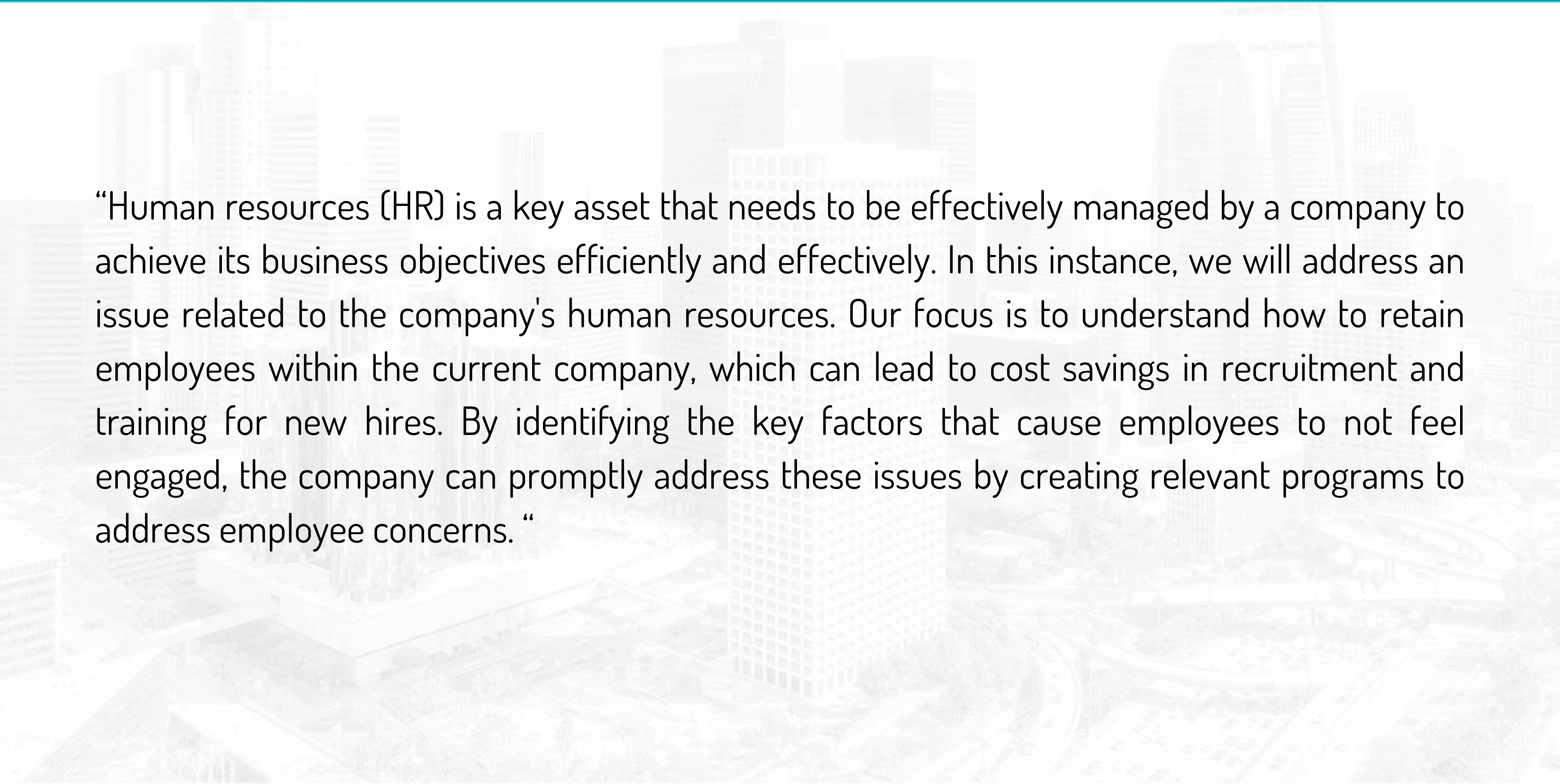
Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Created by:
Anik Maulia Tri Handayani
anikmaulia263@gmail.com
linkedin.com/in/anikmauliatrihandayani

