21-04-2023	3	2	1

ii. Data Attributes

1. Employee Data

Column Name	Description
Employee ID	Unique identifier for each employee in the organization.
First Name	The first name of the employee
Last Name	The last name of the employee
Start Date	The date when the employee started working for the organization.
Exit Date	The date when the employee left or exited the organization (if applicable).
Title	The job title or position of the employee within the organization.
Supervisor	The name of the employee's immediate supervisor or manager.
Email	The email address associated with the employee's communication within the organization.
Business Unit	The specific business unit or department to which the employee belongs.
Employee Status	The current employment status of the employee (e.g., Active, On Leave, Terminated).
Employee Type	The type of employment the employee has (e.g., Full-time, Part-time, Contract).
Pay Zone	The pay zone or salary band to which the employee's compensation falls.
Employee Classification Type	The classification type of the employee (e.g., Exempt, Non-exempt).
Termination Type	The type of termination if the employee has left the organization (e.g., Resignation, Layoff, Retirement).
Termination Description	Additional details or reasons for the employee's termination (if applicable).
Department Type	The broader category or type of department the employee's work is associated with.
Division Description	The division or branch of the organization where the employee works.
DOB	The date of birth of the employee.
State	The state or region where the employee is located.
Job Function	A brief description of the employee's primary job function or role.
Gender	A code representing the gender of the employee.
Location	A code representing the physical location or office where the employee is based.
Race	A description of the employee's racial or ethnic background (if provided).
Marital Status	The marital status of the employee (e.g., Single, Married, Divorced).
Performance Score	A score indicating the employee's performance level (e.g., Excellent, Satisfactory, Needs Improvement).
Current Employee Rating	The current rating or evaluation of the employee's overall performance.

2. Training and Development Data

Column Name	Description
Employee ID	A unique identifier for each employee who participated in the training program.
Training Date	The date on which the training session took place.
Training Program Name	The name or title of the training program attended by the employee.
Training Type	The categorization of the training, indicating its purpose or focus (e.g., Technical, Soft Skills, Safety).
Training Outcome	The observed outcome or result of the training for the employee (e.g., Completed, Partial Completion, Not Completed).
Location	The physical or virtual location where the training session was conducted.
Trainer	The name of the trainer or instructor who facilitated the training.
Training Duration (Days)	The duration of the training program in days.
Training Cost	The cost associated with organizing and conducting the training program.

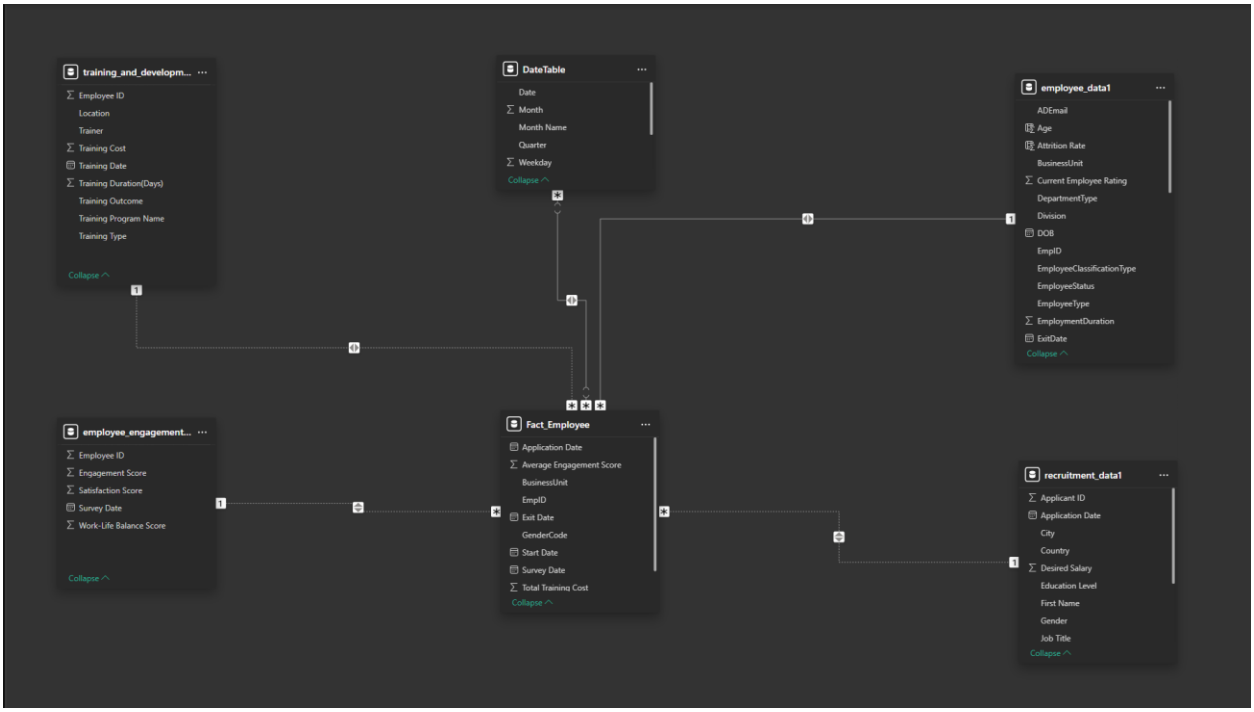
3. Recruitment Data

Column Name	Description
Applicant ID	A unique identifier assigned to each applicant who has submitted their information for a job opportunity.
Application Date	The date on which the applicant submitted their application for the job.
First Name	The first name of the applicant.
Last Name	The last name of the applicant.
Gender	The gender of the applicant.
Date of Birth	The birthdate of the applicant.
Phone Number	The contact phone number of the applicant.
Email	The email address of the applicant for communication purposes.
Address	The street address where the applicant resides.
City	The city where the applicant's address is located.
State	The state or province where the applicant's address is situated.
Zip Code	The postal or ZIP code associated with the applicant's address.
Country	The country where the applicant's address is located.
Education Level	The highest level of education attained by the applicant.
Years of Experience	The number of years of professional experience the applicant has.
Desired Salary	The salary the applicant wishes to receive for the job.
Job Title	The title or designation of the job the applicant is applying for.
Status	The status of the applicant's application.

