FINAL PROJECT PEMBELAJARAN MESIN

ANALISIS DAN PEMODELAN PREDIKSI PERGANTIAN KARYAWAN MENGGUNAKAN MODEL *MACHINE LEARNING* KLASIFIKASI

Kelompok 1

Delai Resgista Setyawan, 5025221221, 2022

Adyuta Prajahita Murdianto, 5025221186, 2022

Muhammad Rifqi Ma'ruf, 5025221060, 2022

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS

DEPARTEMEN TEKNIK INFORMATIKA

2024

BAB I. PENDAHULUAN

1.1 Latar Belakang

Pergantian karyawan adalah salah satu tantangan utama yang dihadapi oleh departemen sumber daya manusia di berbagai organisasi atau perusahaan. Pergantian karyawan yang tinggi tidak hanya berdampak pada biaya rekrutmen dan pelatihan yang meningkat, tetapi juga dapat mengganggu produktivitas, moral tim, dan stabilitas organisasi secara keseluruhan. Oleh karena itu, memahami faktor-faktor yang berkontribusi terhadap *attrition employee* atau pergantian karyawan dan mengembangkan model prediktif untuk mengidentifikasi karyawan yang berisiko tinggi untuk meninggalkan perusahaan adalah langkah penting dalam strategi manajemen sumber daya manusia.

Dalam upaya ini, *Synthetic Employee Attrition Dataset* telah dikembangkan sebagai *dataset* simulasi yang dirancang untuk analisis dan prediksi *attrition* karyawan. *Dataset* ini mencakup informasi terperinci tentang berbagai aspek profil karyawan, termasuk demografi, fitur terkait pekerjaan, dan kondisi pribadi. Dengan total 64,498 sampel yang dibagi menjadi set pelatihan dan pengujian. *Dataset* ini memberikan basis yang kuat untuk pengembangan dan evaluasi model *machine learning* yang bertujuan untuk memprediksi *attrition*.

Setiap baris dalam dataset mencakup ID Karyawan yang unik serta fitur-fitur yang mempengaruhi *attrition* karyawan, Tujuannya adalah untuk memahami faktor-faktor yang berkontribusi terhadap *attrition* dan mengambangkan model prediktif untuk mengidentifikasi karyawan yang berisiko. *Dataset* ini sangat ideal untuk analitik HR, pengembangan model *machine learning*, dan demonstrasi teknik analisis data tingkat lanjut. Dengan memberikan pandangan yang komprehensif dan realistis tentang faktor-faktor yang mempengaruhi retensi karyawan, *dataset* ini menjadi sumber daya yang berharga bagi peneliti dan praktisi di bidang sumber daya manusia dan pengembangannya.

Penggunaan *dataset* ini memungkinkan perusahaan untuk mengidentifikasi pola dan tren yang mungkin tidak terlihat melalui analisis tradisional. Dengan menggunakan metode *machine learning*, perusahaan dapat membuat prediksi yang lebih akurat tentang karyawan mana yang berisiko tinggi untuk meninggalkan perusahaan dan mengambil tindakan proaktif untuk mempertahankan mereka yang berharga. Hal ini

tidak hanya dapat mengurangi biaya yang terkait dengan pergantian karyawan, tetapi juga dapat membantu menciptakan lingkungan kerja yang lebih stabil dan produktif.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, diperloeh rumusan masalah sebagai berikut:

1. Apa model yang paling optimal antara random forest, naive bayes, svm, knn, d3, logistrik regresion, XGBoost, dan adaboost dalam klasifikasi *attrition* karyawan?
2. Apakah penggunaan normalisasi dan hyper parameter tunning mampu meningkatkan hasil klasifikasi?

1.3 Batasan Masalah

Adapun batasan masalah yang perlu diperhatikan antara lain:

1. Projek akhir ini menggunakan dataset sintetik yang tidak sepenuhnya mencerminkan kompleksitas dan variasi kondisi nyata di berbagai tempat. Sehingga hasil dari penelitian tidak dapat digeneralisasi ke semua sektor industri
2. Data yang digunakan mencerminkan kondisi pada saat dataset ini dibuat. Sehingga tidak mencerminkan perubahan tren atau kondisi pasar tenaga kerja yang terjadi pada saat setelah dataset dibuat.

1.4 Tujuan

Tujuan yang diharpakan pada projek akhir ini adalah sebagai berikut:

1. Mengidentifikasi model terbaik dalam klasifikasi *attrition* karyawan menggunakan model random forest, naive bayes, svm, knn, d3, logistrik regresion, XGBoost, dan adaboost.
2. Membuktikan pengaruh normalisasi dan hyper parameter tunning mampu meningkatkan hasil klasifikasi.

