PRAKTIKUM DATA MINING

PENERAPAN METODE FEATURE SELECTION PADA DATASET DATA SCIENCE JOB SALARIES DAN EARTHQUAKE

DISUSUN OLEH:
Kelompok 3

1.	Mujadid Choirus Surya	121450015
2.	Pramudya Wibowo	121450030
3.	A Rafi Paringgom Iwari	121450039
4.	Veni Zahara Kartika	121450075
5.	M. Faqih	121450120

PROGRAM STUDI SAINS DATA
JURUSAN SAINS
INSTITUT TEKNOLOGI SUMATERA
2023/2024

LAPORAN PRAKTIKUM DATA MINING

PENERAPAN METODE FEATURE SELECTION PADA DATASET DATA SCIENCE JOB SALARIES DAN EARTHQUAKE

Mujadid Choirus Surya[1)], Pramudya Wibowo[2)], A Rafi Paringgom Iwari[3)], Veni Zahara Kartika[4)], M.Faqih[5)]

Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera

Email : mujadid.121450015@student.itera.ac.id[1)], pramudya.121450030@student.itera.ac.id[2)], arafi.121450039@student.itera.ac.id[3)], veni.121450075@student.itera.ac.id[4)], mfaqih.121450120@student.itera.ac.id[5)]

Abstrak

Setiap dataset yang diolah pastinya memiliki tujuan yang berdasar seperti untuk menambah wawasan dan sebagai bahan pertimbangan terhadap pengaruh keputusan yang akan diambil kedepannya berdasarkan data dan fakta. Sementara, seringkali ditemukan ketidaksesuaian data dan jumlah data yang diperlukan untuk mewakili fitur-fitur pada dataset sangat besar dan tidak efektif. Pada praktikum kali ini akan dilakukan penerapan Metode feature selection pada dataset "Data Science Job Salaries" dan "Earthquake dataset". Hasil analisis laporan ini didapatkan fitur-fitur penting pada tiap dataset, dimana masing-masing dataset digunakan dua metode yang berbeda, yaitu filter dan decision tree, dengan tujuan menghasilkan analisis yang akurat dari suatu data dengan kualitas data yang baik juga.

Kata Kunci : Feature Selection, Data Preparation, Data Mining**.**

1. Pendahuluan

Pada beberapa dataset seringkali ditemukan ketidaksesuaian data dan jumlah data yang diperlukan untuk mewakili fitur-fitur pada dataset sangat besar dan tidak efektif mengakibatkan redudansi pada data yang mempengaruhi kinerja model yang dibangun dan interpretasi yang cukup sulit dilakukan. Hal ini didasarkan pada fitur dalam dataset yang tidak relevan dan hanya sebagian fitur yang memiliki korelasi hubungan dan signifikan terhadap variabel target dalam dataset. Setiap dataset yang diolah pastinya memiliki tujuan yang berdasar seperti untuk menambah wawasan dan sebagai pertimbangan terhadap pengaruh keputusan yang akan diambil kedepannya berdasarkan data dan fakta.

Berdasarkan permasalahan diatas, pada tahap data preprocessing dapat dilakukan penerapan *feature selection*. Metode *feature selection* merupakan sebuah metode pemilihan subset dari seluruh fitur yang ada dalam sebuah dataset untuk digunakan sebagai proses analisis data serta membuat model prediksi ataupun klasifikasi. *Feature selection* memiliki manfaat dalam meningkatkan kinerja model baik dalam hal prediksi ataupun klasifikasi, mengurangi overfitting, efisiensi komputasi dalam hal waktu dan memori, mengurangi dimensi yang terlalu besar, dan menghilangkan noise pada data[1].

Metode feature selection dapat diterapkan pada dataset “Data Science Job Salaries” dan “Earthquake dataset”. Metode feature selection yang akan diterapkan adalah metode filter dan metode embedded. Metode filter tidak melibatkan algoritma pembelajaran sehingga feature selection dilakukan berdasarkan informasi yang ada pada setiap fitur dan menggunakan peringkat dalam memilih subset fitur. Metode filter berbeda dengan metode embedded karena metode embedded menggunakan algoritma pembelajaran sehingga fitur dipilih berdasarkan keputusan algoritma pembelajaran.

Kedua metode tersebut diharapkan dapat meningkatkan performa model yang dibangun berdasarkan data pada dataset “Data Science Job Salaries” dan “Earthquake dataset” dengan mengidentifikasi fitur paling relevan dan berguna dalam hal analisis ataupun pembuatan model prediksi dan klasifikasi.

Laporan ini dibuat sebagai pertanggungjawaban atas praktikum mata kuliah data mining pertemuan ketiga. Referensi yang digunakan yakni pertama, modul 3 praktikum data mining yang menjelaskan konsep dasar terkait seleksi fitur atau feature selection. Kedua, materi perkuliahan minggu kelima data mining dan ketiga, jurnal “Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa” karangan I made Budi Adnyana.

2. Metode

2.1 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari dua sumber utama yaitu data Science Job Salaries Dan Earthquake

Data Science Job Salaries:

Data set ini berisi informasi tentang pekerjaan di bidang ilmu data, termasuk gaji, pengalaman, pendidikan, dan lokasi geografis dari individu yang bekerja di industri ilmu data.

Atribut Utama:

- Gaji (Salary): Informasi tentang gaji individu.
- Pengalaman (Experience): Lamanya pengalaman kerja individu.
- Pendidikan (Education): Tingkat pendidikan yang dimiliki individu (misalnya, sarjana, magister, doktor).
- Lokasi Geografis (Location): Lokasi geografis di mana individu tersebut bekerja.

Earthquake:

Data set ini berisi informasi tentang gempa bumi, termasuk lokasi, kedalaman, magnitudo, dan waktu kejadian gempa.

Atribut Utama:

- Lokasi (Location): Koordinat geografis (lintang dan bujur) tempat gempa terjadi.
- Kedalaman (Depth): Kedalaman dari episenter gempa bumi.
- Magnitudo (Magnitude): Ukuran besar gempa bumi.

- Waktu Kejadian (Time): Tanggal dan waktu ketika gempa terjadi.

2.2 Proses Pembersihan Data

Langkah pembersihan data dengan menggunakan df1.dropna() untuk menghapus baris dengan nilai kosong dari data set.

2.3 Pra Proses Data

Melakukan praproses data dengan langkah-langkah berikut:

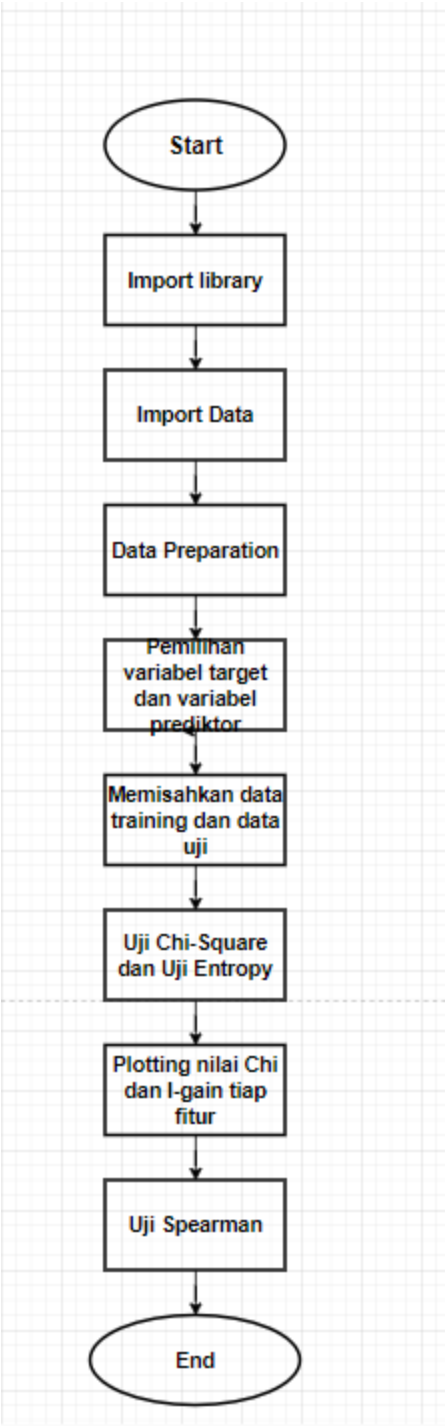
- Menghitung jumlah data pada variabel target 'alert' menggunakan df1['alert'].value_counts().
- Mengidentifikasi kolom dengan tipe data non-numerik (kategori) menggunakan df1.select_dtypes(exclude=[np.number]).columns dan menyimpannya dalam data_column_category.
- Melakukan label encoding pada kolom-kolom kategori tersebut menggunakan LabelEncoder dari scikit-learn untuk mengubah nilai kategori menjadi nilai numerik.

2.4 Metode Feature selection

Konsep dari feature selection yaitu mengidentifikasi variabel yang paling informatif dan relevan untuk dipertahankan, sementara variabel yang tidak relevan akan dihilangkan dalam analisis.dalam praktikum ini kita akan menggunakan metode filter dan embedded

1.) Filter Selection:

Metode Filter Selection adalah salah satu metode Feature Selection yang fokus pada peringkat fitur berdasarkan metrik tertentu, seperti Chi-Square atau Mutual Information. Dalam metode ini, setiap fitur dinilai secara independen berdasarkan matrik tersebut tanpa melibatkan algoritma pembelajaran. Fitur-fitur diberikan peringkat berdasarkan sejauh mana mereka dianggap relevan atau informatif terhadap tujuan analisis. Fitur-fitur yang memiliki peringkat tinggi dipertahankan, sementara yang memiliki peringkat rendah dapat dihapus dari dataset.Metode Filter Selection memberikan cara yang cepat dan efisien untuk mengurangi dimensi dataset dengan mempertahankan fitur-fitur yang paling penting dan relevan, berikut ditampilkan untuk flowchart program yang digunakan pada metode ini.



Gambar 2.4.1.1 Flowchart Program Filter

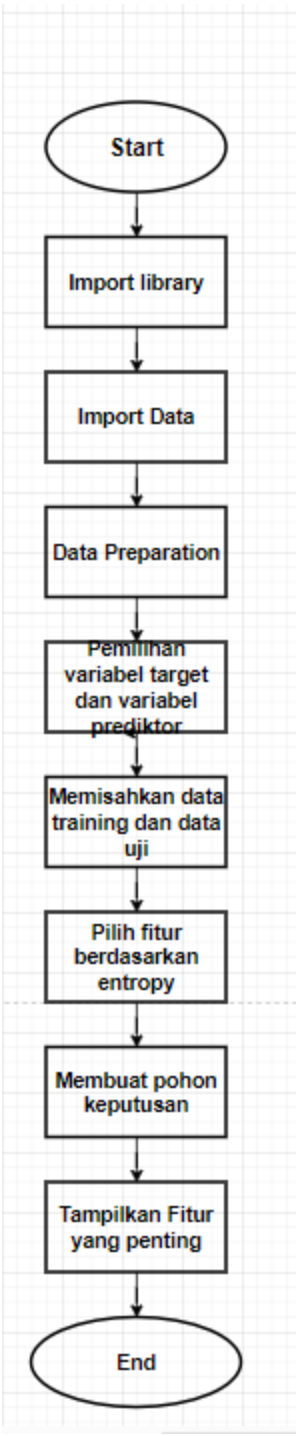
2.) Metode Embedded (Decision Tree):

Metode Embedded adalah metode Feature Selection yang menggunakan proses algoritma pembelajaran untuk memilih fitur yang paling informatif selama pelatihan model. Dalam

praktik ini, kami menggunakan Decision Tree sebagai algoritma pembelajaran yang secara alami melakukan seleksi fitur saat membangun pohon keputusan.

Ketika Decision Tree membangun pohon keputusan, setiap fitur diuji untuk melihat sejauh mana mereka dapat membagi data menjadi subset yang paling homogen dalam hal variabel target. Fitur-fitur yang memiliki kemampuan yang lebih baik untuk memisahkan data akan lebih tinggi dalam hierarki pohon keputusan, sementara fitur-fitur yang kurang informatif akan ditempatkan lebih rendah atau bahkan diabaikan.

Metode Embedded dengan Decision Tree dapat memberikan hasil yang baik karena algoritma ini secara alami mempertimbangkan relevansi fitur terhadap tugas klasifikasi atau regresi yang sedang dilakukan. Berikut ditampilkan flowchart program yang digunakan dalam metode ini.



Gambar 2.4.1.2 Flowchart D-Tree

3. Hasil

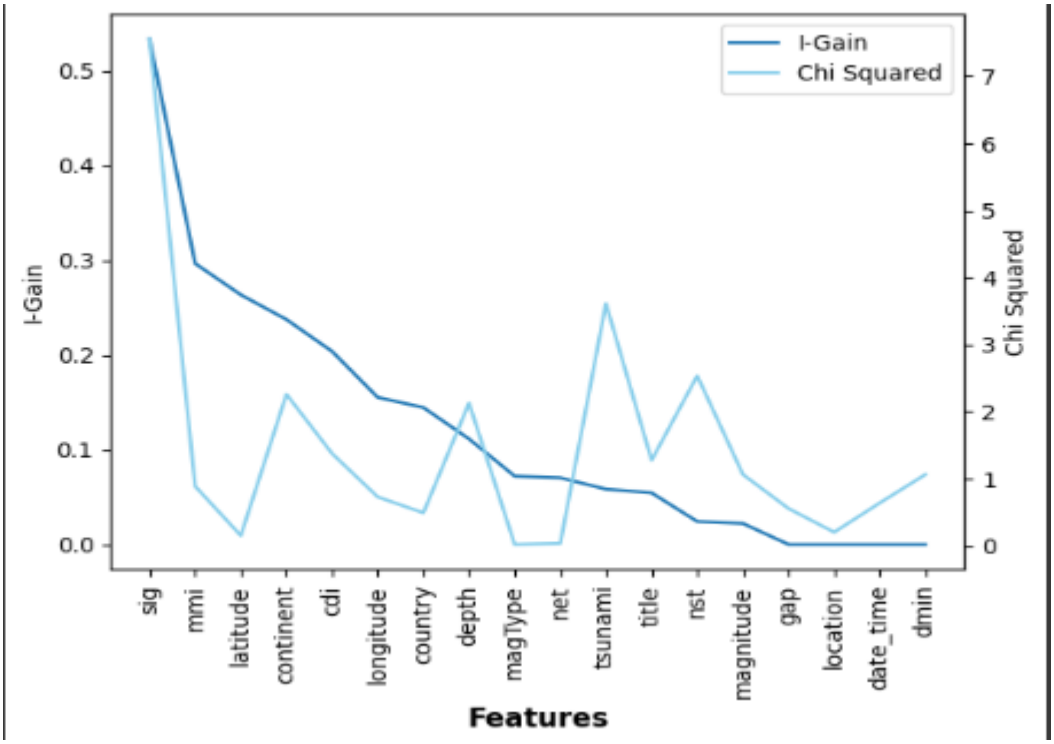
3.1. Data Preprocessing

Pada proses ini terdapat dua dataset yang akan dilakukan pengecekan terhadap noisenya, dataset pertama data mengenai gempa dan dataset kedua mengenai gaji pada pekerjaan yang berkaitan dengan Data Science. Proses ini dilakukan dengan cara mengecek missing value dan melakukan label encoding pada kedua dataset. Pada dataset mengenai gempa dilakukan penghapusan terhadap asumsi adanya missing value pada data, pemilihan variabel target dan kemudian dilakukan proses label encoding pada kolom dengan nilai kategorik dengan tujuan agar pemrosesan dapat dilakukan dengan baik. Kolom yang digunakan sebagai target adalah pada kolom alert dimana kolom ini memiliki niliunik yaitu, green, yellow, orange dan red, yang mewakili status untuk daerah tertentu. Beberapa kolom yang dilakukan transformasi pada data ini adalah title, magnitude, data-time, cdi, mmi, alert, tsunami, net, magtype, location continent dan country.

Untuk dataset kedua, Gaji pekerjaan pada bidang Data Science dilakukan hal yang sama data yang hilang dihapus pada dataset, kemudian kolom yang dipilih sebagai target adalah kolom company_size dengan nilai unik yaitu, M, L dan S, yang mewakili rata-rata pekerja yang bekerja pada suatu perusahaan dalam kurun waktu satu tahun. Beberapa kolom yang dilakukan transformasi untuk label encoding pada data ini antara lain, experice_level, employe_type, job_title, salary_currency, employee_residence, company_location, dan company_size.

3. 2. Pemilihan Fitur Menggunakan Metode Filter

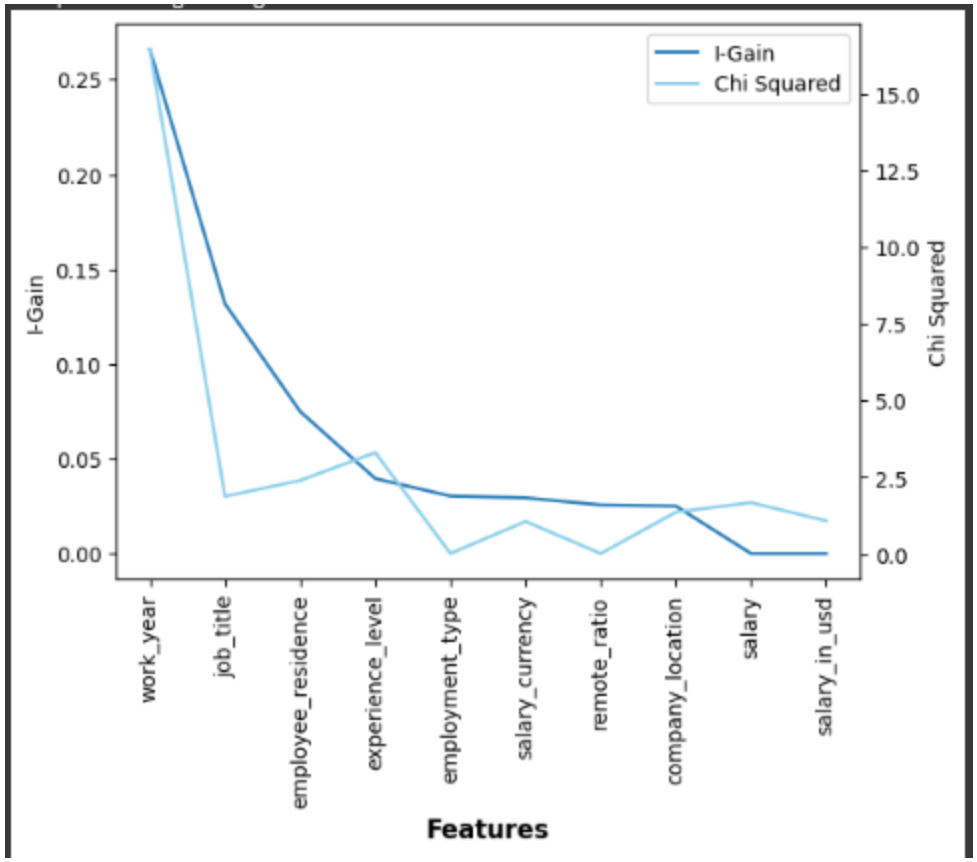
Data pertama yang berkaitan dengan gempa, untuk proses ini didapatkan hasil akhir mengenai variabel yang memiliki hubungan yang kuat untuk menjelaskan variabel target dengan tipe kategorik, serta informasi mengenai fitur-fitur yang memiliki pengaruh yang signifikan untuk menjelaskan variabel target pada tipe kategorik, dalam hal ini variabel target yang maksud ialah terdapat pada kolom alert, artinya model ingin melihat fitur apa saja yang memiliki pengaruh kuat dalam melakukan prediksi status kawasan gempa terhadap variabel-variabel lainnya yang disediakan pada data awal. Berdasarkan model yang didapatkan dari hasil pemodelan pada data latih dan data uji, didapatkan hasil sebagai berikut.



Gambar 3.2.1 Fitur Terbaik earthquake

Berdasarkan gambar 3.2.1. Didapatkan informasi bahwa fitur yang baik menurut pengujian chi-square yaitu sig, continent, depth, tsunami, title dan nst. Sedangkan fitur dengan nilai informasi yang paling penting berdasarkan score I-gain atau uji berdasarkan nilai entropy adalah sig, mmi, latitude continent, cdi dan longitude. Nilai korelasi spearman pada proses ini adalah sebesar 0.19 dan p-value 0.45, nilai ini dapat dengan jelas menyatakan bahwa kedua uji tersebut memiliki korelasi yang rendah.

Pada data gaji pekerjaan pada bidang Data Science, proses yang dilakukan tidak jauh berbeda, untuk mendapatkan fitur yang terbaik dalam pemodelan dilihat pada score untuk uji chi-square dan score pada I-Gain yang didapatkan melalui serangkaian proses pemodelan pada data latih dan data uji. Hasil yang didapatkan pada data ini akan dijelaskan melalui visualisasi berikut.

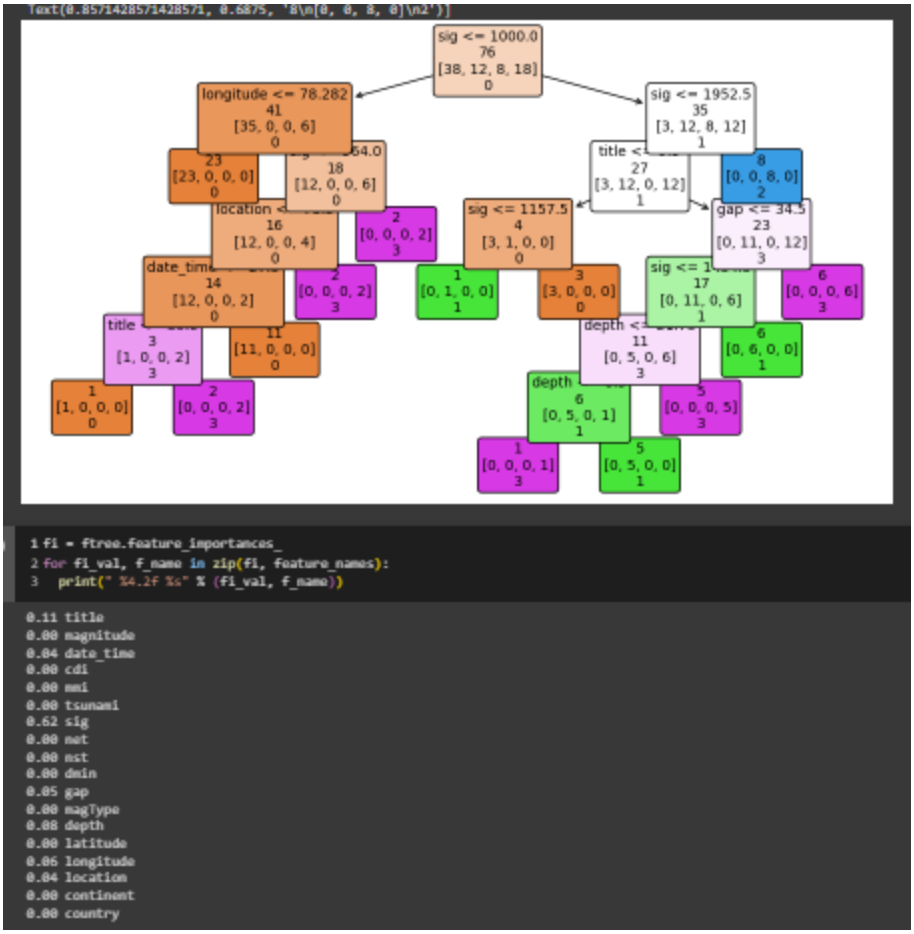


Gambar 3.2.2 Fitur terbaik salary

Pada gambar 3.2.2 didapatkan hasil yaitu fitur yang terbaik dalam menjelaskan variabel target berdasarkan uji chi ialah, work_year, experience_level dan salary. Sedangkan untuk fitur yang dinilai paling memberikan informasi penting yaitu, work_year, job_title dan employee_residence. Nilai korelasi spearman pada proses ini adalah 0. 5 dan p-value 0.07 , terbilang cukup baik namun belum cukup untuk mengatakan bahwa hasil pada kedua uji tersebut dapat dikatakan memiliki keterhubungan yang kuat

3. 3. Pemilihan Fitur Menggunakan Metode Decision Tree

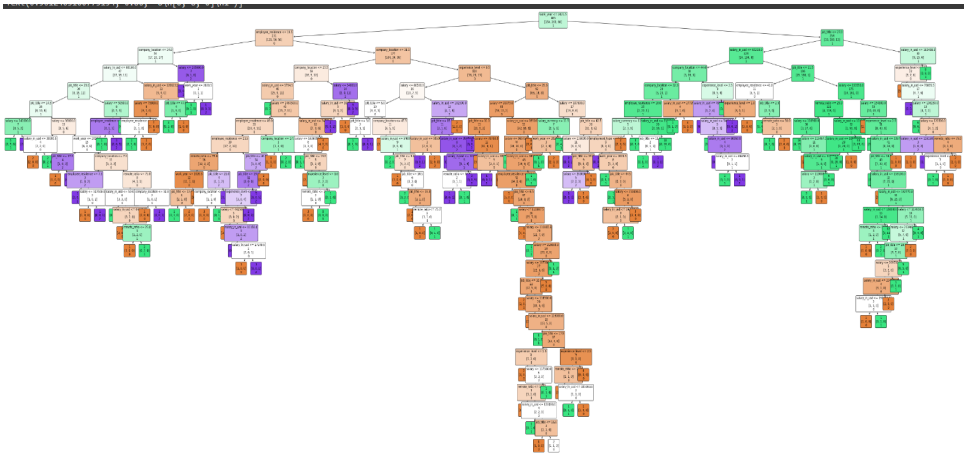
Metode decision tree digunakan untuk melakukan pengambilan keputusan yang lebih logis, sederhana dan terjabar dengan baik. Pada dataset yang berkaitan dengan gempa dilakukan juga proses decision tree dalam menentukan alert atau status pada kawasan tertentu. Proses decision untuk data ini dihasilkan pohon keputusan dengan cabang yang tidak terlalu banyak dan tidak rumit, hal ini dikarenakan pada proses pengklasfikaisan menggunakan fungsi DecisionTreeClassifier, dengan kriteria berdasarkan I-Gain, disini model dengan baik menentukan fitur yang digunakan untuk memprediksi target dengan menghilangkan fitur lain yang tidak signifikan pada entropinya, hasil pohon keputusan pada data ini sebagai berikut.



Gambar 3.3.1 Decision tree earthquake

Terlihat pada gambar 3.3.1. bahwa fitur yang penting dalam pengambilan keputusan pada metode ini ialah depth, sig, longitude, gap, location, data_time dan title, dengan nilai akurasi prediksi 0,65 yang terbilang cukup baik.

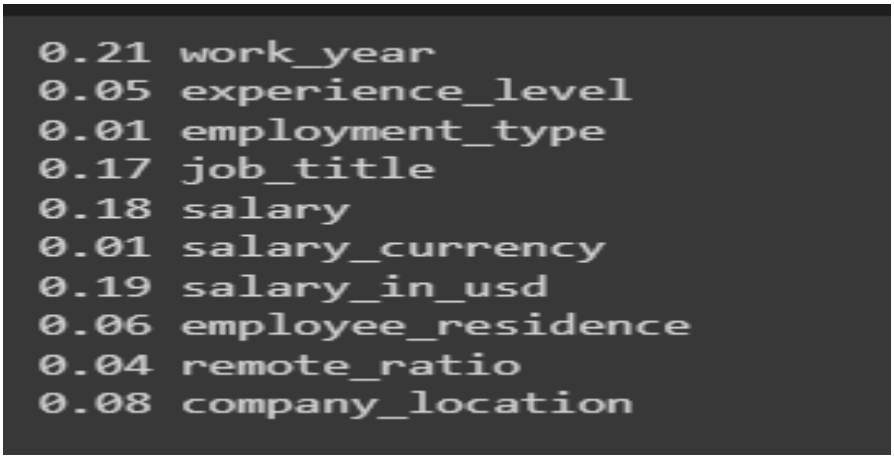
Pada data gaji pekerja pada bidang Data Science, dilakukan hal yang serupa pada data mengenai gempa, penentuan fitur yang digunakan dalam melakukan pembuatan pohon keputusan, dilihat dari nilai entropy atau uji I-Gain. Pada hasil akhir yang divisualisasikan untuk data ini akan sulit dilihat, karena data ini memiliki cukup banyak fitur dengan nilai entropy yang tinggi, untuk hasil visualisasi pohon keputusan pada data ini akan ditampilkan sebagai berikut.



,

Gambar 3.3.1. Decision Tree Salary

Terlihat pada gambar 3.3.1 visualisasi untuk pohon keputusan pada data ini sulit terlihat, diperkirakan terdapat sekitar 156 leaves pada pohon keputusan tersebut, hal ini dapat terjadi karena fitur yang dirasa penting menurut program cukup banyak, kemudian faktor lain dikarenakan keterbatasan media komputasi dalam menampilkan data yang cukup kompleks.



Gambar 3.3.2. Fitur terbaik D-Tree Salary

Berdasarkan gambar 3.3.2.terdapat cukup banyak fitur yang memiliki nilai entropy yang baik, hal inilah menjelaskan kenapa pada visualisasi pada gambar 3.3.1 cukup sulit untuk difahami. Beberapa fitur yang penting berdasarkan nilai entropy yaitu, work_year, salary_in_usd, salary, job_title dan fitur-fitur lainnya. Nilai akurasi model ini dalam melakukan prediksi terhadap target ialah sebesar 0.57 yang terbilang cukup baik.

4. Kesimpulan

Pada suatu data yang didapatkan dari berbagai sumber terkadang memiliki fitur-fitur yang tidak terlalu penting untuk dimasukan pada proses analisis, hal ini akan

membuat program memakan waktu yang lama dan memakan cukup banyak, untuk mengatasi masalah ini dapat dilakukan dengan cara memiliki fitur terpenting dalam data, yaitu dengan dapat dengan metode filter ataupun decision tree.

Pada hasil analisis laporan ini didapatkan fitur-fitur penting pada tiap dataset, dimana masing-masing dataset digunakan dua metode yang berbeda, yaitu filter dan decision tree. Metode pemilihan fitur dengan filter akan menggunakan konsep uji chi-square dan uji entropy, yang kemudian dilihat hubungan keduannya, sedangkan untuk metode decision tree lebih menekankan pada uji entropy tiap fiturnya yang kemudian dibuatkan model pohon keputusannya dan terakhir ditentukan fitur-fitur yang dapat menjelaskan variabel target dengan cukup baik.

Referensi

[1] I. M. B. Adnyana, "Penerapan Feature Selection untuk Prediksi Lama Studi," JURNAL SISTEM DAN INFORMATIKA, vol. 13, pp. 72-76, 2019.

Lampiran

1. **Link Tautan Program :**
   https://colab.research.google.com/drive/1SG7ss1gOW57wYLOdjNvNf07VUZtPB0F1?usp=sharing

2. **Link YouTube:**
   https://youtu.be/yCcrljG6R4k

3. **Link Dataset :**
   https://www.kaggle.com/datasets/warcoder/earthquake-dataset
   https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries