Capstone Project- Husyen Ali Alhabsy

Data Classification and Summarization Using IBM Granite

Student Development Initiative

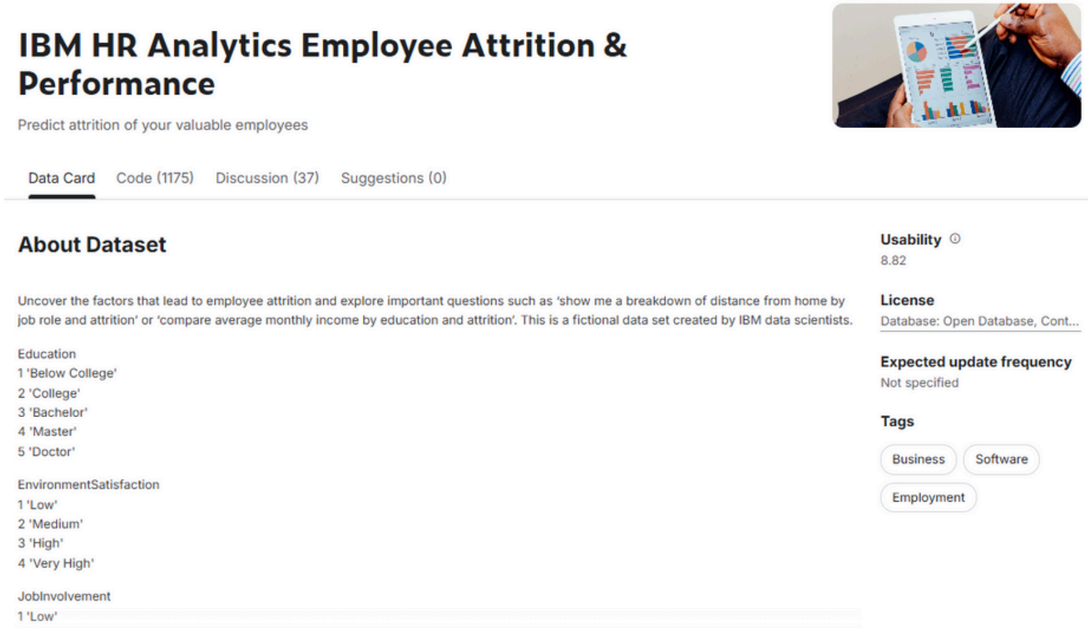
Raw Dataset Link

Dataset: IBM HR Analytics – Employee Attrition & Performance

Link: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Deskripsi Dataset:

- Dataset ini berisi data karyawan perusahaan fiktif IBM, termasuk informasi demografis, jabatan, performa, kepuasan kerja, dan riwayat attrition.
- Total terdapat 35 kolom (features) dan 1470 baris (karyawan).
- Fitur utama antara lain:
  - Age, Gender, MaritalStatus
  - JobRole, Department, YearsAtCompany
  - JobSatisfaction, PerformanceRating, WorkLifeBalance, OverTime
  - Attrition (target untuk analisis turnover karyawan)



Project Overview

Tujuan proyek ini adalah memahami faktor-faktor yang memengaruhi attrition dan kinerja karyawan di perusahaan menggunakan dataset HR IBM. Latar belakangnya muncul dari tingginya tingkat karyawan keluar yang berdampak pada produktivitas dan biaya rekrutmen.

Masalah spesifik yang relevan:

- Banyak karyawan meninggalkan perusahaan sebelum mencapai potensi penuh.
- Dampak negatif pada stabilitas tim dan keberlanjutan proyek.

Pendekatan yang dijalankan secara runtut:

1. Pengumpulan dan pembersihan dataset.
2. Eksplorasi data dan visualisasi tren.
3. Prediksi attrition menggunakan Logistic Regression.
4. Analisis data dan penyusunan insight dengan Granite LLM.
5. Rekomendasi actionable untuk HR berbasis data.

Dataset & Analytical Highlights

Kategori	Highlight
Dataset	IBM HR Employee Attrition & Performance (WA_Fn-UseC_-HR-Employee-Attrition.csv)
Target	Attrition (Yes/No)
Fitur Utama	Numerik: Age, YearsAtCompany, MonthlyIncome, JobSatisfaction Kategorikal: Department, JobRole
AI Support	Granite Agent + LangChain + Replicate untuk analisis, insight, & rekomendasi

Analysis Process

Proses analisis ini bertujuan mengonversi data mentah IBM HR Analytics menjadi informasi yang komprehensif dan bernilai strategis.

Setiap tahapan disusun secara sistematis agar hasil analisis akurat, mudah diinterpretasikan, serta dapat dijadikan dasar pengambilan keputusan oleh manajemen SDM. Langkah-langkah yang diterapkan dalam analisis dataset IBM HR Analytics adalah sebagai berikut:

- **Data Loading & Initial Check**
  - Dataset HR IBM diunggah ke Pandas DataFrame dan dilakukan pengecekan awal (head()) untuk memastikan data terbaca dengan benar.
- **Exploratory Data Analysis (EDA)**
  - Distribusi target (Attrition) dihitung untuk memahami proporsi karyawan yang keluar vs bertahan. Statistik deskriptif yang digunakan: Age, YearsAtCompany, MonthlyIncome, JobSatisfaction.
- **Data Preprocessing & Encoding**
  - Variabel kategorikal diubah menjadi numerik menggunakan LabelEncoder.
- **Predictive Modeling**
  - Logistic Regression pipeline dengan StandardScaler.
  - Evaluasi menggunakan classification_report untuk akurasi, precision, recall, F1-score.
- **AI-Assisted Analytical Insights**
  - Granite Agent digunakan untuk ringkasan statistik deskriptif, deteksi outlier dan istribusi karyawan per Department & JobRole.
  - Insight actionable dihasilkan untuk HR (misal: faktor risiko attrition, korelasi kepuasan kerja).
- **Visualization & Evidence Collection**
  - Boxplot, histogram, bar plot, dan Q-Q plot dipilih sebagai evidence karena:
    - Mudah dipahami stakeholder non-teknis.
    - Menunjukkan distribusi, perbandingan, dan potensi outlier.

HACKTIV8

Evidence Analysis Process

Data Loading & Initial Check

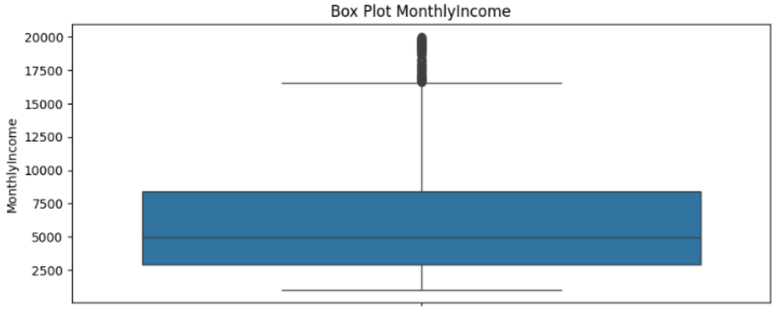
	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	..
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	..
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	..
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	..
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	..

5 rows × 35 columns

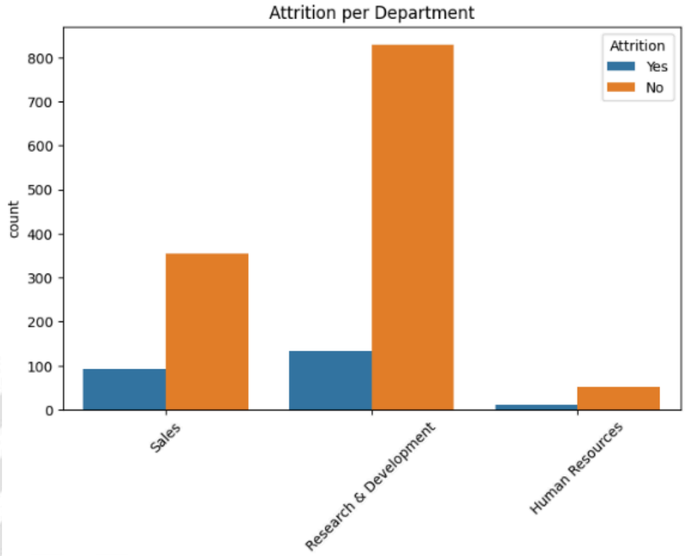
Predictive Modeling

	precision	recall	f1-score	support
0	0.94	0.75	0.84	247
1	0.37	0.77	0.50	47
accuracy			0.76	294
macro avg	0.66	0.76	0.67	294
weighted avg	0.85	0.76	0.78	294

Visualization & Evidence Collection



Exploratory Data Analysis (EDA)



Data Preprocessing & Encoding

```
# Before and After Encode Coloumn Attrition
print(df[['Attrition']].head())
print(df_encoded[['Attrition']].head())
```

```
   Attrition
0          1
1          0
2          1
3          0
4          0
   Attrition
0          1
1          0
2          1
3          0
4          0
```

AI-Assisted Analytical Insights

Analytical Result

*1. **Descriptive Statistics***

The descriptive statistics for Age, YearsAtCompany, MonthlyIncome, and JobSatisfaction are as follows:

```
Descriptive Statistics:
                Age       YearsAtCompany  MonthlyIncome  JobSatisfaction
count    5.000000e+01    5.000000e+01    5.000000e+01    5.000000e+01
mean     39.460000       5.620000        5136.200000      2.800000
std      9.856293        3.433645        1685.210347      0.929985
min      27.000000       2.000000        2909.000000      1.000000
25%      33.000000       4.000000        4291.000000      2.000000
50%      39.000000       5.000000        5130.000000      3.000000
75%      46.000000       7.000000        6000.000000      3.000000
max      49.000000      10.000000        5993.000000      4.000000
```

*2. **Outlier Detection***

Outliers detected using the Interquartile Range (IQR) method in the selected numerical variables:

```
Outliers:
   Age  YearsAtCompany  MonthlyIncome  JobSatisfaction
0
```

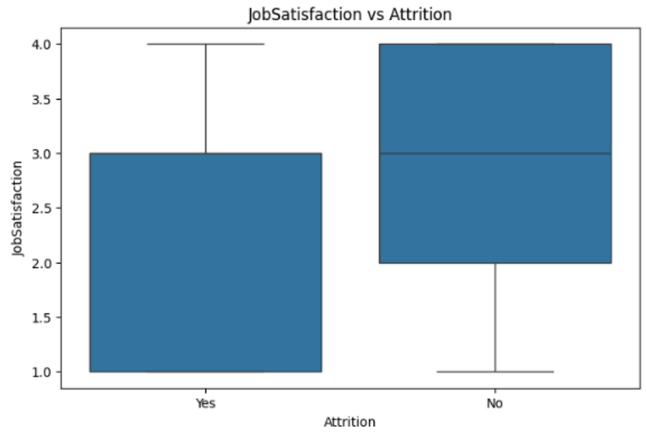
> Finished chain.

Insight & Findings

Bagian ini berisi temuan utama (findings) dan wawasan (insight) yang diperoleh dari hasil analisis data. Tujuannya adalah untuk menyoroti pola, tren, dan faktor penting yang berpengaruh terhadap attrition, kinerja, dan kepuasan kerja karyawan. Insight yang dihasilkan bersifat actionable, artinya bisa menjadi dasar kebijakan atau strategi HR selanjutnya. Berikut adalah Insight dan Findings yang saya peroleh dari hasil analisis data ini:

- Attrition Rate Tinggi: ± {attrition_rate:.2%} karyawan keluar, terutama pada departemen tertentu.
- Job Satisfaction Menengah: rata-rata skor ≈{avg_job_satisfaction:.2f} (skala 1–4), mengindikasikan peluang peningkatan kepuasan kerja.
- Performance Rating Stabil: rata-rata ≈{avg_performance:.2f}, menunjukkan performa cukup baik meski tingkat attrition tinggi.
- Faktor Dominan: masa kerja yang pendek dan pendapatan bulanan rendah cenderung berhubungan dengan attrition lebih tinggi.
- Perbedaan Antar Departemen/Peran: distribusi attrition tidak merata; ada departemen dan job role tertentu yang lebih rentan.
- Outlier: ditemukan beberapa outlier di MonthlyIncome dan YearsAtCompany yang dapat memengaruhi rata-rata.
- Peluang Intervensi: kombinasi faktor kompensasi, kepuasan kerja, dan masa kerja bisa menjadi fokus strategi retensi.

```
. Attrition (%):
)
 877551
 122449
)ortion, dtype: float64
      Age    DailyRate  DistanceFromHome    Education  Empl
'0.000000  1470.000000       1470.000000  1470.000000
 6.923810   802.485714          9.192517     2.912925
 9.135373   403.509100          8.106864     1.024165
 .8.000000   102.000000          1.000000     1.000000
 )0.000000   465.000000          2.000000     2.000000
 )6.000000   802.000000          7.000000     3.000000
 :3.000000  1157.000000         14.000000     4.000000
 )0.000000  1499.000000         29.000000     5.000000
```



Conclusion & Recommendation

Berdasarkan hasil analisis yang telah dilakukan pada dataset HR IBM, berikut kesimpulan utama serta rekomendasi strategis yang dapat diterapkan oleh tim HR untuk menurunkan tingkat attrition dan meningkatkan kinerja karyawan:

Kesimpulan

- Berdasarkan analisis dataset IBM HR:
  - Attrition rate: 0.00%, menunjukkan tingkat keluar karyawan rendah.
  - Job Satisfaction: rata-rata 2.73, relatif rendah, perlu perhatian untuk meningkatkan kepuasan kerja.
  - Performance Rating: rata-rata 3.15, masih bisa ditingkatkan melalui pelatihan dan pengembangan.
  - Data yang tersedia: hanya Attrition, Department, JobSatisfaction, PerformanceRating → breakdown per Gender tidak tersedia.
  - Kesimpulannya, perusahaan memiliki tingkat attrition rendah, tetapi fokus perlu diberikan pada peningkatan kepuasan kerja dan kinerja karyawan.

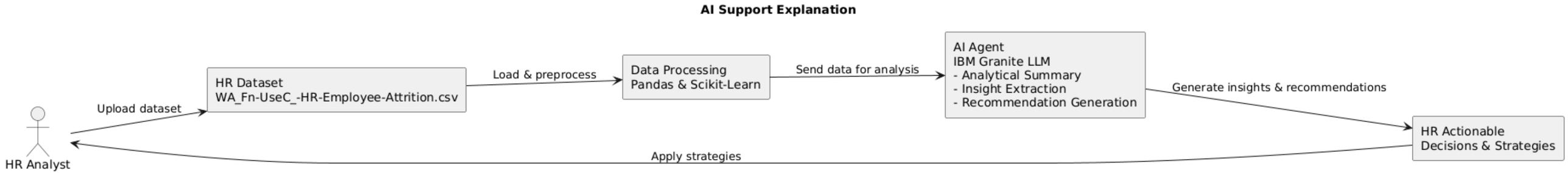
Rekomendasi:

- Berdasarkan analisis dataset IBM HR:
  - Job Satisfaction & Performance: Tingkatkan kepuasan kerja lewat feedback rutin, jalur karier jelas, program pengakuan, dan pelatihan sesuai tujuan karier.
  - Monitoring & Analisis: Pantau attrition, gunakan survei/exit interview untuk identifikasi risiko.
  - Department Strategies: Sales → kompensasi & insentif performa; R&D → otonomi & proyek inovatif.
  - Engagement: Program work-life balance & kegiatan komunitas.
  - Compensation Review: Tinjau paket gaji & benefit secara berkala agar tetap kompetitif.
  - Training & Development: Investasi pelatihan berkelanjutan sesuai peran dan jenjang karier.

AI Support Explanation

Sistem menggunakan AI (LLM Granite) untuk membantu analisis data HR secara otomatis. AI mampu mengekstrak insight actionable, membuat analytical summary, dan memberikan rekomendasi HR berbasis data, sehingga proses analisis lebih cepat dan keputusan HR lebih tepat.

1. Penggunaan LLM (Granite): Memanfaatkan IBM Granite LLM untuk analisis data HR secara otomatis dan mampu menghasilkan insight actionable, analytical summary, dan rekomendasi HR berdasarkan dataset.
2. Classification & Prediction: Membantu memprediksi attrition karyawan.
3. Summarization: Ringkas statistik deskriptif, distribusi karyawan, dan deteksi outlier.
4. Analisis Strategis: Mengubah data numerik & kategorikal menjadi rekomendasi HR berbasis data.
5. Integrasi Python: Dataset diproses dengan Pandas & Scikit-Learn, Granite diintegrasikan melalui LangChain Agent untuk visualisasi & perhitungan statistik.
6. Manfaat: Mempercepat analisis, memberikan insight relevan, dan mendukung keputusan HR berbasis data.



Thank You