This is the **published version** of the bachelor thesis:

Oriol Cabrera, Lara. *Predictive Machine Learning in Human Resources : enhancing employee retention strategies through data-driven insights*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2025 (Administració i Direcció d'Empreses)

This version is available at https://ddd.uab.cat/record/323874

IMPROVING EMPLOYEE RETENTION WITH MACHINE LEARNING

Lara Oriol Cabrera
Business Administration & Management Student
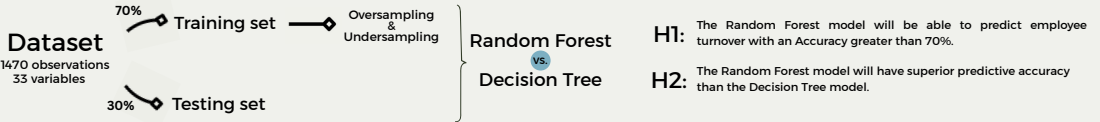Year 2024-2025

The Problem

Employee retention has long been a key topic due to the significant direct and indirect costs it brings to organizations. However, **traditional retention strategies are often reactive**, addressing turnover only after employees decide to leave. In addition, these strategies **usually focus on organizational or team levels**, missing early signs from individual employees. Consequently, many opportunities to proactively retain talent are lost, reducing their overall effectiveness.

The Solution

Using advanced **Machine Learning techniques can help firms become more proactive in employee retention**. By analyzing large amounts of data, the models can identify patterns and predict which employees are at risk of leaving earlier than traditional methods. This allows companies to tailor retention strategies to individual needs, making them more effective and timely.
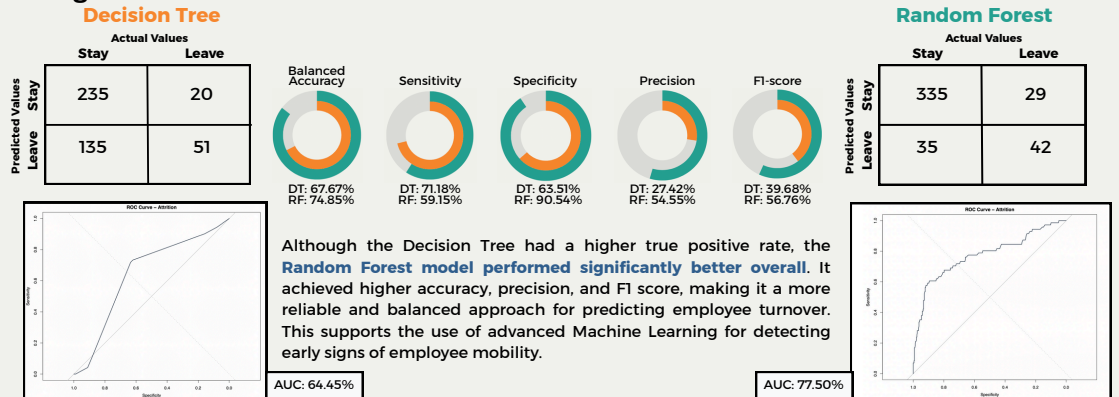
Methodology

This study aims to show that advanced Machine Learning can successfully predict employee turnover, by using a public synthetic dataset with employee data. Two models were compared: a Random Forest, representing the more advanced approach, and a Decision Tree, which is simpler. The goal is to compare them to show that advanced methods predict turnover more efficiently.
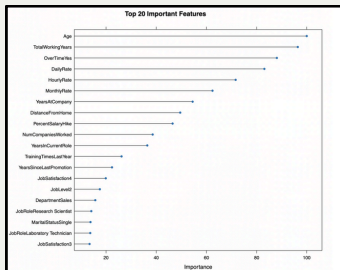
Dataset
1470 observations
33 variables

70% → Training set → Oversampling & Undersampling
30% → Testing set

Random Forest vs. Decision Tree

H1: The Random Forest model will be able to predict employee turnover with an Accuracy greater than 70%.

H2: The Random Forest model will have superior predictive accuracy than the Decision Tree model.

Findings & Outcomes

Decision Tree

	Actual Values	
Predicted Values	**Stay**	**Leave**
Stay	235	20
Leave	135	51

	Balanced Accuracy	Sensitivity	Specificity	Precision	F1-score
DT:	67.67%	71.18%	63.51%	27.42%	39.68%
RF:	74.85%	59.15%	90.54%	54.55%	56.76%

ROC Curve – Attrition
AUC: 64.45%

Random Forest

	Actual Values	
Predicted Values	**Stay**	**Leave**
Stay	335	29
Leave	35	42

ROC Curve – Attrition
AUC: 77.50%

Although the Decision Tree had a higher true positive rate, the **Random Forest model performed significantly better overall**. It achieved higher accuracy, precision, and F1 score, making it a more reliable and balanced approach for predicting employee turnover. This supports the use of advanced Machine Learning for detecting early signs of employee mobility.

Random Forest Interpretation

Global

Top 20 Important Features
(Age, TotalWorkingYears, OverTimeYes, DailyRate, HourlyRate, MonthlyRate, YearsAtCompany, DistanceFromHome, PercentSalaryHike, NumCompaniesWorked, YearsInCurrentRole, TrainingTimesLastYear, YearsSinceLastPromotion, JobSatisfaction, JobLevel2, DepartmentSales, JobRoleResearch Scientist, MaritalStatusSingle, JobRoleLaboratory Technician, JobSatisfaction3)
Importancia

To support global interpretability, Feature Importance reveals which variables contribute most to the model's predictions. **The chart displays the top 20 most important features from the dataset for the model.**

Local

For local interpretability, **SHapley Additive exPlanations was used to understand how individual features influence an employee's predicted risk of leaving**. SHAP values show which characteristics increase or decrease the probability of turnover. Based on these values, **an interactive platform was developed** where users can input employee data to get a predicted risk and the top 5 personalized reasons for leaving.

Check out the interactive dashboard!

Conclusions

This study demonstrates that advanced Machine Learning models, such as Random Forest, can predict employee turnover accurately and efficiently. By combining global and local interpretability techniques, firms gain valuable insights into the key drivers of turnover. These tools showcase how advanced learning models can serve as an ally for more proactive, personalized, and data-driven retention strategies.