COMP6579001 Big Data Processing - Group Assignment

Anggota kelompok:

2501969630 - Albert Juan

2501988056 - Faiz Arya Rabbani

2501970525 - Juan Mike Volney Sandjaja

2501983906 - Nadhif Fathoni Hafiz

2501985760 - Rafi Muhammad Adyatma

bab 1 : latar belakang tujuan dan manfaat.

bab2 : Kerangka metedologi dari analysis

bab 4: Detail dari pengerjaan analysis yang dilakuan

bab 5: kesimpulan

Dataset:

https://www.kaggle.com/datasets/lodetomasi1995/income-classification

Bab 1: Latar Belakang, Tujuan, dan Manfaat

Latar Belakang:
Klasifikasi pendapatan adalah proses mengelompokkan individu berdasarkan tingkat pendapatan mereka. Klasifikasi ini memiliki banyak aplikasi penting dalam berbagai domain, seperti pemodelan pasar, penetapan harga, pemilihan kredit, dan kebijakan sosial. Dalam era big data, analisis data yang luas dapat digunakan untuk mengembangkan model prediksi yang akurat untuk mengklasifikasikan pendapatan individu berdasarkan berbagai fitur dan faktor yang relevan.

Tujuan:

1. Mengembangkan model klasifikasi pendapatan berbasis big data yang akurat.
2. Memahami dan mengidentifikasi faktor-faktor yang mempengaruhi tingkat pendapatan individu.
3. Mengklasifikasikan pendapatan individu dengan akurat berdasarkan fitur-fitur yang relevan

4. Mendukung pengambilan keputusan bisnis yang lebih baik dalam hal penetapan harga, penentuan target pasar, dan alokasi sumber daya.
5. Membantu dalam pengembangan kebijakan sosial dan kebijakan pemerintah yang bertujuan mengurangi kesenjangan pendapatan dan meningkatkan pemberdayaan ekonomi.

Manfaat:
Adapun manfaat dari penelitian ini adalah:

1. Meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi pendapatan individu.
2. Mengembangkan model prediksi yang dapat digunakan untuk mengklasifikasikan pendapatan individu secara efektif.
3. Memungkinkan analisis yang lebih baik dalam pengambilan keputusan bisnis, penetapan kebijakan, dan pemodelan pasar.
4. Meningkatkan keakuratan pemilihan kredit dan evaluasi risiko.
5. Mendukung pengembangan kebijakan sosial dan kebijakan pemerintah yang lebih efektif dalam mengurangi kesenjangan pendapatan.

Bab 2: Kerangka Metodologi dari Analisis

Bab ini akan menjelaskan kerangka metodologi yang digunakan dalam analisis klasifikasi pendapatan berbasis big data. Beberapa aspek yang akan dibahas termasuk:

1. Pengumpulan dan preprocessing data: Metode pengumpulan data yang digunakan, sumber data yang dianalisis, dan teknik preprocessing data untuk mempersiapkan data untuk analisis.
2. Seleksi fitur: Proses pemilihan fitur yang paling relevan dan signifikan untuk klasifikasi pendapatan.
3. Model klasifikasi: Pemilihan model atau algoritma yang paling sesuai untuk mengklasifikasikan pendapatan. Ini dapat mencakup teknik pembelajaran mesin seperti LogisticRegression
4. Pelatihan dan evaluasi model: Pembagian data menjadi set pelatihan dan pengujian, melatih model menggunakan set pelatihan, dan menguji kinerja model menggunakan set pengujian. Evaluasi hasil dan interpretasi dari model yang dikembangkan.
5. Visualisasi data

4. Hasil analisis

Dengan menggunakan Logistic Regression dan BinaryClassificationEvaluator kita mendapatkan accuracy

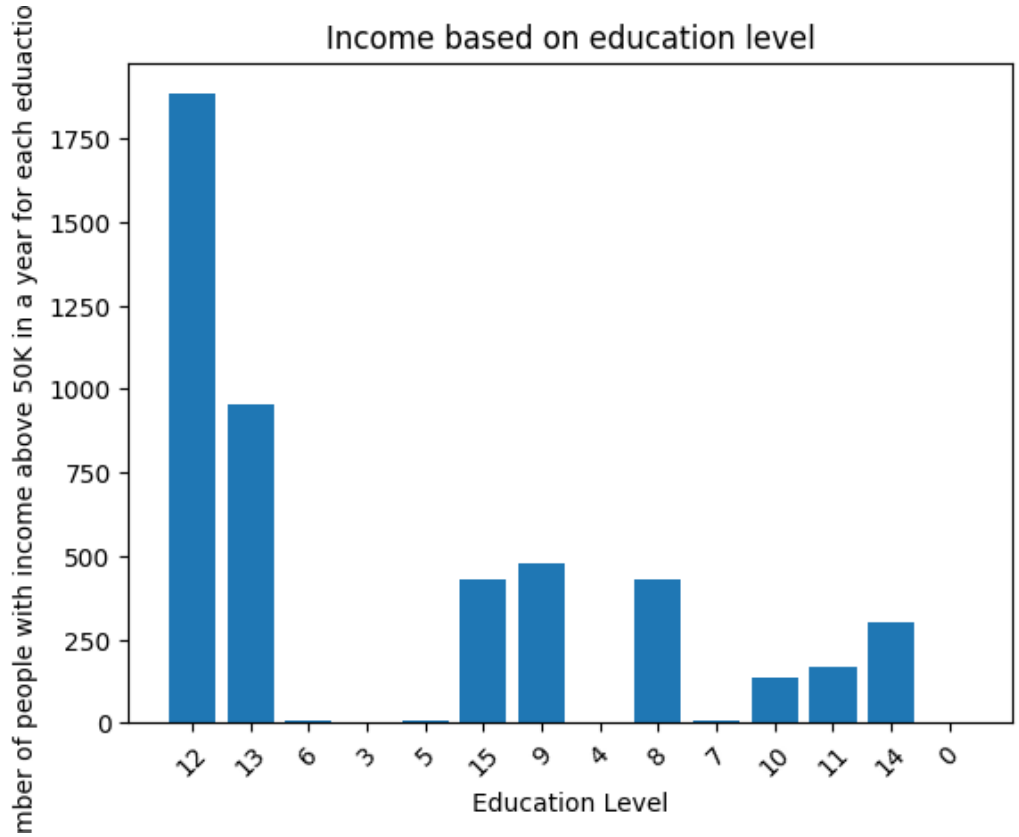
Waktu bekerja yang memiliki income terbesar adalah yang bekerja 40- 60jam setiap minggu

Edukasi yang paling bisa mendapatkan income adalah jejang selesai pendidikan kuliah dan ahli bidang

Negara yang paling menghasilkan capital terbanyak adalah Amerika

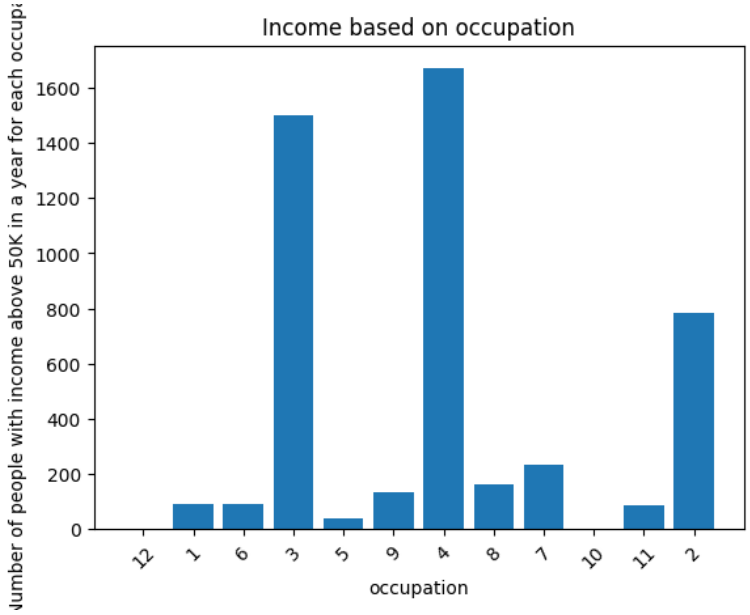
Pekerjaan yang paling menghasilkan income adalah Profesor, Manager, dan Salesman

Berikut adalah visualisasinya:



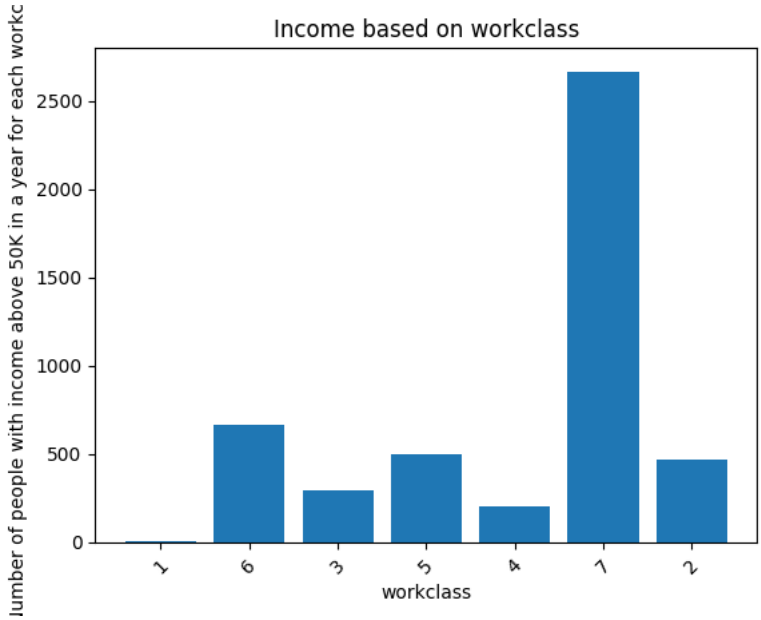
Gambar 1.1

Berdasarkan *education level*-nya, dilihat dari visualisasi data grafik bar gambar 1.1. Kelompok *education-level* yang diklasifikasikan memiliki pendapatan 50.000 terbanyak dari kalangan *Bachelors* atau lulusan sarjana S1.



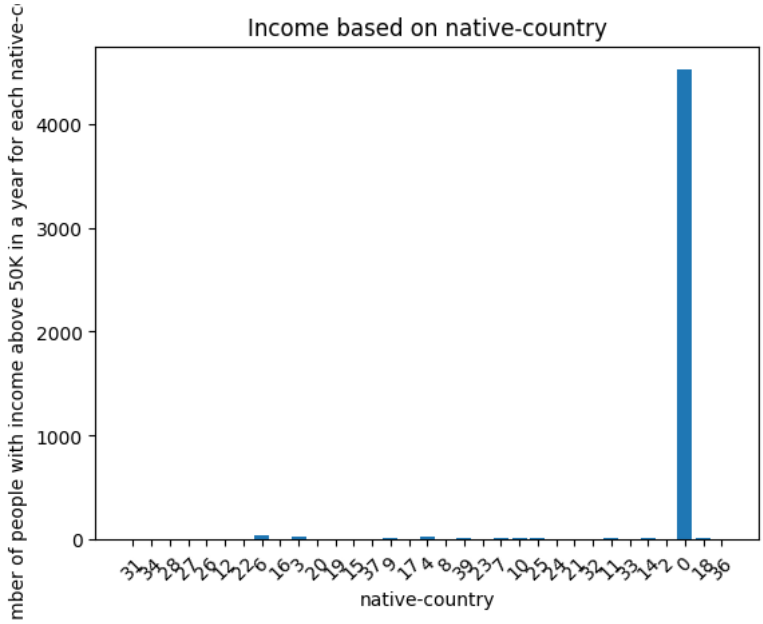
Gambar 1.2

Berdasarkan *education occupation*, dilihat dari visualisasi data grafik bar gambar 1.2. Kelompok *occupation* yang diklasifikasikan memiliki pendapatan 50.000 terbanyak dari kalangan *Prof-specialty* kemudian diikuti oleh Exec-managerial di urutan kedua, dan Sales diurutan ketiga.



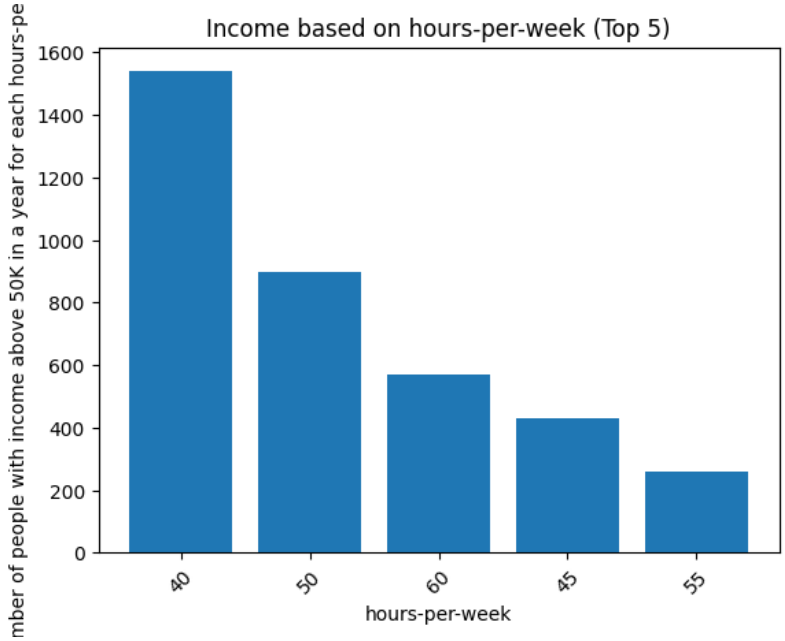
Gambar 1.3

Berdasarkan *workclass*-nya, dilihat dari visualisasi data grafik bar gambar 1.1. Kelompok *workclass* yang diklasifikasikan memiliki pendapatan 50.000 terbanyak dari kalangan *Private* kemudian diikuti oleh Self-emp-not-inc pada nomor dua, dan Self-emp-inc pada nomor tiga.



Gambar 1.4

Berdasarkan *native-country*-nya, dilihat dari visualisasi data grafik bar gambar 1.1. Negara asal yang diklasifikasikan memiliki pendapatan 50.000 terbanyak dari United States (US), dengan perbedaan yang sangat signifikan.



Gambar 1.5

Berdasarkan *hours-per-week*-nya, dilihat dari visualisasi data grafik bar gambar 1.5. *hours-per-week* yang diklasifikasikan memiliki pendapatan 50.000 terbanyak dari kelompok yang bekerja 40 jam per minggu, diikuti dengan kelompok yang bekerja 50 jam perminggu pada nomor 2, dan 60 jam per minggu pada nomor 3.

5.Conclusion

Secara keseluruhan, analisis dataset ini menunjukkan potensi besar dari Big Data dalam membantu pemodelan dan prediksi klasifikasi pendapatan. Dengan penggunaan yang tepat, Big Data dapat menjadi alat yang kuat dalam mendukung pengambilan keputusan yang lebih baik dan pengembangan solusi yang lebih efektif dalam domain ini.