I possess foundational knowledge in data analysis and a proven ability to design and develop Machine Learning solutions. I have applied these skills to a project aimed at improving employee retention by predicting employee attrition. I strongly believe in the pivotal role that data plays in enhancing business performance. I hold relevant certifications and practical experience, which have prepared me to be a dedicated data professional. I am also adept at collaborating within multidisciplinary teams and am ready to advance my career in data analysis across various industries, including telecommunications, commerce, retail, FMCG, finance, and banking.

Overview



“Human resources (HR) is a key asset that needs to be effectively managed by a company to achieve its business objectives efficiently and effectively. In this instance, we will address an issue related to the company's human resources. Our focus is to understand how to retain employees within the current company, which can lead to cost savings in recruitment and training for new hires. By identifying the key factors that cause employees to not feel engaged, the company can promptly address these issues by creating relevant programs to address employee concerns. “

Data Preprocessing

Data Splitting

Data splitting occurs prior to other preprocessing steps to prevent information leakage between the training and testing datasets. An 80:20 ratio is applied to allocate data for training and testing due to data volume constraints, optimizing its utilization in model construction.

Missing Values

- Missing values in the **AlasanResign** column will be replacing with values from data entries that share the same **TanggalResign** value.
- Missing values in the **JumlahKeikutsertaanProjek, JumlahKeterlambatanSebulanTerakhir, SkorKepuasanPegawai,** and **JumlahKetidakhadiran** columns will be imputed using the median value.
- The **IkutProgramLOP** column will be dropped since the number of missing values exceeds 89%, and imputation might introduce bias into the data.

Duplicated Values

The datasets contain no duplicates, so there is no need to handle duplicated values.

