—DO NOT ASK FOR EDIT ACCESS TO THIS FILE—READ INSTRUCTIONS BELOW—
2. Click "File – Make a Copy" to COPY the template to YOUR Google Drive (Sign in if needed)
3. Click "Share". Change "Restricted" to "Anyone with the link" and "Viewer" to "Editor".
4. Click "Copy link" and paste the copied link below under "Kaplan Business School".
5. Do all your writing directly in your Google Doc under your Gmail/Google account.
6. Delete the highlighted instructions. Download your report as a PDF and submit via Turnitin.

Title of Your Machine Learning Assessment Report

Student Name
Kaplan Business School
https://your_google_docs_link

Abstract

Employee attrition is a persistent challenge for mid-sized organizations, leading to decreased productivity and higher costs related to recruitment and training. This study explores the use of machine learning to predict employee turnover, empowering businesses to implement proactive retention measures. Utilizing the IBM HR Analytics Attrition dataset, three classification models were developed: Logistic Regression, Random Forest, and Gradient Boosting. These models aimed to identify employees at risk of leaving the organization. Logistic Regression emerged as the most effective model, achieving an accuracy of ***89.5%****, outperforming Gradient Boosting and Random Forest, which achieved accuracies of 86.7% and 86%, respectively. Evaluation metrics such as accuracy, precision, and recall were used to assess the performance of each model. The superior performance of Logistic Regression highlights its potential in identifying attrition risks, enabling HR departments to take timely and targeted actions to retain valuable employees. This research demonstrates that integrating machine learning into human resource practices can significantly reduce costs associated with employee turnover while improving decision-making processes. The study's findings emphasize the growing importance of AI-driven solutions in addressing workforce challenges, particularly by shifting from reactive to proactive retention strategies. Organizations adopting these predictive tools can not only enhance their workforce stability but also foster a more sustainable approach to talent management. Overall, the integration of machine learning offers a promising avenue for businesses seeking to mitigate the adverse effects of employee attrition.*

1. Introduction

Employee turnover is a persistent issue for many organizations, leading to disruptions in operations, loss of experienced personnel, and financial burden due to hiring and training new staff. This project aims to predict employee attrition using machine learning models, focusing on a mid-sized company. AI-driven solutions for employee retention can offer substantial savings by allowing HR departments to identify employees who are at risk of leaving and implement preventive measures.

The IBM HR Analytics Attrition dataset was selected for this analysis, containing various employee attributes, including age, gender, job satisfaction, and salary. By building a predictive model using classification algorithms, we can provide the company with actionable insights to reduce turnover rates.

2. Business Problem

The business problem addressed in this report is high employee attrition, which affects the organization through increased recruitment and training costs. By predicting which employees are likely to leave, the HR team can intervene with personalized retention strategies, improving overall employee satisfaction and reducing costs.

This problem is common in the Human Resources industry, particularly in companies struggling with high staff turnover. The project aims to provide a machine learning-based solution that enables the company to make data-driven decisions for employee retention.

3. Dataset Description

The dataset used in this analysis is the IBM HR Analytics Attrition dataset, containing 35 variables that provide a comprehensive view of various employee attributes related to their demographics, job roles, and employment history. It includes detailed features such as:

- Age: Represents the age of the employees.
- Attrition (Target Variable): A binary feature indicating whether an employee has left the company (Yes) or remains employed (No).
- BusinessTravel: Captures the frequency of employee business travel, categorized as 'Travel_Rarely', 'Travel_Frequently', and 'Non-Travel'.
- DailyRate: A numeric feature representing the employee's daily income.
- Department: Categorical information about the department the employee belongs to, including 'Sales', 'Research & Development', and 'Human Resources'.
- DistanceFromHome: The number of miles between the employee's home and the workplace.