4. Employee Engagement Survey Data

Column Name	Description
Employee ID	A unique identifier assigned to each employee who participated in the employee engagement survey.
Survey Date	The date on which the engagement survey was administered to employees.
Engagement Score	A calculated numerical score representing the level of employee engagement based on survey responses.
Satisfaction Score	A numerical score indicating employee satisfaction with various aspects of their job and workplace.
Work-Life Balance Score	A numerical score reflecting employee perceptions of the balance between work and personal life.

iii. Dimensional Modeling:



For our project, we have utilized Power BI to create the dimensional model. Power BI's modelling features made it easy for us to establish different relationships between the tables, create a fact table, and connect it with the tables for swift analysis.

The objectives of creating this dimensional model were to simplify the analysis of employee data from several datasets, create a unified data model that facilitates dashboards and reporting, and designing dimension tables that promotes more efficient querying.

a. Overview of the Dimensional Model

This **dimensional model** consists of:

1) Fact Table:

We have created the fact table (Fact_Employee) by combining key metrics from the dimensions that we have, to ensure a unified view for analysis. The (Fact_Employee) table functions as the main table for analysis as it has critical metrics, such as: average_engagement_score, attrition_rates, and total_training_cost. It also links to the other tables' foreign keys an example would be the Employee_ID. The fact table has one-to-many relationships with other dimensions

2) Dimension Tables:

Employee Data:

Has detailed information about employees, like the features mentioned above in the attributes. This table can help us in categorizing and filtering employee-related analysis.

Training and Development Data:

Tracks information related to employees' training programs/costs, etc, it supports analysis related to training outcomes and costs.

Recruitment Data:

This dimension facilitates recruiting performance analysis.

Employee Engagement Survey Data:

This dimension can help us in analyzing the sentiment of the employees.

Date Table:

For temporal analysis, a standard date dimension was created to link all date fields (training date, application date and survey date).

This dimensional model served as the basis for all of the project's dashboards and analysis, allowing for actionable insights concerning employee performance, engagement, recruitment and training.

Data Preprocessing

In data science projects, preprocessing is a crucial step for us to guarantee that the data is clear, clean, and organized, which allows it to be in a good state for the model to be trained on, in other words, it transforms raw data into a format which can be suitable for the machine learning model that will be used. The overall performance is increased, noise is decreased, and the model correctness is improved all by prepping the data. In our project we will be dealing with 4 distinct datasets, each of which presents different types of information that could be relevant to the problem. We will be aiming to prep up our data for it to reach its best possible format and ready for them to be merged into one dataset.

i. Data Cleaning

a. Handling Missing Values

We have checked if there were any missing values across all 4 datasets, we only found in the first dataset, however, it was dealt with by dropping them in later steps because it was considered irrelevant.

Employee Dataset: There were two features that had missing values which are *ExitDate and TerminationDescription. ExitDate* had 1533 out of 3000 but that was normal because it means that not all employees had left the company, some are still employed.

b. Handling Inconsistent Categorical Values

Machine learning model accuracy may suffer from inconsistent categorical values. Some data entries variation such as additional spaces, irregular capitalization, or even a category may have various spellings which may introduce noise and error in the data. These problems can result in duplicated categories in an inconsistent way.

We examined all our datasets unique categories and found some inconsistencies.

Employee Dataset: We found out in one of the features named *Title*, because of the excess spaces, the same category was mentioned twice, but one with multiple spaces. It gives the impression of them being two distinct categories which were ‘Data Analyst’ and ‘Data Analyst ’. We addressed this issue by eliminating duplicate entries by removing spaces.

c. Feature Engineering

We decided on creating new features from existing ones that we thought might be beneficial in order to increase the prediction power of each dataset. It is important because it allows the model to extract more relevant patterns.

Employee Dataset:

- ***EmploymentDuration:*** We calculated the overall duration of employment in years by subtracting the start date from the exit data. If the exit date was missing, indicating the employee is still active, we used the current date. We added this feature because it gives information into how long employees stay in the company, which is relevant to our peoblem.
- ***Age:*** We calculated the age of each employee by subtracting their date of birth (DOB-datetime object) from the current date and converting the result into years.
- ***IsActive:*** We created our target class from existing features because there was no clear feature that could be considered as a target. This feature is a categorical feature, *IsActive*, that indicates if an employee is currently 'Active' which means the employee stays within the organization otherwise 'Not Active' meaning that the employee left the organization.

d. Dropping Irrelevant Features

There were columns in all 4 datasets that were irrelevant and had nothing to do with our business goal in which they don't add value to our data. We examined each dataset separately in order for us to find and remove these columns. There exists one column from each dataset that uniquely identifies them, but we won't be removing it at the moment because it will be used to merge on it later, which is the *EmpID* but it is names differently in each dataset.

Employee Dataset: From this dataset we decided to drop columns that are considered as unique identifiers and won't bring value into our model. Those features were dropped *FirstName, LastName, StartDate, ExitDate, Supervisor, ADEmail, LocationCode, TerminationDescription, DOB.* We had 26 columns before dropping and adding new features, and now we have 21 columns.

Engagement Dataset: In this dataset there was only one column which was irrelevant. We dropped the *Survey Date* since we won't make use of it later. We had 5 columns, and now we are left with 4.

Recruitment Dataset: We dropped the following columns *Application Date, First Name, Last Name, Date of Birth, Phone Number, Email, State, Address, Zip Code*. We had 18 columns, and now we have 9 columns.

Training Dataset: We dropped those features *Training Date, Trainer*. We had 9 columns, and now we have 7 columns.

ii. Dataset Merging

a. Merging Process

We decided to merge them after implementing some preprocessing techniques because if we combined them from the start, it would be harder for us to deal with all the features all combined in one dataset which will make it even harder not to miss or leave out important stuff. We just wanted to deal with each separately and then combine them after we have reduced their dimensionality, and it would be in a state where it is manageable.

Once the datasets cleaned and prepped, we used a common key to combine the datasets into a single dataset. We will merge them because it allows us to combine all relevant features into a one cohesive structure ready to be fed by the model, after performing another round of cleaning, As I mentioned previously, since the key identifier in all 4 datasets – **employee_df, engagement_df, recruitment_df, training_df** – had the same meaning but different names, so we standardized the key identifier across all datasets for them to have a unified column name and then merged them.

1. **Standardizing Key Column Names:**

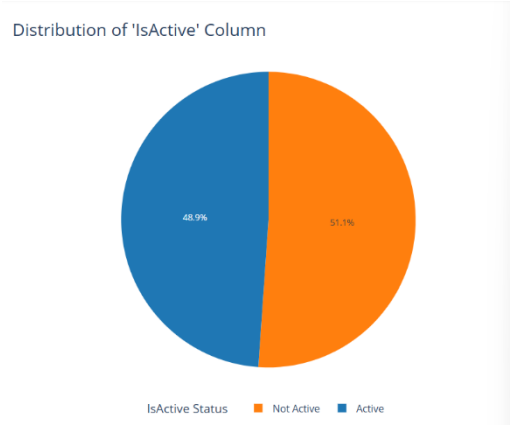
   We renamed *EmpID* in employee_df, *Employee ID* in engagement_df, *Applicant ID* in recruitment_df, *Employee ID* in training_df. All as *EmpID*.

2. **Merging the Datasets:**

   We performed an inner merge using *EmpID* as a common key between all 4 datasets. Now we have 3000 rows and 37 columns after the merge.

b. Post-Merge Cleaning

After merging, we performed additional cleaning steps such as rechecking if their exits missing values that might have been generated after the merge, duplicates, inconsistencies. We couldn't find any errorneous in the data, it was all well. We have also to check if there might be an imbalance in the target class *IsActive* but the class was well balanced, 48.9% of the data was labeled as 'Active' and 51.1% of the data was labeled as 'Not Active' which is not a big difference or considered an imbalance issue. Here is a visual illustration of our target feature distribution.



iii. Data Transformation

a. Encoding Categorical Variables

To prepare categorical input for machine learning models, those features should be encoded which means converting the information into numerical values. Since the majority of the models need numerical inputs, this procedure is crucial. This is how we went about encoding:

1. **Label Encoding Categorical Columns:** We first identified all object data type columns in the emp_df dataset after the merge. Label encoding works by assigning each category a distinct numerical value. It is used when features have few unique categories or limited number of unique values. Therfore, we decided to encode all object data type features with label encoding except for *Job Title, Country, City* and *Location* because those two columns had high cardinality.

2. **Target Encoding Categorical Columns:** When features have many unique values known as high cardinality, and if the features have a strong relationship with the target column. And if we used label encoding for high cardinality it can cause overfitting, therefore target encoding reduces the risk by generating meaningful numbers. Target encoding works by mapping unique categories to their mean target values. We first started by identifying features with high cardinality which cannot be label encoded. After that we plotted the correlation between all features and the target. We encoded *Job Title, Country, City* and *Location* using the target encoding technique.

3. **Binary Mapping** for the Target Feature (*IsActive*): We mapped Active class to 1 and Not Active class to 0.

b. Feature Scaling

To ensure that all features were on a comparable scale, we identified the features that have a scale of values that is way larger or smaller than the rest of the features scale. It is crucial because if we don't scale features, the machine learning sees features with high values, it will consider them much more relevant than the rest even though that might not be the case, therefore it will be biased. We started by getting some statistics for all features and plotted their distribution using either a bar plot or a box plot in order for us to see how the scale varies from one feature to the other.

After a thorough understanding of the features, we applied standard scaling to specific features for the following features: *Age, Desired Salary, Training Cost*. Standard scaling converts the features into a

standard range with a mean of 0 and standard deviation of 1 so that they become comparable with the rest of the features, this will ensure these features contribute fairly to the model.

iv. Feature Selection

We used feature selection to keep only the most important features for the model after merging and transforming the dataset. Some features may be irrelevant or don't help the model in predicting the target class much. It also reduces the risk of overfitting which might result in better generalization.
In order to find the most relevant features for our model, we tested out 3 distinct feature selection techniques:

1. **Recursive Feature Elimination (RFE):** RFE is a wrapper technique that recursively removes the least important features based on the model's performance. The process is repeated until all important features are left. It works by training the model and ranks the features by their importance to the model and removes the ones that are not Important. It is useful when working with a model like SVM, which is in our case.

2. **Tree-Based Model:** This technique works by examining how each feature divides the data into decision trees and which features contribute most to the decision tree. Tree-based techniques such as Random Forest and Gradient Boosting access and evaluate the significance of the features by providing the feature importance score.

3. **Regularizer L1 (Lasso):** Less significant features coefficients are penalized using this technique known as lasso. Zero-coefficient features are removed, however the non-zero will be kept.

Features Selected by RFE

	Feature	Rank
0	Title	1
1	BusinessUnit	2
2	EmployeeStatus	3
3	TerminationType	4
4	DepartmentType	5
5	Division	6
6	JobFunctionDescription	7
7	EmploymentDuration	8
8	Age	9
9	Satisfaction Score	10
10	Work-Life Balance Score	11
11	City	12
12	Country	13
13	Years of Experience	14
14	Desired Salary	15
15	Job Title	16
16	Status	17
17	Training Outcome	18
18	Location	19
19	Training Cost	20

After evaluating all those three distinct techniques used for feature selection, we decided to choose the **Recursive Feature Elimination** method because after testing each technique on the model, RFE had the best positive impact on the model's performance, it offered the best balance between interpretability and performance. Those were the selected features with their rank.

Model Selection and Performance Analysis

i. Model Description

In our employee attrition prediction and prevention system we aim to classify employees who are active and inactive (more likely to leave the company) based on their titles, business unit, status and many other key features. This classification case can help the HR team to target retention strategies and avoid having a high employee turnover rate. In order to get the most accurate and well-performing results, we implemented and tested 3 different machine learning classification techniques:

1. **Random Forest Classifier:** because it uses a bagging approach, it reduces the risk of overfitting. This ensemble-based method aggregates the predictions of multiple decision trees to have a better performance.

2. **Gradient Boosting Classifier:** this technique is also ensemble based that builds decision trees iteratively aiming to optimize the errors in the previous iteration.

3. **Support Vector Machine:** it uses a hyperplane in a high-dimensional space for a better separation of data points. It's a great approach to manage the imbalance of the data.

After testing these three classifiers, we found that the support vector classifier (SVM) has a better ability to have a strong generalization on our dataset. Due to its better performance compared to the other classifiers we chose SVM as our final machine learning model that will be integrated into our employee attrition and prevention system.

ii. State-of-the-Art Review

The have been various studies about our chosen model, Support Vector Machine (SVM), including employee attrition prediction. Some of these studies that highlighted the strength of SVCs in handling complex relationships within the data:

- **Ali Dabas (2024):** he emphasized the flexibility of SVM in binary classification tasks. In this study, *application of support vector machines in machine learning,* he highlighted the SVC's ability to handle nonlinear relationships and large datasets which perfectly align with our data we are using for the employee attrition prediction and prevention system.

- **IEEE (2020):** *A Systematic Literature Review on Support Vector Machines Applied to Classification.* This paper reviewed the extensive applications of SVM in a variety of domains, where it also highlighted the strength of SVC in capturing complex patterns and relationships using its kernel trick, ensuring accurate predictions.

- **Predicting Employee Attrition Using Machine Learning Approaches (2022):** in this paper the study is focused on our case, predicting employee attrition using various machine learning techniques. It also mentioned the significance of feature selection and preprocessing in general and how they are impactful on the results and performance. It discussed the use of SVC and its ability to benefit from the selected features, which align well with our preprocessing steps discussed earlier.

iii. How Does the Model Work?

After training and testing multiple models, our chosen model was **Support Vector Machine (SVM)** classifier, which delivered the best possible performance without bias or overfitting. SVM is a powerful supervised machine learning model, which is primarily used for classification tasks. It works by determining the best hyperplane in a higher dimensional space that separates the data points that belong to different classes. Its main goal or idea is to maximize the margin, or the distance between the support and the nearest points from each class which are known as support vectors. Both linear and non-linear can be handled by SVM. It uses a kernel method to convert data points into a higher dimension when data is not linearly separable.

We used a Radial Basis Function (RBF) kernel to handle nonlinear data correlations in the Support Vector Classifier (SVC) in order to train our model.

iv. How Is the Model Evaluated?

We evaluated our SVM model with several performance metrics obtained from the classification report and the confusion matrix which will be used to assess our model:

Classification Report:

Accuracy: It measures the proportion of the accurately predicted values to all occurrences, which indicates the overall correctness of the model. In our case, the SVM model performed very well overall, by properly classifying ***96.83%*** of the instances.

Precision: It focuses on the quality or how well the positive predictions are by determining the proportion of true positives (correctly predicted positives) out of all predicted positives whether it was true or false positives. In our case, the class 'InActive' has a precision of 0.99, which indicates that 99% of the cases that were predicted as 'InActive' were actually 'InActive'. However, the Active class precision was 0.94. indicating that 94% of the cases that we predicted as 'Active' were actually 'Active'. This metric is crucial when the cost of false positive is significant.

Recall: This metric measures the model's ability to identify all actual positive instances. It is computed by calculating the ratio of true positives to the actual positives (false negative and true positives). The class InActive recall is 0.94 which indicates that 94% of the real class instances were correctly classified by out model. And for the Active class the recall was 0.99.

F1-score: It is a fair assessment of our model performance which is provided by the harmonic mean of precision and recall. Which is particularly used when you need one metric to consider both false positives and negatives, especially if the data is imbalanced, which is not our case. Both classes gave a score of 0.97 indicating that the model maintains a good balance between precision and recall which also makes it reliable for identifying positive cases and reducing false positives.

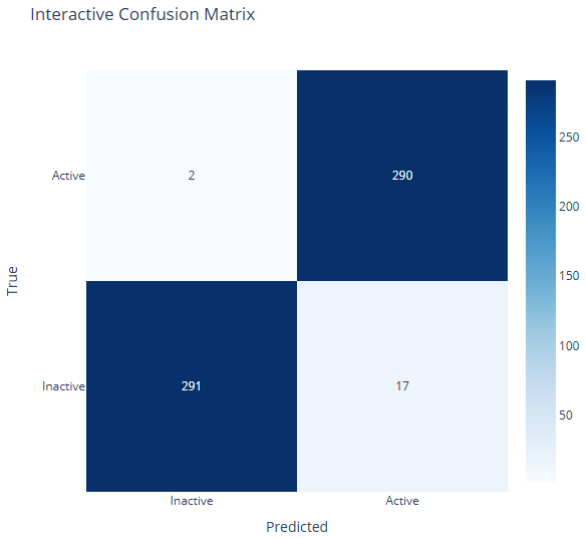
Confusion Matrix:

This confusion matrix is a table that compares predicted against actual labels to provide an overview of our model's performance:

	Predicted: Inactive	**Predicted: Active**
Actual: Active	False Negative (FN): 2	True Positive (TP): 290
Actual: InActive	True Negative (TN): 291	False Positive (FP): 17

Active class: 290 of the instances were predicted as *Active* and were truly *Active* which means they are correctly predicted as class *Active*. However, 2 of the instances were predicted as *InActive* and their true label is *Active*, meaning that they are incorrectly predicted to the other class *InActive*.

InActive class: 291 of the instances were predicted as *InActive* and were truly *InActive* which means they are correctly predicted as class *InActive*. However, 17 of the instances were predicted as *Active* and their true label is *InActive*, meaning that they are incorrectly predicted to the other class *Active*.



v. Advantages and Disadvantages of the Chosen Model

Advantages:

- **Effective for High-Dimensional Data:** SVM works effectively with high- dimensional data, meaning they can deal with data that has a large number of features. In our case, after doing feature selection, we came up with top 20 features, which is still considered a lot.
- **Robust to overfitting:** It works by maximizing the margin, which will in turn reduce the risk of the model overfitting the data.
- **Flexible Kernel Functions:** It can handle nonlinear data using various different kernel methods. We chose RBF as our kernel technique.

Disadvantages:

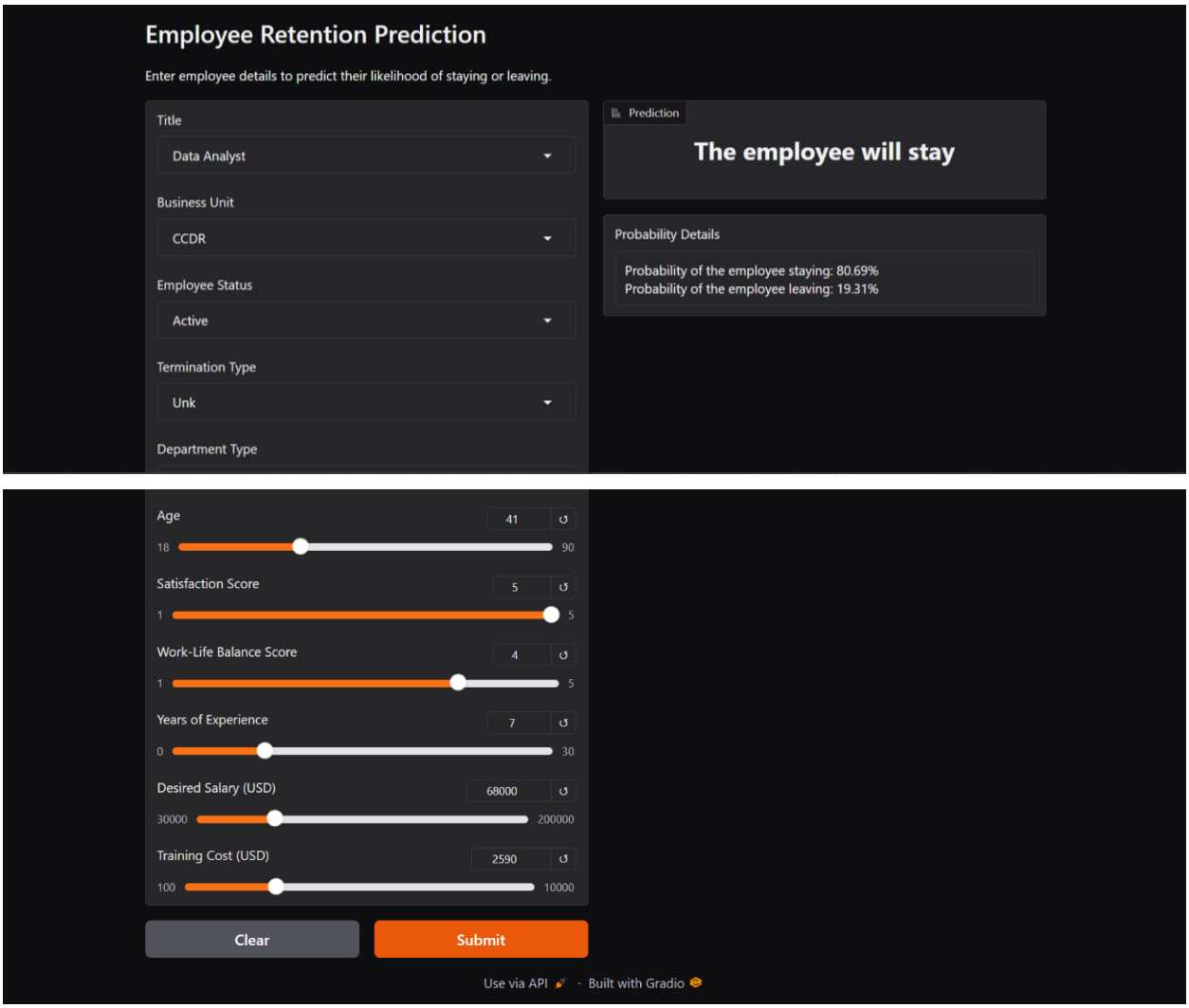
- **Computationally Intensive:** While training the model, it can be slow if we had a larger dataset.
- **Sensitive to Parameter Selection:** The models performance can vary based on our selection to the kernel used and hyperparameters such as gamma and C (regularization).
- **Difficult to interpret:** It is harder to visualize and understand intuitively the support vectors and hyperplane, unlike decision trees.

Regardless of its advantages and the challenges we might face, SVM was selected because it outperformed other models in terms of performance.

vi. User Interface for the SVM Model

In order to make the SVM more approachable and useful for the users, we not only have accessed in developing and evaluating the model, but we decided on making an interface for our model that could be reachable by any user which is also user friendly. All of that was achieved by using **Gradio** which is a python library that allows us to create and interactive user interface, as well as **Hugging Face** website that is an open-source platform. It is used for multiple purposes but in our case, we used it to deploy our interface. Hugging Interface integrates with gradio, enabling us to create an interactive demo for our model that can be accessed by anyone. Since the interface will be used in our HR management system, the interface will allow the HR to insert some of the required features (20 features which are the ones the SVM model was trained on) belonging to a specific employee (This employee could be newly registered one or applicants that would like to be employed or old but still employed employees etc…) and the system will predict wether the employee is most likely to leave or stay with a probability under the prediction label .

Here is an example of our **demo** in **HuggingFace**



Dashboards

Employee Attrition Dashboard:

This dashboard displays data on retention rates, employee attrition, and turnover-influencing variables. Statistics including employee retention, turnover, and breakdowns by corporate and demographic characteristics are highlighted. Key insights from the charts are:

- 51.1% of employees have stayed in the company, while 48.9% have left as we have observed from the pie chart, A high turnover percentage, which could suggest deeper issues with retention strategies, is demonstrated by an almost equal split.

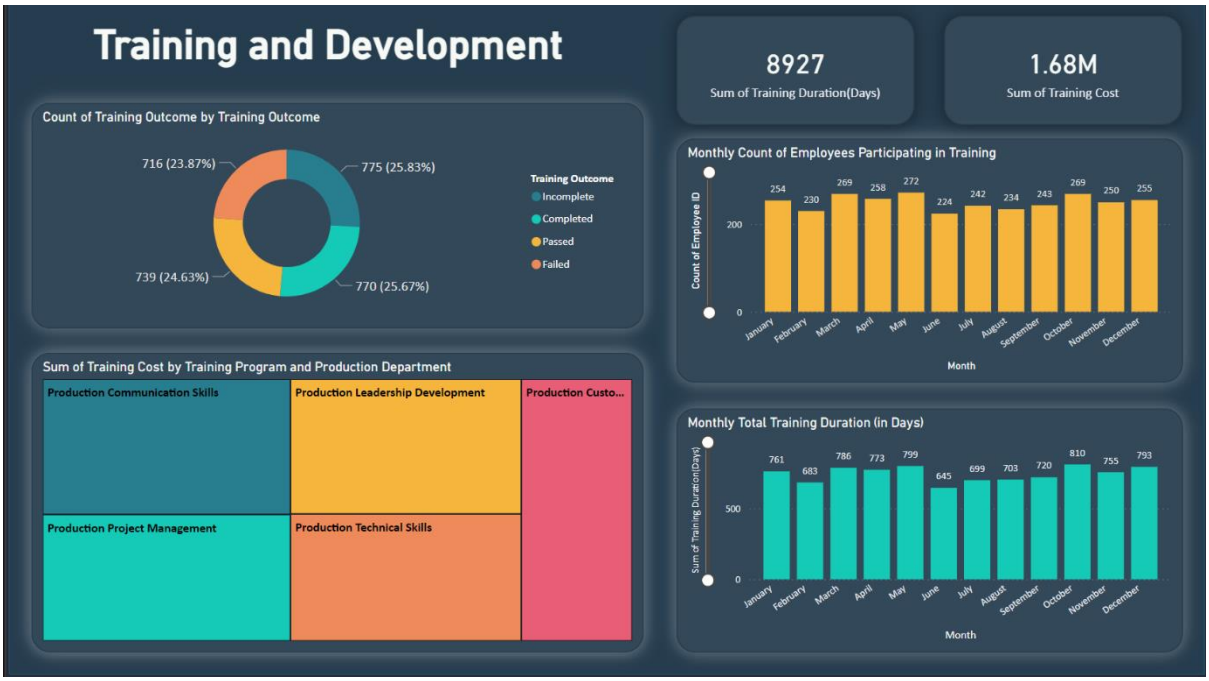
- The highest attrition rates have been observed among part-time employees, especially those who are widowed. This brings attention to a possible problem with part-time workers. Attrition may be decreased by implementing specific regulations to increase their level of involvement.

- The workforce is diverse, with the highest representation of personnel are from Hispanic and Black employees (20.6% and 20.97%, respectively).

- Younger employees (20–30 age range) show lower attrition rates compared to older employees (40–50). Also, in certain age ranges, female employees tend to have higher attrition rates. The retention of senior employees and the difficulties faced by female employees in higher attrition groups could be the primary objectives of customized engagement efforts and programs.

- Production Technician I and II roles show significantly higher attrition compared to other titles, which indicates that employees that work in these roles may need better work conditions or career development plans.

Training and Development Dashboard:



This dashboard analyzes training and development programs, which includes participation trends, training results, costs, and effectiveness of the programs. Some of the key insights are:

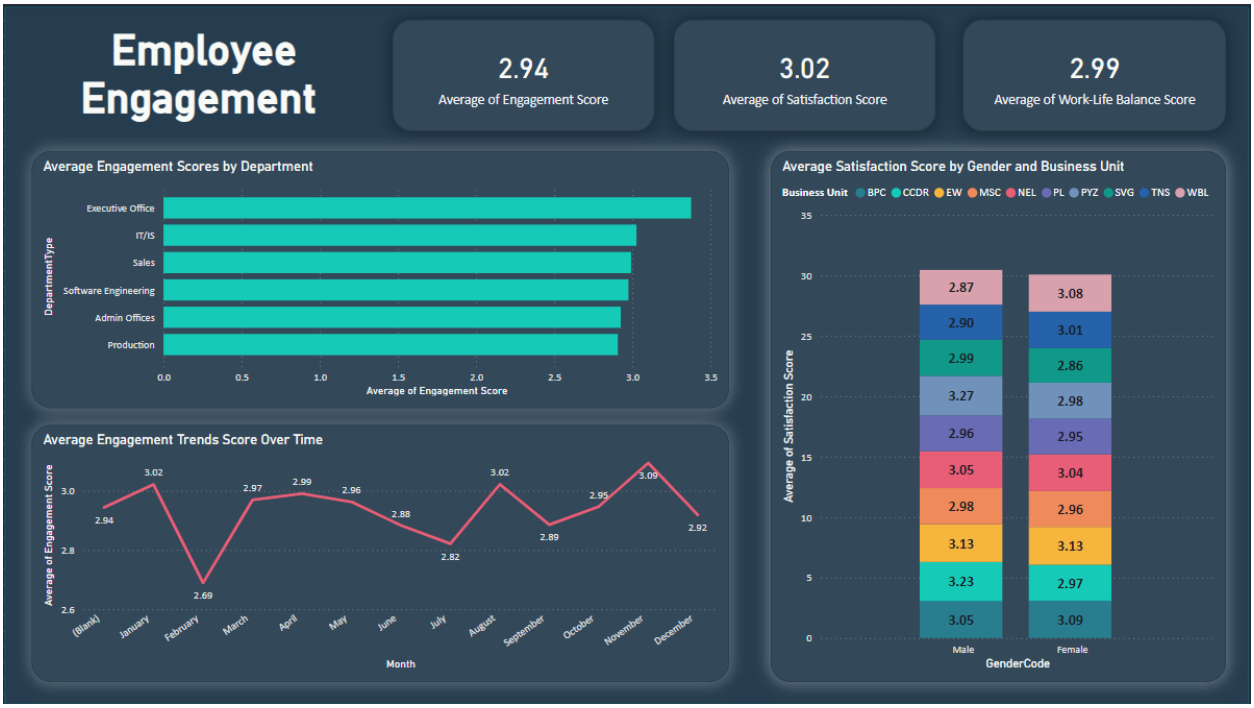
- Training completion rates are high, with 25.83% employees completing their training: however, 23.87% getting incomplete outcomes, illustrating opportunity for growth. The actionable insight would be that the reason why some trainings are incomplete, to improve the attendance.

- Some of the more costly programs are Production Leadership Development and Production Technical Skills. The organization must evaluate if high-cost programs are delivering a benefit to the company in terms of improved performance or retention.

- Training participation increases in May with 272 employees and drops in July with 224 employees. This suggests possible scheduling trends. Ensure that training is evenly dispersed throughout the year to avoid employee overload during periods of greatest demand.

These dashboards offer valuable information regarding the organization's workforce dynamics. The organization may boost employee retention and engagement by dealing with excessive turnover in certain areas, increasing training outcomes, and optimizing expenses.

Employee Engagement Dashboard:



In this dashboard we highlighted some key employee metrics to offer a detailed analysis of employee engagement such as: average of engagement scores which was 2.94, average satisfaction scores which was 3.02 and average work life balance score 2.99.