For more details, you can refer to the Jupyter notebook here

Data Preprocessing

Rakamin Academy

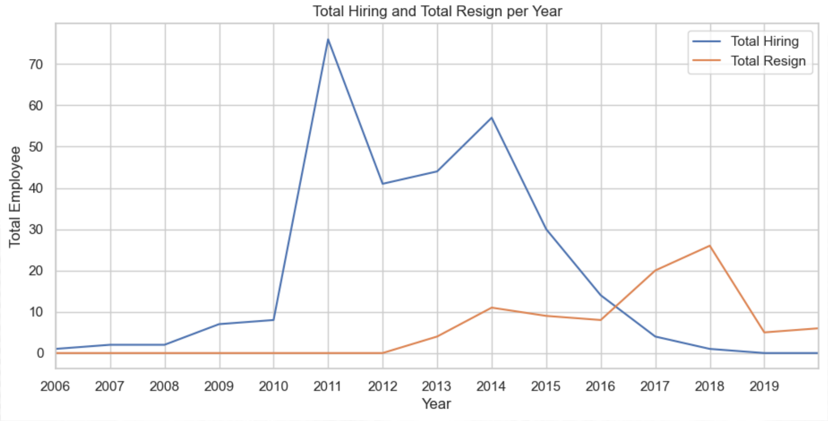
Invalid Values

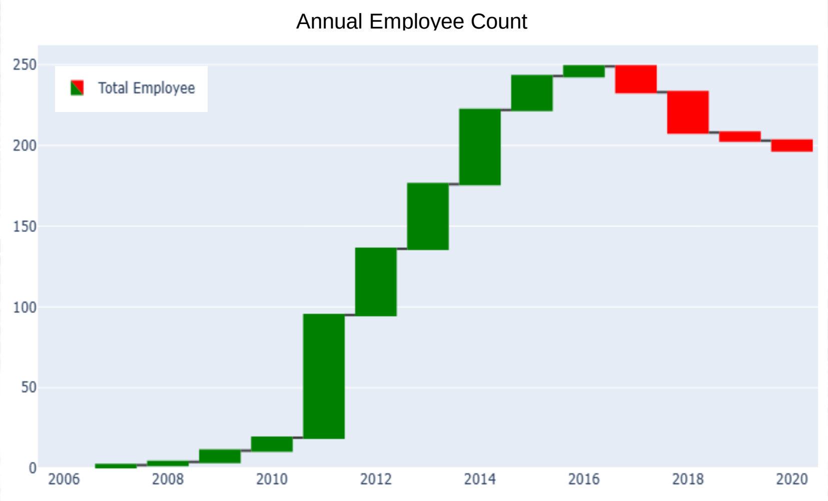
- Swapping invalid values in the **TanggalHiring** and **TanggalResign** columns for employees with EnterpriseIDs 106480 and 111209, as their resignation dates are earlier than their hiring dates.
- Changing invalid values in the **StatusPernikahan** column for entries with Lainnya and - to Belum_menikah, assuming that these employees do not plan to get married.
- Updating the **AlasanResign** for Product Design (UI & UX) to ganti_karir for roles other than Product Design (UI & UX) and kejelasan_karir for Product Design (UI & UX) roles. This is based on the assumption that Product Design (UI & UX) employees may lack career clarity and other employees want to switch careers to Product Design (UI & UX).
- Converting data types in the **JumlahKeikutsertaanProjek**, **JumlahKeterlambatanSebulanTerakhir**, **SkorKepuasanPegawai**, and **JumlahKetidakhadiran** columns to integers, as the values are in units.
- Changing data types in the **TanggalLahir**, **TanggalHiring**, **TanggalPenilaianKaryawan**, and **TanggalResign** columns to datetime.
- Dropping the **PernahBekerja** column as it has a constant value of 1, indicating it's not an important feature for distinguishing between employees who resign and those who do not.

Outliers

Handling outliers will be performed after the feature extraction process to anticipate the presence of outliers in the extracted features. This approach aims to ensure that outlier handling is conducted only once, thus optimizing the time required to achieve the analytical goals.

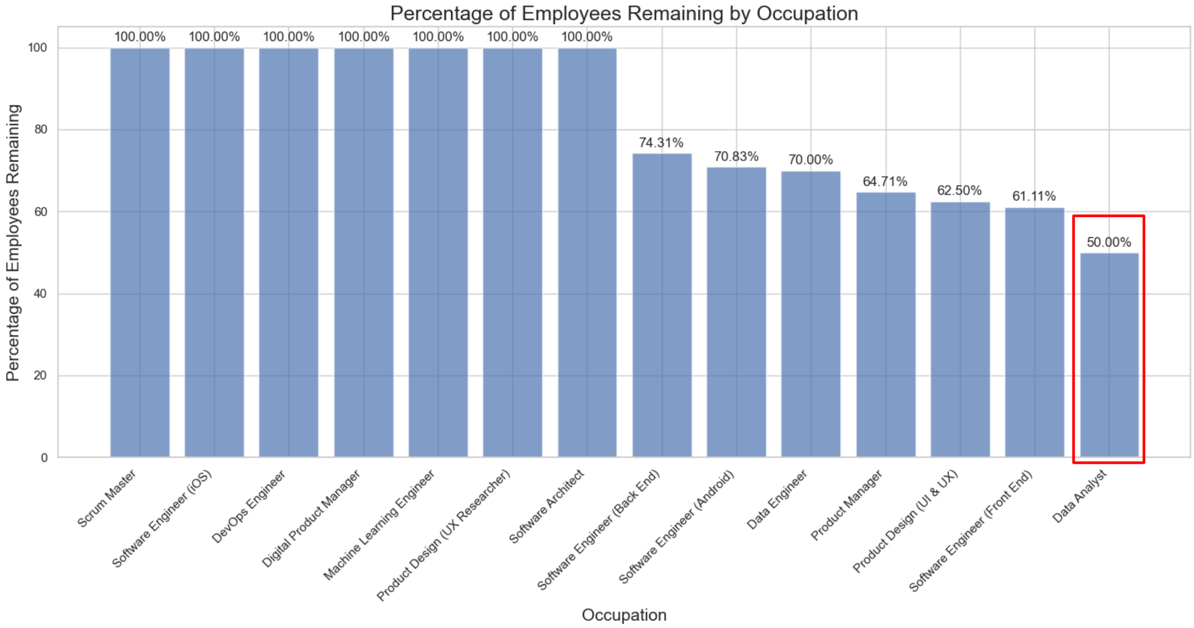
Annual Report on Employee Number Changes

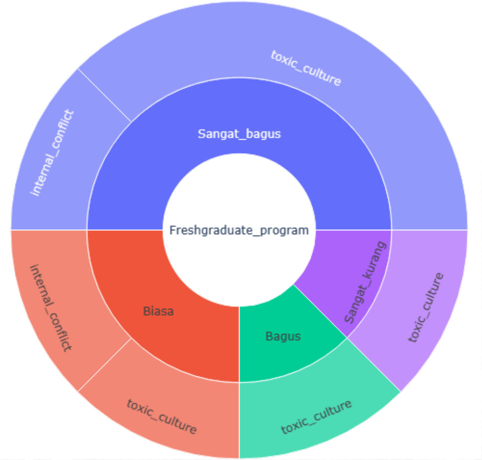

Total Hiring and Total Resign per Year


Annual Employee Count

- The company experienced consistent employee growth from 2006 to 2014, with a significant peak in 2011. A drop in new hires in 2012 was likely due to reaching full staffing levels after several active recruitment years and no employee resignations from 2006 to 2012. However, the rise in new hires in 2013 and 2014 coincided with the beginning of employee resignations.
- The number of resigning employees continued to increase, peaking in 2018. From 2017 to 2020, employee resignations surpassed new hires, indicating a concerning shift compared to the prior healthy years when new hires exceeded resignations.

Resign Reason Analysis for Employee Attrition Management Strategy



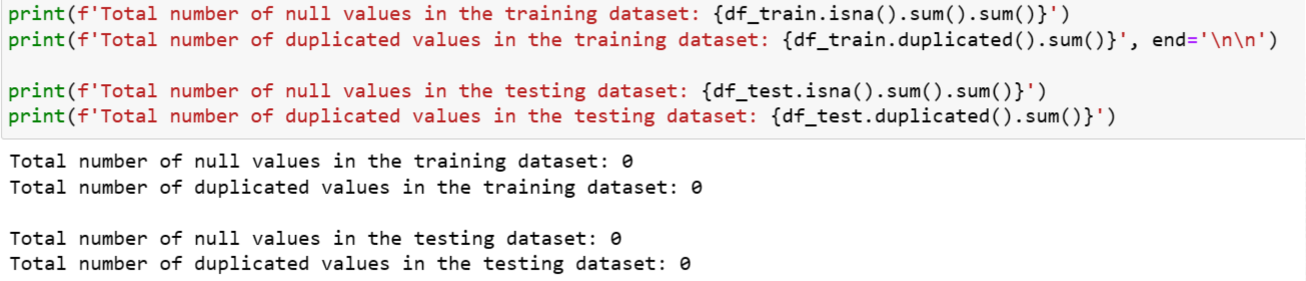


Data Analyst Resignation Based on JenjangKarir, Performance, and AlasanResign

Data Analyst has the highest resignation rate, and it's evident that 38% of the resigning employees have an excellent performance rating due to an unhealthy work culture. Recommendations include the need for improved communication, creating a safe working environment both emotionally and physically, and making improvements based on employee feedback (source: wheniwork.com).

Build an Automated Resignation Behavior Prediction using Machine Learning

Check Dataset



```
print(f'Total number of null values in the training dataset: {df_train.isna().sum().sum()}')
print(f'Total number of duplicated values in the training dataset: {df_train.duplicated().sum()}', end='\n\n')

print(f'Total number of null values in the testing dataset: {df_test.isna().sum().sum()}')
print(f'Total number of duplicated values in the testing dataset: {df_test.duplicated().sum()}')
```

```
Total number of null values in the training dataset: 0
Total number of duplicated values in the training dataset: 0

Total number of null values in the testing dataset: 0
Total number of duplicated values in the testing dataset: 0
```

Both of training and testing datasets contains no duplicates and missing values

Feature Extraction

The following features are the result of the extraction process:

1. **Umur**, which represents the employee's age at the time of assessment.
2. **LamaBekerja**, which is the number of years an employee has worked.
3. **Kepuasan-Keluhan**, is a combined column derived from the **SkorKepuasanPegawai** column and an extraction from the **AlasanResign** column. The reason from for resignation is extracted to obtain complaints data, allowing it to be available when assessments are conducted not only for resigning employees but also for those who do not resign. This enables the company to perform evaluations and improvements as a preventive measure to retain existing employees.

Build an Automated Resignation Behavior Prediction using Machine Learning

Handling Outliers

Handling outliers is performed using the **manually trimmed method**, as it provides a more selective approach to outlier handling and is designed to anticipate potential data loss when dealing with more diverse datasets.

Feature Scaling

Feature scaling is performed using **StandardScaler (standardization),** with the consideration that columns with scores and label encoding results are not scaled, and their data range is from 0-5. This is done to minimize the impact of significant scale differences between features and ensure that it does not affect the performance of the model later.

Feature Encoding

Feature encoding will be performed using **label encoding** and **one-hot encoding**.

- Label encoding will be applied to columns with categorical values that have ordinal levels,
- Meanwhile, one-hot encoding will be applied to other categorical columns that do not have ordinal levels.

Build an Automated Resignation Behavior Prediction using Machine Learning

Feature Selection

Feature selection was carried out using a variety of methods, including **ANOVA, Variance Threshold, Mutual Information, Select K-Best, Feature Importance by ExtraTreesClassifier,** and **Heatmap**.

Handling Imbalance

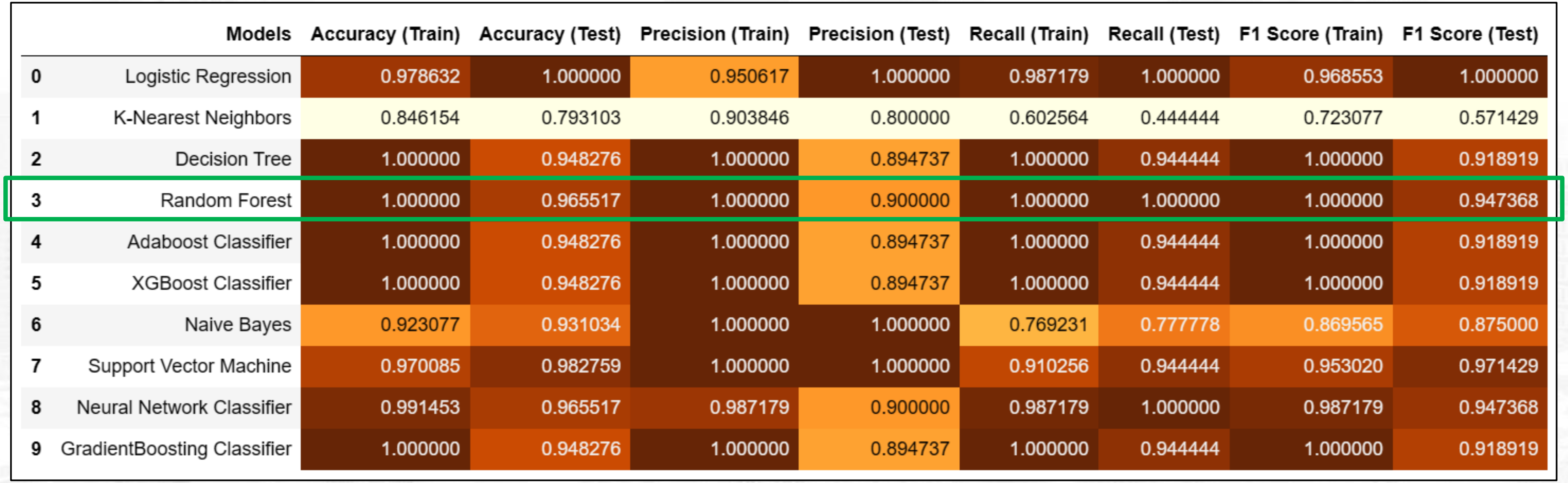
Handling the imbalance is performed using **oversampling with SMOTE**, as there is a significant difference between employees who resign (30%) and those who do not (70%). The use of SMOTE also aims to avoid potential overfitting that may occur with regular oversampling.