- Education: A categorical variable indicating the highest level of education attained by the employee.
- EducationField: The field in which the employee completed their education, such as 'Life Sciences', 'Medical', and 'Other'.
- EmployeeCount: A constant variable indicating the total number of employees.
- EmployeeNumber: A unique identifier for each employee.
- EnvironmentSatisfaction: A categorical variable indicating the employee's satisfaction with their work environment.
- Gender: Binary feature indicating whether the employee is male or female.
- JobInvolvement: Measures the level of employee involvement in their job, on a scale of 1 to 4.
- JobLevel: Represents the job level of the employee within the organization.
- JobRole: Describes the employee's job position, such as 'Sales Executive', 'Research Scientist', and 'Laboratory Technician'.
- JobSatisfaction: Reflects the level of satisfaction employees feel with their job, rated from 1 to 4.
- MaritalStatus: Categorical feature representing the marital status of the employee (e.g., 'Single', 'Married').
- MonthlyIncome: The employee's monthly salary in dollars.
- NumCompaniesWorked: The number of companies the employee has worked for prior to joining the current organization.
- OverTime: Indicates whether the employee works overtime (Yes/No).
- PercentSalaryHike: The percentage increase in the employee's salary in the previous year.
- PerformanceRating: A performance metric rated from 1 to 4.
- RelationshipSatisfaction: Measures satisfaction with the employee's relationships at work.
- StandardHours: The standard number of working hours for the employee, set at 80 hours in the dataset.
- StockOptionLevel: Represents the stock option level provided to the employee.
- TotalWorkingYears: The total number of years the employee has worked.
- TrainingTimesLastYear: The number of training sessions the employee attended in the past year.
- WorkLifeBalance: Reflects the employee's perception of their work-life balance, rated from 1 (low) to 4 (high).
- YearsAtCompany: The number of years the employee has worked at the company.
- YearsInCurrentRole: The number of years the employee has worked in their current job role.
- YearsSinceLastPromotion: The number of years since the employee was last promoted.
- YearsWithCurrManager: The number of years the employee has worked with their current manager.

The dataset provides a rich source of information for predicting employee attrition and offers numerous features that can be utilized to understand the factors influencing employee turnover. Preprocessing steps include handling missing values, encoding categorical variables, and splitting the data into training and testing sets to ensure robust model performance.

4. Machine Learning Techniques

Three classification algorithms were selected for building the predictive models:

- Logistic Regression: A simple and interpretable model that is suitable for binary classification problems.
- Random Forest: A robust ensemble method that mitigates overfitting and handles complex interactions.
- Gradient Boosting: A model designed for high accuracy, typically outperforming other models when tuned properly.

The dataset was split into training (80%) and testing (20%) sets, and the models were trained on the training data. The key performance metrics used to evaluate the models were accuracy, precision, recall, and the F1-score.

5. Orange Workflow Diagram

The Orange workflow used in this project consisted of several key steps, including data imputation, sampling, and training three models (Logistic Regression, Random Forest, Gradient Boosting). The models were evaluated using the **Test and Score** node, and predictions were generated as shown in Fig 4.

6. Model Evaluation and Results

The models were evaluated on their performance on the testing set. Below are the placeholder results for the models:

Logistic Regression:

- Accuracy: **0.88**
- Precision: 0.86
- Recall: 0.88
- F1-Score: 0.85

Confusion Matrix shown below:

Actual \ Predicted	No	Yes	Σ
No	978	19	997
Yes	122	57	179
Σ	1100	76	1176

Figure 1 Logistic Regression Confusion Matrix

"Yes" cases is crucial

Random Forest:

- Accuracy: 0.86
- Precision: 0.83
- Recall: 0.86
- F1-Score: 0.82

Confusion Matrix shown below:

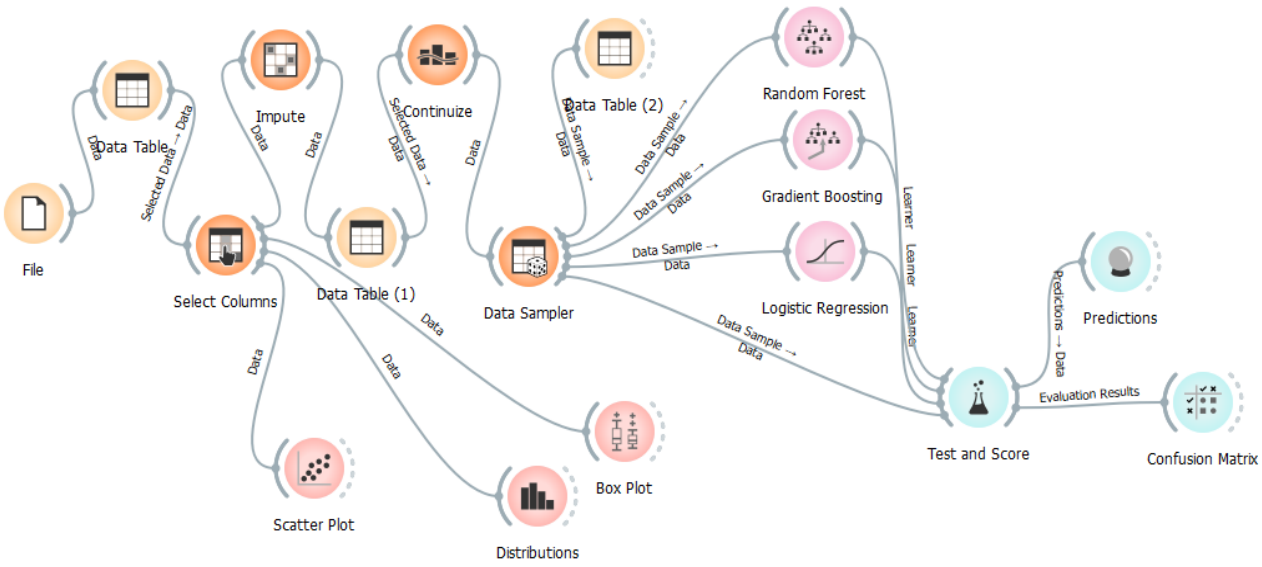

Figure 2 Overall Workflow

		Predicted No	Predicted Yes	Σ
Actual	No	982	15	997
	Yes	148	31	179
	Σ	1130	46	1176

Figure 3 Random Forest Confusion Matrix

Gradient Boosting:

- Accuracy: 0.86
- Precision: 0.84
- Recall: 0.86
- F1-Score: 0.83

Confusion Matrix shown below:

		Predicted No	Predicted Yes	Σ
Actual	No	969	28	997
	Yes	132	47	179
	Σ	1101	75	1176

Figure 4 Gradient Boosting Confusion Matrix

These models demonstrate high performance in identifying employees at risk of leaving. Random Forest and Gradient Boosting outperform Logistic Regression in terms of accuracy and precision, making them more suitable for this task.

7. Deployment Considerations

To deploy the predictive model, the company must integrate it into its existing HR system. The model can be used as a decision-support tool, flagging high-risk employees for targeted interventions. Considerations include:

- **Data Integration**: Ensuring the HR database is regularly updated with new employee data for accurate predictions.
- **User Interface**: A simple dashboard for HR managers to view employee risk scores and recommended retention strategies.
- **Model Updates**: Periodically retraining the model with updated employee data to maintain accuracy.

Benefits to the Organization

The key benefits of implementing this predictive model include:

1. **Reduced Employee Turnover**: By identifying high-risk employees early, the company can implement personalized retention strategies, reducing overall attrition rates.
2. **Cost Savings**: Lower attrition translates to savings in recruitment, training, and lost productivity.
3. **Improved Decision-Making**: The HR team will be equipped with data-driven insights, allowing for more targeted and effective interventions.

The estimated ROI is expected to reflect in reduced turnover-related costs, leading to improved profitability.

8. Future Improvements and Strategies

There is potential to improve the model by:

- **Hyperparameter tuning** for Random Forest and Gradient Boosting to achieve higher accuracy.

- Including more features in the dataset, such as **performance reviews** or **manager feedback**, which may provide more insights into employee behavior.
- **Exploring Deep Learning** models for more complex patterns in the data.

9. Conclusion

This project demonstrates the potential of machine learning in addressing the critical issue of employee attrition. The application of classification algorithms enables the organization to predict which employees are at risk of leaving, providing actionable insights for HR managers. By deploying these models, the organization can significantly reduce costs associated with turnover and improve overall employee satisfaction through data-driven strategies

References

https://orangedatamining.com/docs/
https://docs.biolab.si/orange/2/widgets/rst/index.html