Big Data Kelompok

KELOMPOK 7

Income Classification

Kaggle

The data set contains
3999 rows and 14 cols.

Latar Belakang

Income Classification adalah proses mengelompokkan individu berdasarkan tingkat pendapatan mereka. 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

Manfaat

1

Meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi pendapatan individu.

2

Memungkinkan analisis yang lebih baik dalam pengambilan keputusan bisnis,

3

Meningkatkan keakuratan pemilihan kredit dan evaluasi risiko.

Method

Data Gathering
Load & Transfrom
Modeling
Google colab
Visual
Analysis

Select features

- mengurangi dimensi data,
- meningkatkan kualitas model,
- mempercepat waktu pemrosesan,
- mengurangi risiko overfitting,

workclass ,education , marital-status, occupation, relationship , captal-gain ,capital-loss,hours-per-week, native-country , income

Modeling

```
model = LogisticRegression(featuresCol = "Vector Features", labelCol = " income", maxIter = 10).fit(income)
prediction = model.transform(income)
```

Memperoleh pemahaman yang lebih baik tentang hubungan dan pola yang ada di dalamnya.

pemahaman data

Kita dapat menggunakan teknik statistika untuk mengidentifikasi hubungan dan korelasi antara variabel-variabel tertentu

analisis

Analysis memungkinkan kita untuk membuat prediksi dan estimasi yang lebih akurat tentang hasil masa depan

prediksi

Visual -education level (1/5)

```
when(df[" education"] == " Preschool", 0).
when(df[" education"] == " 1st-4th", 1).
when(df[" education"] == " 5th-6th", 2).
when(df[" education"] == " 7th-8th", 3).
when(df[" education"] == " 9th", 4).
when(df[" education"] == " 10th", 5).
when(df[" education"] == " 11th", 6).
when(df[" education"] == " 12th", 7).
when(df[" education"] == " HS-grad", 8).
when(df[" education"] == " Some-college", 9).
when(df[" education"] == " Assoc-voc", 10).
when(df[" education"] == " Assoc-acdm", 11).
when(df[" education"] == " Bachelors", 12).
when(df[" education"] == " Masters", 13).
when(df[" education"] == " Doctorate", 14).
when(df[" education"] == " Prof-school", 15))
```

Visual -native country (2/5)

```
native-country"] == " United-States", 0).
native-country"] == " Cambodia", 1).
native-country"] == " Puerto-Rico", 2).
native-country"] == " Canada", 3).
native-country"] == " Germany", 4).
native-country"] == " Outlying-US(Guam-USVI-etc)", 5).
native-country"] == " India", 6).
native-country"] == " Japan", 7).
native-country"] == " Greece", 8).
native-country"] == " China", 9).
native-country"] == " Cuba", 10).
native-country"] == " Iran", 11).
native-country"] == " Honduras", 12).
native-country"] == " Philipines", 13).
native-country"] == " Italy", 14).
native-country"] == " Poland", 15).
native-country"] == " Jamaica", 16).
native-country"] == " Vietnam", 17).
native-country"] == " Mexico", 18).
native-country"] == " Portugal", 19).
```

Visual -workclass (3/5)

```
["workclass"] == " Never-worked", 0).
["workclass"] == " Without-pay", 1).
["workclass"] == " Local-gov", 2).
["workclass"] == " State-gov", 3).
["workclass"] == " Federal-gov", 4).
["workclass"] == " Self-emp-inc", 5).
["workclass"] == " Self-emp-not-inc", 6)
["workclass"] == " Private", 7))
```

Visual -occupation (4/5)

```
"] == " Tech-support, Craft-repair", 0).
"] == " Other-service", 1).
"] == " Sales", 2).
"] == " Exec-managerial", 3).
"] == " Prof-specialty", 4).
"] == " Handlers-cleaners", 5).
"] == " Machine-op-inspct", 6).
"] == " Adm-clerical", 7).
"] == " Farming-fishing", 8).
"] == " Transport-moving", 9).
"] == " Priv-house-serv", 10).
"] == " Protective-serv", 11).
"] == " Armed-Forces", 12))
```

Visual -hours-per-week (5/5)

Analysis

```python
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction", labelCol=" income")
accuracy = evaluator.evaluate(prediction)

print(accuracy)
```

```
0.8870010778360999
```

0.8870010778360999

Conclusion

- 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

THANK YOU

HTTPS://WWW.KAGGLE.COM/DATASETS/LODETOMASI1995/INCOME-CLASSIFICATION