A black background with a black square

Description automatically generated with medium confidence

**EMPLOYE ATTRITION**

*Master’s degree in computer engineering for Robotics and Smart Industry*

Subject:

**Machine Learning & Artificial Intelligence**

Submitted to:

Vittorio Murino , Cigdem Beyan & Andrea Avogaro

ACADEMIC YEAR 2023/2024

VERONA UNIVERSITY, ITALY

Submitted by:

**Lokesh Kumar**

**VR511556**

Index

[Abstract 3](#_Toc169981657)

[1. Introduction 3](#_Toc169981658)

[2. Motivation and Rationale 3](#_Toc169981659)

[3. State of the Art 4](#_Toc169981660)

[4. Objectives 4](#_Toc169981661)

[5. Methodology 4](#_Toc169981662)

[5.1 Data Exploration and Preprocessing 4](#_Toc169981663)

[5.1.1Exploratory Data Analysis 5](#_Toc169981664)

[5.2 Feature Selection 7](#_Toc169981665)

[5.3 Machine Learning Algorithms 7](#_Toc169981666)

[5.3.1 K-Nearest Neighbors (KNN): 7](#_Toc169981667)

[5.3.2 Support Vector Machine (SVM): 7](#_Toc169981668)

[5.3.3 Decision Tree: 7](#_Toc169981669)

[5.3.4 Mean Shift Clustering: 7](#_Toc169981670)

[5.3.4 Neural Network: 7](#_Toc169981671)

[6. Experiments and Results 7](#_Toc169981672)

[6.1 KNN Algorithm 8](#_Toc169981673)

[6.2 Support Vector Machine 8](#_Toc169981674)

[6.3 Decision Tree 9](#_Toc169981675)

[6.4 Mean Shift Clustering 9](#_Toc169981676)

[6.5 Neural Network 10](#_Toc169981677)

[6.6 Model Comparison 10](#_Toc169981678)

[7. Conclusions 11](#_Toc169981679)

[References 11](#_Toc169981680)

# Abstract

The process by which employees depart an organization known as employee attrition presents serious difficulties for human resource managers. Organizations can lower turnover costs and sustain productivity by using proactive retention tactics with the use of accurate employee prediction. This study investigates the use of several machine learning algorithms to forecast staff churn. The study illustrates how well several models predict attrition using a large dataset that include personal, professional, and salary variables.

The methodology encompasses five machine learning algorithms: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Mean Shift Clustering, and Neural Network. It also covers feature selection and data preprocessing. Precision, accuracy, recall, and F1 scores are used to analyze the results and offer information on the best methods for predicting degradation.

# 1. Introduction

Employee turnover affects overall corporate performance, employee morale, and operational continuity, it is an important issue for firms. Excessive turnover rates can result in significant expenses for hiring, training, and knowledge loss within the firm. HR departments can identify at risk employees and carry out focused interventions to retain talent by using employee attrition forecasts. This study analyzes a dataset of employee qualities and predicts the chance of attrition using machine learning techniques.

A room with many computers and a few people walking

Description automatically generated

# 2. Motivation and Rationale

The increasing requirement for businesses to control labor expenditures and personnel efficiently served as our driving factor for this project. Layoffs of employees may cause major expenses for recruiting and onboarding new staff members as well as disturbances to corporate operations. Organizations may enhance employee retention, job happiness, and overall productivity by precisely identifying the people who are most likely to depart.

The use of predictive models in HRM is one of the main forces behind this. HR managers may identify employees who are at risk and take proactive measures to address their issues with the use of automated solutions. This can lower attrition rates and raise employee engagement.

Another motivation is the practical importance of accurate uncertainty forecasting in improving organizational decision-making. With advances in machine learning and data availability, developing robust models to predict employee absenteeism has become increasingly feasible and valuable.

# 3. State of the Art

Numerous machine learning models and methods for preparing data have been employed in earlier research on employee leave predictions. Research indicates that ensemble and decision tree techniques are frequently applied in this context. These studies have pinpointed important variables that affect attrition, including pay, work-life balance, and job satisfaction. Nonetheless, there is a need to incorporate cutting-edge methods like deep learning and continuously increase the accuracy of the models.

Promising outcomes in a range of categorization tasks have been observed in recent developments in deep learning, namely in neural networks. It has also been demonstrated that transfer learning, which makes use of models that have already been trained on sizable data sets, is useful for enhancing model performance when there is a shortage of training data.

# 4. Objectives

Our objectives for this project are as follows:

Accurate Prediction create algorithms that use both conventional and deep learning techniques to reliably predict an employee's likelihood of leaving the company.

Explore transfer learning by implementing transfer learning techniques with neural network by fine-tuning pre-trained models to predict employee turnover.

Comparing Performance Based on measures like accuracy, precision, recall, and F1 score, compare the classification performance of cutting-edge neural network models with those of conventional machine learning techniques (SVM, KNN).