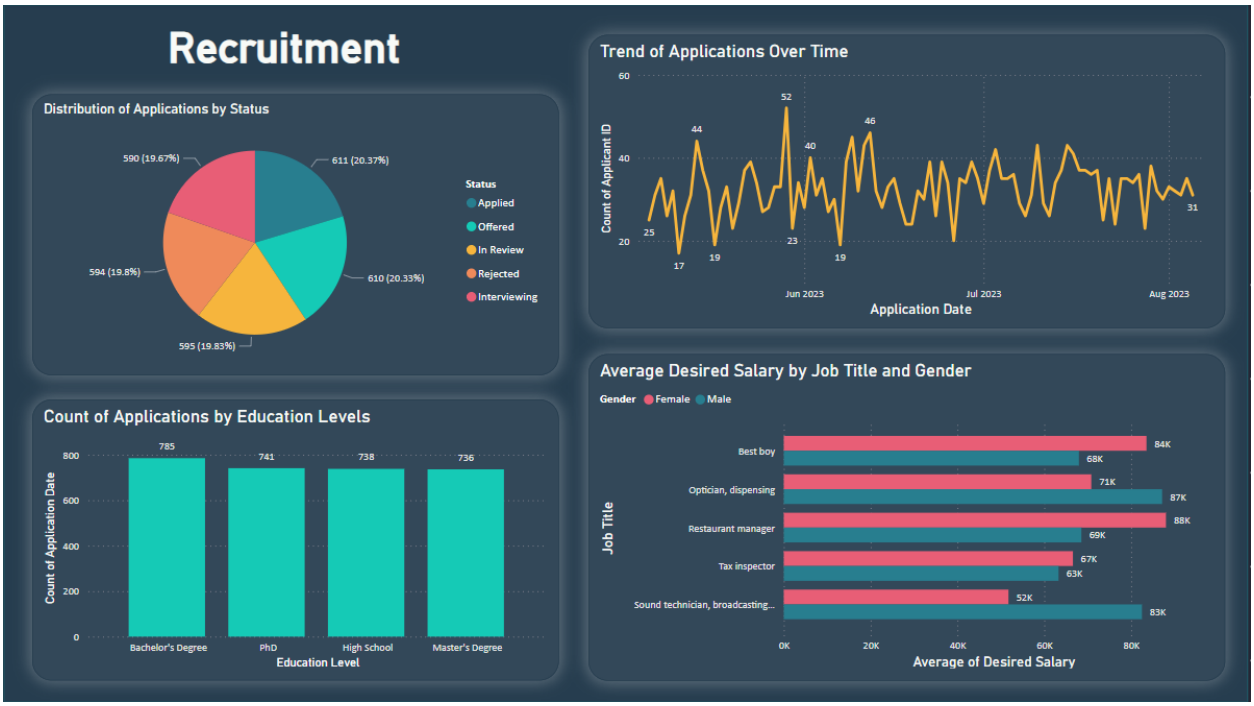
Some Key insights found from our analysis on employee engagement includes:

- **Engagement variations across departments:** we noticed that the executive office has the highest average employee engagement score which was 3.38, followed by the IT/IS department which had a score of 3.03, followed by Sales and Software Engineering departments with scores of 2.99 and 2.97 respectively. The department with the least average engagment score was the Production department and it has a score of 2.91.

- **Average Engagement Trends:** checking how the engagement changes over time was insightful since we noticed peaks in January (3.02) , in August (3.02) and in Novermber (3.09) which maybe because of events or other reasons. However, we also noticed a major drop in the engagement score in February where it had a score of 2.69 which is less than the overall average and it must be analysed further to know the reason behind that.

- **Gender and Business Unit Satisfaction Scores:** eventhough our data is balanced in terms of the gender, we noticed minor variations in some units between the male and female. For instance, in units like SVG, males had a higher satisfaction score (3.27) than females (2.98). However, in units like EW they had an equal satisfaction score which was 3.13.

Recruitment Dashboard:



The recruitment dashboard highlighted some key insights regarding the recruitment pipeline such as, trends of job applications, education levels and salary expectations from both genders. Some of the highlighted observations include:

- **Application by Status:** in our data, we noticed a balanced distribution of applications across 5 different status which include: applied, offered, in review, rejected and interviewing, each one of them was representing 19-20% of the total number of applications.

- **Application Trends:** the number of applications submitted to the system varied from May 2023 to August 2023, noticing some peaks such as 52 new applications submitted on 29th of May 2023 which was the highest among that period of time. The reason behind that maybe because of specific hiring campaigns or the way that the job was posted.
- **Applications of different education levels:** similar number of applications was seen despite the difference of the education level the applicant was holding. Applicants holding a bachelor's degree were representing the largest group (785 applications), followed by applicants holding a PhD, High School certification and Master's degree which represented the least number of applications (736).
- **Expected Salary by Gender and Job titles:** overall, male candidates reported a higher salary expectations compared to females, in roles such as Restaurant Manager and Tax Inspector. However, female applicants reported a higher expectation in roles like Best Boy.

Conclusion

In Conclusion, what makes our employee attrition prediction and prevention system unique from other HR management systems, is utilizing an efficient machine learning algorithm like Support Vector Classifier (SVC) making it easier for HR managers to classify the employees into “Active” and “In Active”. Identifying employees who are at risk of leaving the company is crucial for those managers since it can affect their reputation.

Moreover, the integration of our insightful interactive dashboards to the system, made it easier for the managers to make decisions that can improve the retention rate and minimize the employee turnover rate. For instance, when monitoring the overall engagement, it makes it easier to address the areas of concern and then taking the right actions based on that.

To sum up, our system combines all the tools needed to translate real time data into actionable insights that helps in a better decision-making process.

References

[1] *Employee/HR Dataset (All in One)*. (2023, August 13). Kaggle. https://www.kaggle.com/datasets/ravindrasinghrana/employeedataset/data

[2] *Employee Retention Prediction - a Hugging Face Space by Haniyahalzaben*. (n.d.). https://huggingface.co/spaces/Haniyahalzaben/Employee_Retention_Prediction

[3] Lucas, S. (2024, June 12). *Employee Attrition: Meaning, Impact & Attrition Rate Calculation*. AIHR. https://www.aihr.com/blog/employee-attrition/

[4] *1.4. Support Vector Machines*. (n.d.). Scikit-learn. https://scikit-learn.org/1.5/modules/svm.html

[5]https://d197for5662m48.cloudfront.net/documents/publicationstatus/216532/preprint_pdf/0712c8e4f08648a45d2de5c43f47e9e6.pdf

[6] *A Systematic Literature Review on Support Vector Machines Applied to Classification*. (2020, October 21). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9254028

[7] Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences*, *12*(13), 6424. https://doi.org/10.3390/app12136424