Modeling

Machine Learning models to be created include: Logistic Regression, k-Nearest Neighbors (kNN), Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, XGBoost Classifier, Gaussian Naive Bayes, Support Vector Machine, Neural Network Classifier, and Gradient Boosting Classifier.

Build an Automated Resignation Behavior Prediction using Machine Learning

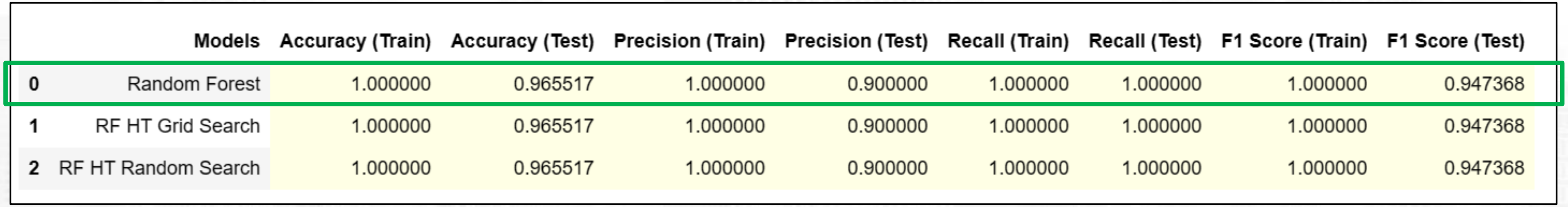
The table below displays the performance of machine learning classification for each model.



	Models	Accuracy (Train)	Accuracy (Test)	Precision (Train)	Precision (Test)	Recall (Train)	Recall (Test)	F1 Score (Train)	F1 Score (Test)
0	Logistic Regression	0.978632	1.000000	0.950617	1.000000	0.987179	1.000000	0.968553	1.000000
1	K-Nearest Neighbors	0.846154	0.793103	0.903846	0.800000	0.602564	0.444444	0.723077	0.571429
2	Decision Tree	1.000000	0.948276	1.000000	0.894737	1.000000	0.944444	1.000000	0.918919
3	Random Forest	1.000000	0.965517	1.000000	0.900000	1.000000	1.000000	1.000000	0.947368
4	Adaboost Classifier	1.000000	0.948276	1.000000	0.894737	1.000000	0.944444	1.000000	0.918919
5	XGBoost Classifier	1.000000	0.948276	1.000000	0.894737	1.000000	0.944444	1.000000	0.918919
6	Naive Bayes	0.923077	0.931034	1.000000	1.000000	0.769231	0.777778	0.869565	0.875000
7	Support Vector Machine	0.970085	0.982759	1.000000	1.000000	0.910256	0.944444	0.953020	0.971429
8	Neural Network Classifier	0.991453	0.965517	0.987179	0.900000	0.987179	1.000000	0.987179	0.947368
9	GradientBoosting Classifier	1.000000	0.948276	1.000000	0.894737	1.000000	0.944444	1.000000	0.918919

The top-performing model is the **Random Forest Classifier**, as it exhibits the best recall performance on both the training and testing data.

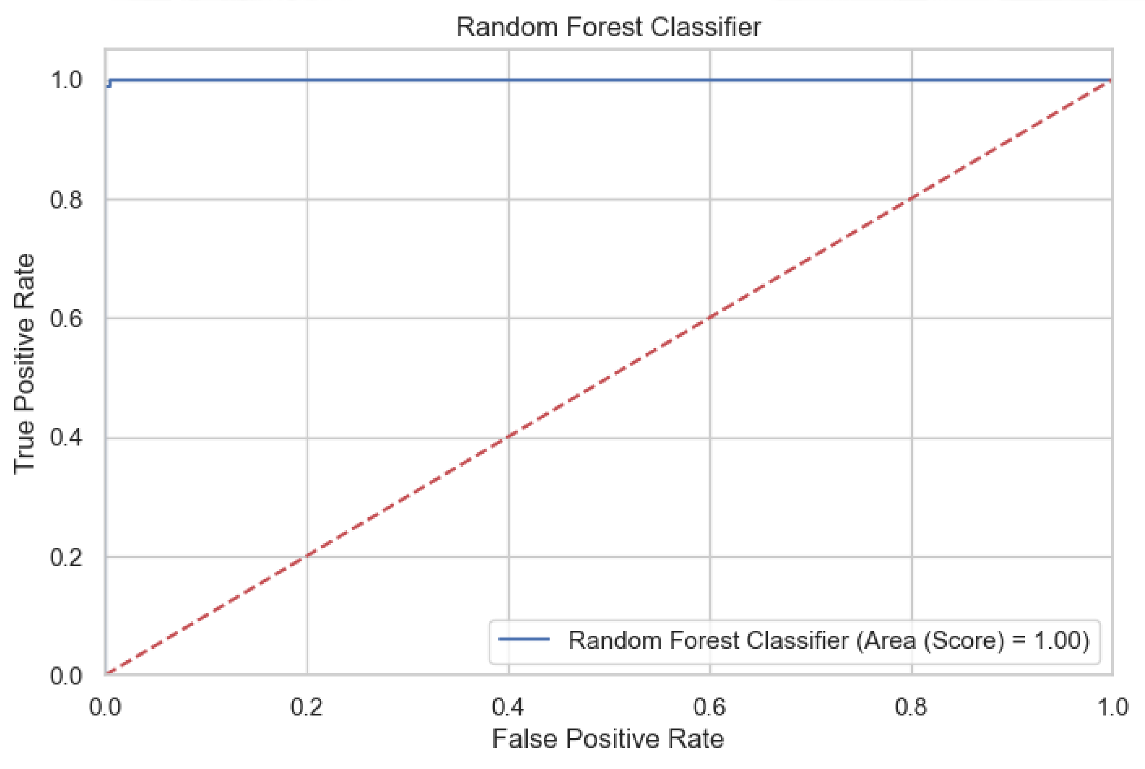
The table below displays the performance of Random Forest Classifier before and after tuning.



	Models	Accuracy (Train)	Accuracy (Test)	Precision (Train)	Precision (Test)	Recall (Train)	Recall (Test)	F1 Score (Train)	F1 Score (Test)
0	Random Forest	1.000000	0.965517	1.000000	0.900000	1.000000	1.000000	1.000000	0.947368
1	RF HT Grid Search	1.000000	0.965517	1.000000	0.900000	1.000000	1.000000	1.000000	0.947368
2	RF HT Random Search	1.000000	0.965517	1.000000	0.900000	1.000000	1.000000	1.000000	0.947368

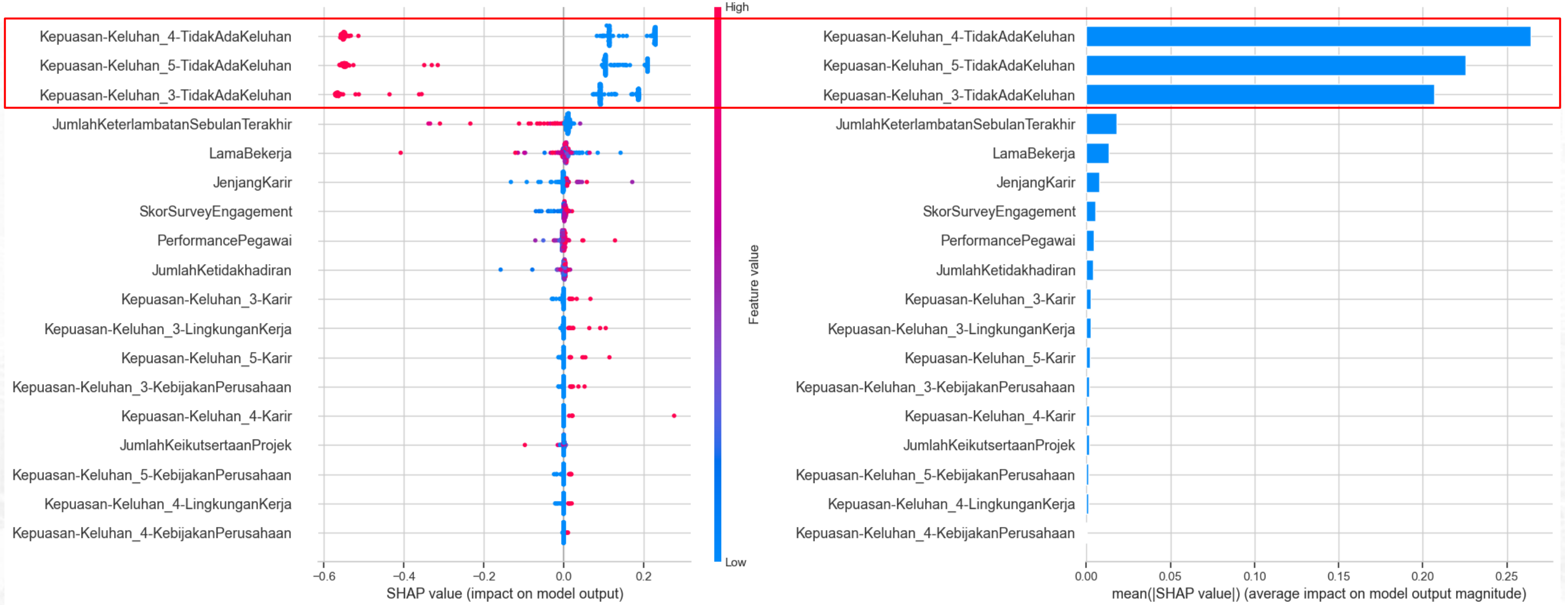
Both before and after hyperparameter tuning using grid search and random search, the models exhibit similar performance across all metrics. This is because from the beginning, the models were configured to use cross-validation, ensuring that the performance of the models before tuning was already optimized. **The chosen model is the untuned random forest**, as this decision was made to streamline the analysis process and save time.

Build an Automated Resignation Behavior Prediction using Machine Learning



The ROC AUC value of the selected model is 1, indicating that the model performs very well in distinguishing between positive and negative classes.

Presenting Machine Learning Products to the Business Users



The three most important features obtained from the model are **Kepuasan-Keluhan_4-TidakAdaKeluhan**, **Kepuasan-Keluhan_5-TidakAdaKeluhan**, and **Kepuasan-Keluhan_3-TidakAdaKeluhan**. Their feature values lean to the left, indicating that if there are no complaints and the satisfaction scores given by employees range from 3 to 5, employees tend to avoid attrition.

- Therefore, the recommendation that can be provided is to **Increase Employee Satisfaction**: Focus on improving employee satisfaction by actively seeking input and feedback from them. Actively seek input from employees and promptly address any necessary improvements. Ensure they feel heard, and that their complaints are taken seriously.