1.5 Manfaat

Adapun manfaat penelitian ini adalah sebagai berikut:

1. Sebagai mahasiswa, penelitian ini membantu pemahaman penggunaan *machine learning* dalam tugas klasifikasi untuk penerapan di dunia nyata.
2. Sebagai pemangku manajemen perusahaan, penelitian ini membantu perusahaan memahami faktor-faktor utama yang mempengaruhi *attrition* karyawan dan menggunakan model prediktif untuk mengoptimalkan strategi *retensi* karyawan.

3. Sebagai peneliti, penelitian ini menjadi dasar bagi penelitian lebih lanjut di bidang *attrition* karyawan, baik dalam konteks metodologi *machine learning* maupun eksplorasi faktor-faktor tambahan yang mempengaruhi retensi karyawan.

BAB II. TINJAUAN PUSTAKA

2.1 *State of The Art*

Adapun pada Tabel 2.1 merupakan penelitian terdahulu yang berhubungan dengan perancangan model dengan konsep *atrition* employee.

Judul	Cakupan Pembahasan dan Hasil
Explaining and Predicting Employees Attrition: A Machine Learning Approach	Penelitian ini menggunakan dataset yang mencakup berbagai fitur terkait karyawan, seperti data demografis dan informasi pekerjaan. Metodologi yang diterapkan berfokus pada analisis univariat dan bivariat serta penggunaan svm, decision tree, dan random forest untuk memahami dan memprediksi pergantian karyawan. Hasil penelitian menunjukkan bahwa random forest memberikan akurasi mencapai 99% dibandingkan metode lainnya dalam mengidentifikasi fitur-fitur seperti gaji dan jumlah proyek yang ditangani sebagai indikator penting untuk prediksi attrition. Kelebihan utama dari pendekatan ini adalah visualisasi data yang komprehensif, yang membantu dalam pemahaman mendalam tentang hubungan antar variabel. Namun, kekurangannya adalah bahwa model ini mungkin kurang akurat dalam situasi dengan data yang sangat dinamis atau tidak terstruktur dengan baik
Predicting Employee Attrition Using Machine Learning Techniques	Penelitian ini memanfaatkan dataset IBM yang berisi 35 fitur dan sekitar 1500 sampel. Beberapa algoritma klasifikasi diterapkan, termasuk Gaussian Naïve

	Bayes, yang memberikan performa terbaik dengan recall rate sebesar 54 persen dan false negative rate sebesar 4.5%. Kelebihan dari algoritma ini adalah kemampuannya dalam mengidentifikasi semua instance positif, namun recall rate yang relatif rendah menunjukkan bahwa beberapa instance positif mungkin terlewat.
A Machine Learning Approaches for Employee Retention Prediction	Penelitian ini menggunakan dataset yang berisi berbagai atribut terkait karyawan seperti pendidikan, pengalaman dan lainnya. Menggunakan knn, random forest dan svm, menghasilkan hasil terbaik dengan akurasi training, testing dan overall berturut turut 99.1%, 84.6% dan 91.8%. Kelebihan dari penelitian ini adalah eksplorasi mendalam terhadap berbagai teknik machine learning dan deep learning untuk meningkatkan akurasi prediksi. Namun, kekurangannya terletak pada kurangnya variasi dalam jenis data yang dapat membatasi generalisasi model ke industri lain
Turnover Prediction in a Call Center: Behavior Evidence of Loss Aversion Using Random Forest and Naive Bayes Algorithm	Penelitian ini menggunakan dataset dari call center yang fokus pada perilaku karyawan. Random Forest dan Naive Bayes digunakan cukup kompetitif untuk memprediksi turnover berdasarkan data perilaku, dengan random forest berhasil memberikan hasil terbaik dalam hal akurasi sebesar 85% pada random forest. Kelebihan utama dari model ini adalah efektivitasnya dalam menangkap pola

	perilaku yang kompleks, meskipun analisis perilaku membutuhkan data yang sangat terperinci yang mungkin tidak terlalu tersedia.
Prediction of Employee Turnover in Organization Using Machine Learning Algorithms	Penelitian ini menggunakan dataset yang berisi berbagai informasi terkait demografi dan atribut pekerjaan. Algoritma Support Vector Machine digunakan untuk memprediksi niat karyawan untuk keluar, menunjukkan performa yang baik dengan akurasi yang tinggi. XGBoost sangat efektif untuk dataset dengan margin yang jelas antara kelas-kelas yang berbeda, tetapi bisa menjadi kurang efektif pada dataset yang sangat besar atau tidak terstruktur dengan baik. Performa terbaik AUC XGBoost mencapai 88%.

BAB III. METODOLOGI

3.1 Pengumpulan Data

Dataset diambil dari platform kaggle dengan nama *employee attrition*. Dataset bisa dilihat dari tautan berikut:

Employee Attrition Classification Dataset (kaggle.com)

Dataset dengan format .csv ini berisi train dan test set untuk melakukan prediksi. Dataset memiliki ukuran 9.55 MB, setiap baris dari dataset ini menyimpan tentang detail informasi dari variasi atribut profil karyawan, termasuk demografi, dan kebutuhan personal. Dataset memiliki total 74,498 sampel dan 24 kolom dengan detail fitur sebagai berikut:

1. *Employee ID:* Angka unik sebagai pengidentifikasi setiap karyawan
2. *Age:* Umur Karyawan (18-60 tahun)
3. *Gender:* Jenis kelamin (Laki-laki / perempuan)
4. *Years at Company:* Jumlah tahun karyawan telah bekerja
5. *Montly Income:* Gaji bulanan dalam dolar
6. *Job Role:* Peran tempat karyawan bekerja
7. *Work-Life Balance:* Keseimbangan dalam berkerja (Buruk, di bawah rata-rata)
8. *Job Satisfaction:* Kepuasan dalam bekerja (Rendah, sedang, tinggi)
9. *Performance Rating:* Peringkat kinerja (Rendah, Tinggi)
10. *Number of Promotions:* Total promosi yang diterima
11. *Overtime:* Kerja lembur
12. *Distance from Home:* Jarak dari rumah
13. *Education Level:* Jenjang pendidikan
14. *Martial Status:* Status perkawinan
15. *Number of Dependents:* Jumlah tanggungan
16. *Job Level:* Tingkat pekerjaan (Menengan, senior)
17. *Company Size:* Ukuran pekerjaan
18. *Company Tenure:* Jumlah total tahun karyawan bekerja
19. *Remote Work:* Apakah karyawn bekerja dari ajrak jauh (Ya atau Tidak)
20. *Leadership Opportunities:* Apakah karyawan memiliki peluang kepemimpinan
21. *Invovation Opportunities:* Apakah karyawan mempunyai peluang untuk berinovasi

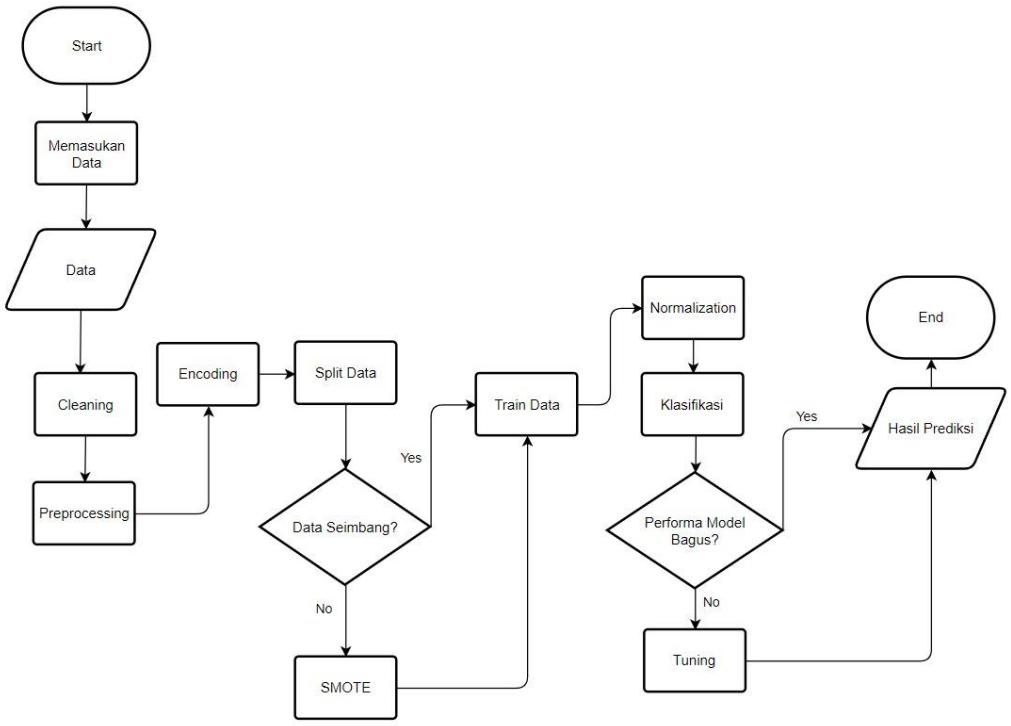
22. *Company Repotation:* Persepsi karyawan terhadap reputasi perusahaan
23. *Employee Recognition:* Tingkat pengakuan yang diterima (Sangat rendah, rendah)
24. *Attrition:* Apakah karyawan telah keluar dari perusahaan, jika 0 (tetap) dan 1 (keluar)

Dataset memiliki 0 nilai null dan 0 nilai duplicated value. Dengan masing masing persebaran nilai unik dan tipe data masing-masing kolom adalah sebagai berikut:

Fitur	Tipe Data	N-Unique
Employee IDE	Int64	74498
Age	Int64	42
Gender	Object	2
Years at Company	Int64	51
Job Role	Object	5
Monthly Income	Int64	9842
Work-Life Balance	Object	4
Job Satisfaction	Object	4
Performance Rating	Object	4
Number of Promotions	Int64	5
Overtime	Object	2
Distance from Home	Int64	99
Education Level	Object	5
Martial Status	Object	3
Number of Dependents	Int64	7
Job Level	Object	3
Company Size	Object	3
Company Tenure	Int64	127
Remote Work	Object	2
Leadership Opportunities	Object	2
Inovation Opportunities	Object	2
Company Reputation	Object	4
Employee Recognition	Object	4

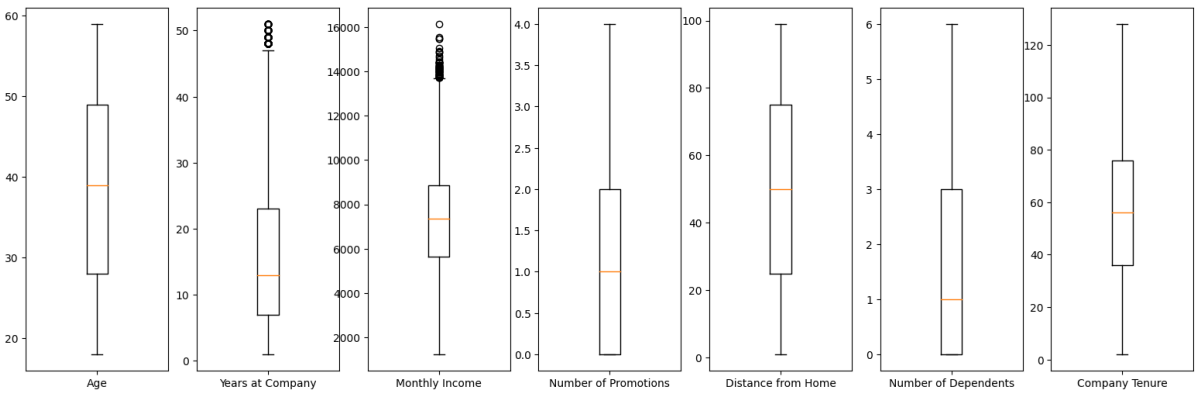
Attrition	Object	2

3.2 Diagram Alir

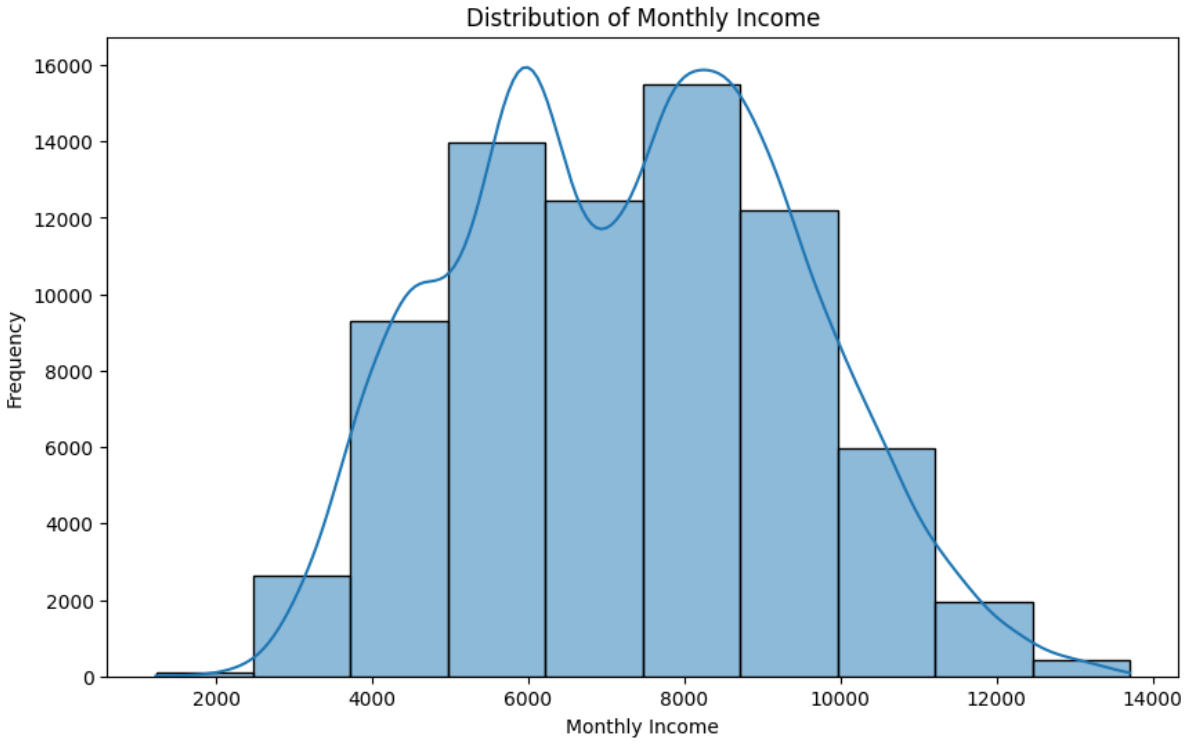


3.3 Tahap Cleaning

Dataset yang digunakan tidak memiliki nilai null dan duplicate value. Pertama yang dilakukan adalah drop kolom “Employee ID” karena tidak digunakan dalam proses training data. Kemudian tahapan ini dilakukan handling outlier pada fitur dengan tipe data numeric.



Terlihat bahwa di kolom "**years at company**" dan "**montly income**" terdapat beberapa data poin yang dianggap sebagai *outlier* karena sedikit menjauhi batas atasnya. Pada awalnya kami menganggap ini tidak begitu mempengaruhi data, karena bisa dianggap sebagai kasus khusus seperti gaji untuk pekerjaan spesial dan karyawan yang masih belum mengambil masa pensiunnya. Namun, pada kasus ini kami membersihkannya menggunakan perhitungan kuartil untuk menjaga kesempurnaan data. Perhitungan kuartil, menghitung batas bawah dan batas atas untuk setiap kelom, dengan mengurangi 1,5 kali IQR dari Q1 dan menambah 1,5 kali IQR ke Q3



3.4 Tahap Preprocessing

3.4.1 Encoding

Dataset yang sudah dibersihkan masih memiliki data kategori, sedangkan komputer hanya memahami bentuk bilangan bulat. Sehingga perlu dilakukan encodinng untuk data kategorical. Metode encoding yang kita pakai adalah label encoding. Label encoding dipakai karena sebagian besar fitur-fitur kategori memiliki hubungan ordinal atau urutan seperti "rendah", "sedang", "tinggi" seperti pada kolom "Performance", "Work Life Balance", "Job Satisfaction". Kolom lainnya seperti gender, job role dan remote work juga

dilakukan label encoding dengan tujuan effisiensi dan menyederhanakan bentuk data.

Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions
0	0	0	2	0	2
3	0	3	0	3	3
2	2	2	0	3	0
0	0	2	0	2	1
0	0	1	3	0	0

3.4.2 Feature Scaling

	Age	Gender	Years at Company	Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions	Overtime	Distance from Home	Education Level
0	31	1	19	0	5390	4	2	3	2	0	22	3
1	59	0	4	3	5534	1	3	1	3	0	21	4
2	24	0	10	2	8159	3	3	1	0	0	11	2
3	36	0	7	0	3989	3	3	4	1	0	27	1
4	56	1	41	0	4821	2	4	3	0	1	71	1

Perhatikan gambar di atas, terlihat bahwa setiap kolom memiliki range nilai yang berbeda. Model tanpa fitur scaling dapat menghasilkan prediksi yang didominasi oleh fitur yang memilikinilai yang tersebar. Sehingga fitur scaling akan menyeimbangkan terhadap hasil prediksi model. Berikut adalah hasil fitur scaling dengna standar scaler, standard scaler dipilih karena baik digunakan untuk distribusi data yang sudah seimbang.

	Age	Gender	Years at Company	Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions	Overtime
0	-0.623178	0.908078	0.292028	-1.440117	-0.887902	1.493622	-0.921920	0.066919	1.172770	-0.696701
1	1.694050	-1.101226	-1.044399	0.598637	-0.820726	-1.705498	0.223677	-2.635719	2.177579	-0.696701
2	-1.202486	-1.101226	-0.509828	-0.080948	0.403834	0.427248	0.223677	-2.635719	-0.836848	-0.696701
3	-0.209388	-1.101226	-0.777113	-1.440117	-1.541467	0.427248	0.223677	1.418238	0.167961	-0.696701
4	1.445776	0.908078	2.252121	-1.440117	-1.153340	-0.639125	1.369275	0.066919	-0.836848	1.435336

3.5 Tahap Training

Data dieskpor ke dalam dua file .csv berbeda. Dataset pertama berisi dataset yang dilakukan feature scaling yang akan digunakan untuk skenario 1 dan 3. Kemudian,

file kedua berisi dataset tanpa dilakukan feature scaling yang akan digunakan skenario 2. Ketiga skenario tersebut akan dilakukan training dengan 7 model .

1. Logistik Regression

   Logistik regression adalah model statistik yang digunakan untuk memprediksi probabilitas kejadian biner. Model ini menggunakan fungsi logit untuk mengubah nilai input menjadi probabilitas Fungsi logit tersebut didefinisikan sebagai berikut.

   $$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

2. Naive Bayes

   Naive bayes adalah algoritma berbasis probabilitas yang mengasumsikan independesi antar fitur dengan menghitung probabilitas masing-masing kelas. Menggunakan Teorema Bayes untuk menghitung probabilitas posterior dari setiap kelas untuk instance baru dan memprediksi kelas dengan probabilitas tertinggi.

3. KNN

   KNN (K-Neirest Neighbor) menyimpan semua instance dari training data untuk mengklasifikasikan instance baru. KNN menghitung jarak antara instance baru dan semua instance dalam training data. Instance baru akan diklasifikasi berdasarkan mayoritas kelas dari k tentangga terdekatnya.

4. Decisiton Tree

   Model prediktif yang menggunakan struktur pohon keputusan untuk membuat prediksi. Cara kerja model ini memisahkan fitur yang memaksimalkan gain informasi atau mengurangi impurity seperti Gini impurity atau entropy). Kemudian, membangun tree berdasarkan pemisahan tersebut dan membuat prediksi berdasarkan fitur input.

5. Random Forest

   Salah satu metode ensemble learning yang mengadopsi konsep decision tree. Model ini akan mengambil sampelacak dari dataset. Setelah itu membangun decision tree pada setiap sampel, dimana tiap tree dilatih pada suset acar dan menggabungkan prediksi dari semua pohon melalui majority fote untuk klasifikasinya.

6. XGBoost

Termasuk algoritma boosting yang meningkatkan akurasi model dengan menggabungkan banyak model sederhana. Pertama XGBoost akan membangun model secara bertahap dan menggabungkannya. Lalu, setiap model baru akan memberikan perbaikan dari kesalahan model sebelumnya. Terakhir mengunakan fungsi objective untuk mengoptimalkan loss function dan regularization untuk mencega overfitting.

7. AdaBoost

   Algoritma boosting yang mengkombinaksikan banyak weak learners. Model ini bekerja dengan membangun model secara bertahap, di setiap model baru berfokus pada instance yang salah diklasifikasikan oleh model sebelumnya. Kemudian memberikan bobot lebih pada instance yang sulit untuk meningkatkan akurasi keseluruhan. Dan terakhir menggabungkan semua model dengan bobot yang ditentukan berdasarkan kinerjanya.

Ada 3 skenario yang akan digunakan untuk menentukan model optimal untuk mengklasifikasn attrition employe. Skenario pertama model akan ditraining menggunakan dataset yang sudah melalui scaling. Skanario kedua akan ditraining tanpa menggunakan scaling. Skenario terakhir model akan ditraining menggunakan dataset yang sudah scaling dan melalui hyper parameter tuning.

2.6 Testing dan Evaluasi

Untuk setiap skenario, kita akan menggunakan evaluasi metrik Accuracy Score, Precision Scorem Recall, dan F1 score sebagai parameter hasil model terbaik.

a. Akurasi

   Proporsi total prediksi yang benar (TAPI dan TN_) dari keseluruhan prediksi

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

b. Presisi

   Kemampuan model untuk tidak membuat kesalahan positif yang palsu, ini dihitung sebagai rasio TAPI terhadap total prediksi positif.

$$\text{Precision} = \frac{TP}{TP + FP}$$

c. Recall

Kemampuan model untuk mendeteksi semua kasus positif, dihitung sebagai rasio TAPI terhadap semua kasus yang sebenarnya positif.

$$\text{Recall} = \frac{TP}{TP + FN}$$

d. F1 Score

F1 score didefinisikan sebagai hamonic mean dari presisi dan reacall.