By fulfilling these goals, our analysis hopes to advance knowledge about how well deep learning and conventional methods predict employee turnover.

# 5. Methodology

## 5.1 Data Exploration and Preprocessing

The study's data set comprises comprehensive personnel data, such as job roles, performance indicators, demographics, and remuneration information. In preliminary data exploration, the quality and completeness of the data are evaluated. Categorical variables were label encoded and columns with more than 20% missing values were eliminated. After that, the data set was standardized to make sure that every feature was on the same scale, which improved the machine learning algorithm's effectiveness.

### 5.1.1 Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to understand the distribution of key attributes and their potential impact on attrition. The analysis revealed that the dataset was imbalanced, with a larger proportion of employees who did not leave the organization. Histograms and bar plots were used to visualize the distribution of attributes such as age, monthly income, and job satisfaction. These visualizations helped in identifying trends and patterns that could influence the prediction models.

A graph with a green rectangular bar

Description automatically generated

A graph of a distribution of age

Description automatically generatedA graph of a distribution of income

Description automatically generated

A graph of a city

Description automatically generated with medium confidence

## 5.2 Feature Selection

Two methods dimensionality reduction and feature significance analysis were used to choose features. The most significant variance in the data was captured by reducing the number of attributes from 35 to 3, using dimensionality reduction techniques such as Principal Component Analysis (PCA). Furthermore, to choose the most pertinent features for the prediction models, feature importance scores were computed. To get rid of low-variance features, attributes having a standard deviation less than 0.8 were disregarded.

## 5.3 Machine Learning Algorithms

Five machine learning algorithms were implemented to predict employee attrition:

### 5.3.1 K-Nearest Neighbors (KNN):

A simple, instance-based learning algorithm that classifies data points based on their proximity to other labeled points.

### 5.3.2 Support Vector Machine (SVM):

A powerful classifier that finds the optimal hyperplane to separate different classes in the feature space.

### 5.3.3 Decision Tree:

A tree-based model that splits the data into branches to make predictions based on the values of input features.

### 5.3.4 Mean Shift Clustering:

A clustering algorithm that groups data points based on density, used here to explore unsupervised learning techniques.

### 5.3.4 Neural Network:

A multi-layered model that mimics the human brain's structure, capable of capturing complex relationships in the data.

# 6. Experiments and Results

The performance of each model was evaluated using metrics such as accuracy, precision, recall, and F1-score. The results are summarized in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm name** | **Accuracy** | **Precision** | **Recall** | **F1\_Score** |
| KNN | 61.62% | 61.24% | 63.88% | 62.53% |
| SVM | 62.70% | 62.47% | 64.15% | 63.30% |
| Decision Tree | 58.11% | 58.64% | 55.80% | 57.18% |
| Mean Shift Clustering | 49.86% | 100% | - | - |
| Neural Network | 64.59% | 65.98% | 60.65% | 63.20% |

## 6.1 KNN Algorithm

The KNN algorithm exhibited moderate performance with balanced precision and recall, indicating a fair ability to classify both attrition and non-attrition cases. Its simplicity makes it easy to implement but may not capture complex patterns in the data.

A blue squares with white text

Description automatically generated

## 6.2 Support Vector Machine

SVM performed slightly better than KNN, with a higher recall, suggesting it is more effective in identifying employees who are likely to leave. The model's robustness in handling high-dimensional data contributed to its performance.

A blue squares with white text

Description automatically generated

## 6.3 Decision Tree

The Decision Tree algorithm had lower performance compared to KNN and SVM. While it provided interpretable results, its tendency to overfit the data may have reduced its generalization capability.

A blue and white chart

Description automatically generated

## 6.4 Mean Shift Clustering

Mean Shift Clustering was less effective in this context, reflecting its unsuitability for classification tasks where labeled data is crucial. The model's focus on clustering data points based on density did not translate well to the attrition prediction problem.

A blue squares with white text

Description automatically generated

## 6.5 Neural Network

The Neural Network achieved the highest accuracy and precision, indicating its robustness in handling the complexity of the dataset. Its ability to model non-linear relationships and interactions between features contributed to its superior performance.

A blue squares with white text

Description automatically generated

## 6.6 Model Comparison

A graph showing a comparison of a model

Description automatically generated with medium confidence

# 7. Conclusions

This study demonstrates the applicability of machine learning techniques in predicting employee attrition. The results indicate that while simpler models like KNN and Decision Tree provide moderate performance, more sophisticated models like SVM and Neural Network offer better predictive capabilities. Future work will focus on refining these models through hyperparameter tuning, exploring ensemble methods, and integrating additional features such as employee engagement metrics. Implementing dynamic data integration and developing interactive dashboards will enhance the practical application of these predictive models.

# References

Kaggle Employee Attrition Dataset Available at:

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset