



Stage 2 - Data Pre-Processing

Team:

1. Timothy Agalliasis
2. Bernita Berliana Noor Maghfiroh
3. Intan Gunawan
4. Mia Agustina Nurfadilah

Submission:

1. **Report:**  Stage 2 - Data Pre-Processing
 2. **Notebook:**  Stage 2 - Data Pre-Processing.ipynb
 3. **Powerpoint :** [Stage 2 Data Pre-Processing](#)
-

Data Cleansing

1. Handle missing values

	0				
EmployeeID	0				
EnvironmentSatisfaction	25	Gender	0		
JobSatisfaction	20	JobLevel	0		
WorkLifeBalance	38	JobRole	0		
Age	0	MaritalStatus	0	TotalWorkingYears	9
Attrition	0	MonthlyIncome	0	TrainingTimesLastYear	0
BusinessTravel	0	NumCompaniesWorked	19	YearsAtCompany	0
Department	0	Over18	0	YearsSinceLastPromotion	0
DistanceFromHome	0	PercentSalaryHike	0	YearsWithCurrManager	0
Education	0	StandardHours	0	JobInvolvement	0
EducationField	0	StockOptionLevel	0	PerformanceRating	0
EmployeeCount	0				

Terdapat beberapa kolom yang memiliki nilai kosong:

- EnvironmentSatisfaction: 25 nilai kosong
- JobSatisfaction: 20 nilai kosong
- WorkLifeBalance: 38 nilai kosong
- NumCompaniesWorked: 19 nilai kosong
- TotalWorkingYears: 9 nilai kosong

Nilai yang kosong dari EnvironmentSatisfaction, JobSatisfaction, dan WorkLifeBalance diisi dengan modus dari kolom tersebut karena cocok untuk kolom kategorik ordinal yang mewakili sebagian besar data. Sedangkan nilai yang kosong dari NumCompaniesWorked dan TotalWorkingYears diisi dengan mediannya karena lebih robust terhadap outliers.

2. Handle duplicated data

```
[ ] df_merge.duplicated().any()
```

```
➞ False
```

```
[ ] print('Jumlah data duplikat:', df_merge.duplicated().sum())
```

```
➞ Jumlah data duplikat: 0
```

Tidak terdapat duplicated data pada data set, sehingga tidak perlu di-handle.

3. Handle outliers

Tidak dilakukan penanganan pada outliers karena model yang akan digunakan adalah model yang robust terhadap outliers.

4. Feature encoding

- Mengubah tipe data menjadi kategorikal untuk feature 'EnvironmentSatisfaction', 'JobSatisfaction', 'WorkLifeBalance', 'Education', 'JobLevel', 'StockOptionLevel', 'JobInvolvement' dan 'PerformanceRating'.
- Kolom yang di OneHot Encoding: 'BusinessTravel', 'Department', 'EducationField', 'JobRole', 'MaritalStatus', dan 'Over18'.
- Kolom yang di Label Encoding: 'Gender' dan 'Attrition'.

5. Feature transformation

Sebelum feature transformation, terlebih dahulu dilakukan split antara data test dan data train. Fit scaling dengan standard scaler hanya dilakukan pada data train, kemudian transformasi data test menggunakan parameter scaling yang telah dipelajari dari data train.

6. Handle class imbalance

Terjadi class imbalance dengan rasio dari kelas Yes terhadap kelas No adalah sekitar 19.45%. Untuk mencegah kebocoran data (data leakage), akan dilakukan SMOTE terhadap data train saja, tidak pada data test.

Feature Engineering

1. Feature selection

Beberapa feature yang di-drop:

- Kolom EmployeeID di-drop karena hanya sebagai identify
- Kolom EmployeeCount, StandardHours dan Over18 dapat di-drop karena tidak memiliki variabilitas

2. Feature extraction

Beberapa feature baru yang dibangun dari feature yang sudah ada:

- **PromotionGap** (YearsAtCompany - YearsSinceLastPromotion): feature ini jika nilainya besar bisa menunjukkan karyawan tersebut sudah lama tidak dipromosikan dan mungkin merasa kurang dihargai, yang bisa memengaruhi tingkat kepuasan atau kemungkinan mereka keluar dari perusahaan.
- **AverageYearsPerCompany**: feature ini bisa memberikan wawasan tambahan tentang stabilitas pekerjaan karyawan. Jika karyawan sering berpindah perusahaan dalam waktu yang singkat, ini mungkin mengindikasikan ketidakpuasan dengan pekerjaan mereka atau pencarian untuk peluang yang lebih baik.

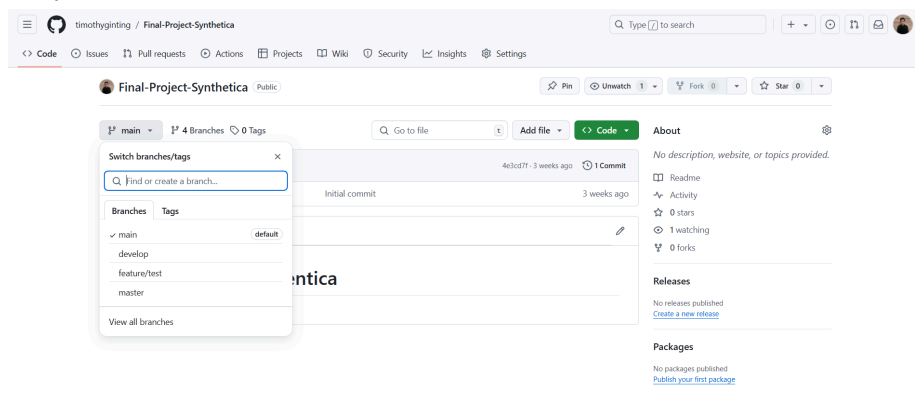
3. Beberapa feature tambahan yang mungkin akan sangat membantu membuat performansi model semakin bagus:

- **JobFitIndicator**: perbandingan antara level pekerjaan dengan jumlah pengalaman kerja yang dimiliki. Feature ini menggambarkan kesesuaian antara posisi yang dijabat karyawan dengan keterampilan dan latar belakang mereka. Jika seseorang memiliki banyak pengalaman tapi berada di level pekerjaan yang lebih rendah, mereka mungkin merasa kurang dihargai atau tidak sesuai dengan potensi mereka, yang bisa mendorong attrition.
- **Annual_Leave**: berapa kali cuti yang diambil karyawan selama setahun. Feature ini bisa memberikan gambaran bahwa karyawan yang merasa nyaman untuk mengambil cuti tahunan atau merasa perusahaan mendukung work-life balance dan kehidupan pribadi mereka, kemungkinan lebih puas dan lebih sedikit berisiko untuk meninggalkan perusahaan.

- **Vacation_Leave:** berapa kali karyawan mengambil cuti liburan selama setahun. Feature ini bisa memberikan gambaran bahwa karyawan yang merasa nyaman untuk mengambil cuti liburan atau merasa perusahaan mendukung work-life balance dan kehidupan pribadi mereka, kemungkinan lebih puas dan lebih sedikit berisiko untuk meninggalkan perusahaan.
- **Liabilities:** jumlah tanggungan yang perlu dibayar karyawan per bulan seperti hutang atau cicilan. Karyawan dengan kewajiban finansial yang besar mungkin lebih cenderung berpindah ke perusahaan lain yang menawarkan gaji lebih tinggi, tunjangan yang lebih baik, atau lebih banyak peluang untuk memperoleh bonus.

Git

Repository Git



Upload ke git