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Berikut adalah penjelasan mengenai konsep hasil prediksi kelas target:

1. True Positive (TAP)

   True positive terjadi ketika model klasifikasi dnegan benar memprediksi instansi positif sebagai positif. Contoh dalam sebuah tes model untuk mendeteksi attrition, TP berarti model benar benar mengidentifikasi karyawan yang leave sebagai leave.

2. True Negative (TN)

   True negative terjadi ketika model klasifikasi dengan benar memprediksi instansi negatif sebagi negatif. Contoh dalam konteks ini. TN berarti model benar mengidentifikasi karyawan yang stayed sebagai stayed.

3. False Positif (FP)

   False positif terjadi ketika model klasifikasi salah memprediksi instansi negatif. Dalam kasus ini, FN berarti model salah mengidetifikasi karyawan yang leave sebagai not stayed

4. False Negatif (FN)

   Fakse negatif terjadi ketika model salah memprediksi instansi postif sebagai positif. Dalam kasus ini, FN berarti model salah mengidentifikasi indifiu yang stayed sebagai leave.

BAB III. HASIL DAN PEMBAHASAN

3.1 Skenario I

Training 1, masing masing model dilakukan training menggunakan data yang sudah di scaling dan prediksi data dengan 80% training data, dan 20% test data, kemudian dihitung metriks evaluasinya.

No.	Model	Accuracy	Precision	Recall	F1
1.	Logistik Regression	0.73	0.71	0.72	0.71
2.	KNN	0.69	0.68	0.67	0.67
3.	Random Forest	0.75	0.73	0.74	0.74
4.	Decision Tree	0.66	0.64	0.65	0.64
5.	AdaBoost	0.76	0.75	0.75	0.75
6.	Naive Bayes	0.72	0.69	0.76	0.72
7.	XGBoost	0.75	0.73	0.74	0.74

3.2 Skenario II

Training 1, masing masing model dilakukan training menggunakan data tanpa dilakukan scaling dan prediksi data dengan 80% training data, dan 20% test data, kemudian dihitung metriks evaluasinya.

No.	Model	Accuracy	Precision	Recall	F1
1.	Logistik Regression	0.65	0.66	0.57	0.61
2.	KNN	0.52	0.49	0.47	0.48
3.	Random Forest	0.75	0.73	0.74	0.74
4.	Decision Tree	0.66	0.64	0.65	0.65
5.	AdaBoost	0.76	0.75	0.75	0.75
6.	Naive Bayes	0.74	0.69	0.76	0.72
7.	XGBoost	0.75	0.73	0.74	0.74

3.3 Skenario III

Training 1, dari skenario I dan II diambil 3 model yang dianggap paling stabil dan optimal dalam melakukan klasifikasi. Tiga model tersebut adalah Random Forest,

XGBoost dan AdaBoost. Lalu, 3 model ini akan dilakukan hyper parameter tunning dan hasil metriks nya sebagai berikut.

1. Random Forest

```python
# Hyperparameter Tuning on Random Forest Classifier
rf = RandomForestClassifier()

# Define the hyperparameter space
param_distributions_rf = {
    'n_estimators': randint(100, 1000),
    'max_depth': randint(1, 20),
    'min_samples_split': randint(2, 20),
    'min_samples_leaf': randint(1, 20),
    'max_features': ['auto', 'sqrt', 'log2']
}

# Initialize RandomizedSearchCV for Random Forest
random_search_rf = RandomizedSearchCV(rf, param_distributions_rf, n_iter=100, cv=5,
                                      random_state=42, n_jobs=-1)
random_search_rf.fit(X_train, y_train)
best_rf = random_search_rf.best_estimator_
y_pred_rf = best_rf.predict(X_test)

print("Random Forest Best Hyperparameters:", random_search_rf.best_params_)
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("Random Forest Classification Report:\n", classification_report(y_test, y_pred_rf))
```

2. XGBoost

```python
# Define the XGBoost model
xgb = XGBClassifier(use_label_encoder=False, eval_metric='logloss')

# Define the hyperparameter space
param_distributions_xgb = {
    'n_estimators': randint(100, 1000),
    'max_depth': randint(1, 20),
    'learning_rate': uniform(0.01, 0.3),
    'colsample_bytree': uniform(0.5, 1.0)
}
# Initialize RandomizedSearchCV for XGBoost
random_search_xgb = RandomizedSearchCV(xgb, param_distributions_xgb, n_iter=100,
                                       cv=3, random_state=42, n_jobs=-1)
random_search_xgb.fit(X_train, y_train)
best_xgb = random_search_xgb.best_estimator_
y_pred_xgb = best_xgb.predict(X_test)

print("XGBoost Best Hyperparameters:", random_search_xgb.best_params_)
print("XGBoost Accuracy:", accuracy_score(y_test, y_pred_xgb))
print("XGBoost Classification Report:\n", classification_report(y_test, y_pred_xgb))
```

3. AdaBoost

```python
# Define the AdaBoost model
ada = AdaBoostClassifier()

# Define the hyperparameter space
param_distributions_ada = {
    'n_estimators': randint(50, 500),
    'learning_rate': uniform(0.01, 2.0)
}

random_search_ada = RandomizedSearchCV(ada, param_distributions_ada, n_iter=100,
                                       cv=3, random_state=42, n_jobs=-1)
random_search_ada.fit(X_train, y_train)
best_ada = random_search_ada.best_estimator_
y_pred_ada = best_ada.predict(X_test)

print("AdaBoost Best Hyperparameters:", random_search_ada.best_params_)
print("AdaBoost Accuracy:", accuracy_score(y_test, y_pred_ada))
print("AdaBoost Classification Report:\n", classification_report(y_test, y_pred_ada))
```

Berikut adalah tabel hasil matrik evalusai untuk 3 model yang telah dilakukan hyper parameter tunning.

No.	Model	Accuracy	Precision	Recall	F1
1.	Random Forest	0.75	0.74	0.74	0.74
2.	XGBoost	0.76	0.74	0.75	0.75
3.	AdaBoost	0.76	0.75	0.75	0.75

3.4 Analisis Hasil

BAB IV. KESIMPULAN

4.1 Kesimpulan

Setelah dilakukan percobaan didapat kesimpulan yang menjawab permasalahan antara Lain:

1. Model Random Forest, XGBoost, dan AdaBoost cukup bersaing untuk melakukan prediksi.
2. Normalisasi berpengaruh pada hasil akurasi model dan Hyper Parameter Tunning tidak serta merta akan memberikan peningkatan akurasi model.

4.2 Saran

Adapun saran yang kami berikan dari penelitian ini adalah:

1. Menggunakan teknik feature enginineer atau feature selection
2. Menggunakan dataset yang lebih beragam
3. Menggunakan model deep learning untuk mendapatkan model yang optimal

BAB 5. REFERENSI

Jain, P. K., Jain, M., & Pamula, R. (2020). Title of the paper. *SN Applied Sciences, 2*, 757. https://doi.org/10.1007/s42452-020-2519-9

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers, 9*(4), 86. https://doi.org/10.3390/computers9040086

G. Marvin, M. Jackson and M. G. R. Alam, "A Machine Learning Approach for Employee Retention Prediction," *2021 IEEE Region 10 Symposium (TENSYMP)*, Jeju, Korea, Republic of, 2021, pp. 1-8, doi: 10.1109/TENSYMP52854.2021.9550921.

S. Zhang, "Challenges in KNN Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4663, Oct. 2022.

Wickramasinghe, I., Kalutarage, H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Comput* **25**, 2277–2293 (2021). https://doi.org/10.1007/s00500-020-05297-6

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *IT Department, Technical College of Informatics Akre, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq*. Vol. 02, No. 01, pp. 20–28. ISSN: 2708-0757.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86.

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee turnover prediction based on random forests and survival analysis. In *Web Information Systems Engineering–WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II 21* (pp. 503-515). Springer International Publishing.

Alsheref, F. K., Fattoh, I. E., & M. Ead, W. (2022). Automated prediction of employee attrition using ensemble model based on machine learning algorithms. *Computational Intelligence and Neuroscience*, *2022*(1), 7728668.

Lazzari, M., Alvarez, J. M., & Ruggieri, S. (2022). Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, *14*(3), 279-292.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, *134*, 113290.

Zhu, H. (2021). Research on human resource recommendation algorithm based on machine learning. *Scientific programming*, *2021*(1), 8387277.

Muncie, T. (2020). *Using machine learning models to predict student retention: Building a state-wide early warning system* (Doctoral dissertation, Morehead State University).

Paul, T., & Bommu, R. (2024). Strategic Employee Performance Analysis in the USA: Leveraging Intelligent Machine Learning Algorithms. *Unique Endeavor in Business & Social Sciences*, *3*(1), 113-124.

Sainju, B., Hartwell, C., & Edwards, J. (2021). Job satisfaction and employee turnover determinants in Fortune 50 companies: Insights from employee reviews from Indeed. com. *Decision Support Systems*, *148*, 113582.