Pandasquad

- Annisa Ulfa Rahma
- Edwin Reyhan D.R.R.
- M Fikri Setiawan
- Muhammad Rizqi

Dokumen Laporan Final Project

Latar Belakang Masalah

- *Income* atau pendapatan bersifat pribadi.

- *Income* digunakan sebagai tolak ukur untuk penentuan lolos atau tidak, misal untuk pinjaman, cicil rumah, dan pembuatan/pengembangan produk baru.

Studi Kasus

- Sebuah perusahaan produksi barang yang mempunyai proses bisnis mengembangkan produknya butuh tool yang bisa membantu memprediksi pendapatan orang (jika belum menemukan data pendapatan) di suatu tempat dimana produk tersebut akan dikembangkan untuk menentukan harga produk akan diterima dengan pertimbangan pendapatan orang.

- Sebuah badan statistik mempunyai tugas : melaksanakan tugas pemerintahan dibidang statistik sesuai peraturan perundang-undangan, ingin memprediksi pendapatan warga di suatu tempat.

Latar Belakang Masalah

Data

- *Income* atau pendapatan bersifat pribadi. => Tertuang pada UU No. 14 Tahun 2008

Headline Berita

Pre-processing

- Baris duplicated kami drop. Namun kami juga akan mencoba bagaimana jika duplicated tidak didrop.

- Missing value berupa string “?” yang untuk saat ini kami jadikan “other”, karena rownya lumayan banyak. Namun berdasarkan hasil mentoring, kami akan mempertimbangkan untuk mendrop row yang mempunyai value “?” karena jika dibandingkan jumlah row yang mempunyai “?” dengan total row, row tersebut bisa dipertimbangkan untuk didrop. Namun juga sebelum kami drop, kami akan memastikan terlebih dulu kenapa value itu berupa “?”. Kami juga akan mencoba bagaimana jika “?” diisi modus.

Pre-processing

- Feature dengan outliers seperti age, education.num, hours.per.week kami transformasi menggunakan log karena banyak machine learning model yang asumsinya adalah data yang digunakan berdistribusi normal. Kami juga akan mencoba scenario lain selain log untuk membuat data yang kami punya berdistribusi normal.

- Feature workclass, marital.status, occupation untuk saat ini kami grouping karena masing-masing feature memiliki unik value yang banyak dan menurut kami bisa di-grouping. Namun berdasarkan hasil mentoring yang dilakukan, kami mempertimbangkan untuk tidak melakukan grouping pada features tersebut. Karena menurut mentor, performa model machine learning akan baik-baik saja jika features tersebut menghasilkan kolom yang banyak jika diencode.

Pre-processing

- Feature workclass, marital.status, occupation, relationship, race, sex, dan income kami encode dengan metode one hot encoding karena data features ini berupa categorical (nominal) dan kami encode agar siap untuk digunakan oleh machine learning model.

- Feature education kami hapus karena ada feature education.num (ordinal) yang menggambarkan berapa lama Pendidikan atau tingkatan Pendidikan feature education. Hal ini sama seperti kita melakukan label encoding terhadap feature education.

- Semua kolom (features) selain target income kami scale menggunakan StandardScaler karena MinMaxScaler sensitive terhadap outliers. Feature scaling butuh dilakukan agar features yang memiliki skala besar tidak mendominasi features yang skala kecil.

Pre-processing

- Kami mencoba undersampling dan oversampling untuk class imbalanced.

- Undersampling atau oversampling perlu dilakukan karena jika terjadi class imbalance, score accurary yang dihasilkan model machine learning akan misleading.

EDA,Insights and visualization

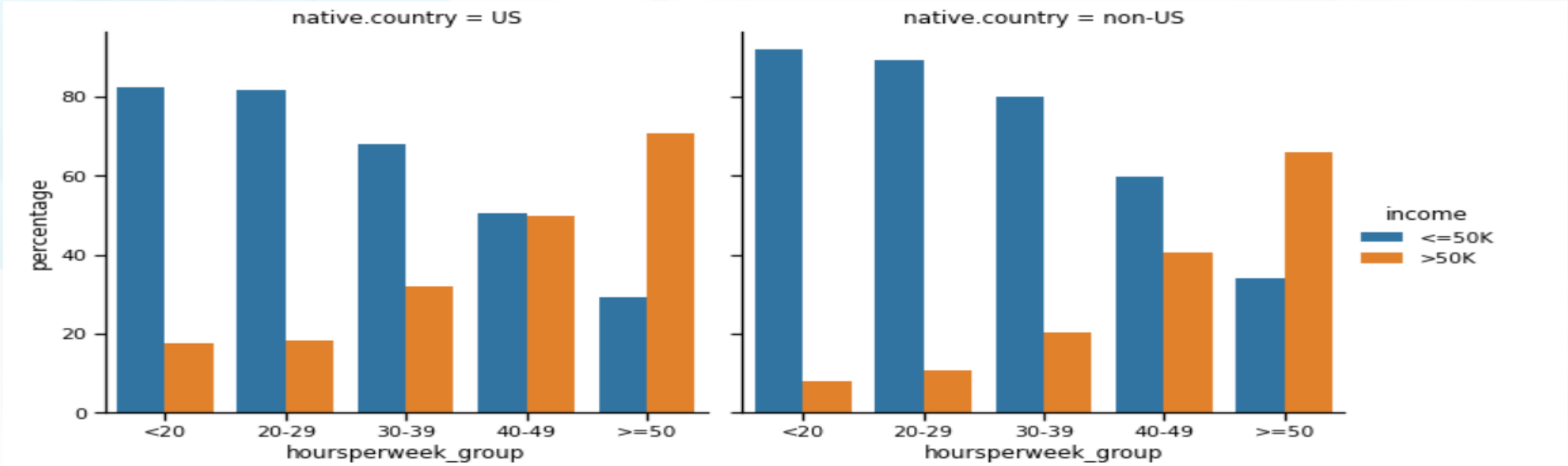
- Sebagian besar dari masing-masing golongan penduduk yang bekerja di pemerintahan dan self-employee mendapatkan gaji di atas 50K. Sedangkan untuk golongan private sebagian besar mereka bergaji dibawah atau sama dengan 50K.
- Pendidikan yang sebagian besar penduduk yang memilikinya mendapatkan gaji di atas 50K adalah penduduk yang berpendidikan Assoc-voc sampai Doctorate.
- Rata-rata umur penduduk yang bergaji lebih dari 50K adalah 44 tahun dan rata-rata umur penduduk yang self-employee dan yang bekerja di pemerintahan adalah 43 – 46 tahun, maka dari itu rata-rata penduduk yang bergaji lebih dari 50K adalah self employee dan yang bekerja di pemerintahan .

EDA,Insights and visualization

Insights and visualization

Insights:

Pada range jam kerja per minggu 40-49 jam, penduduk asli US lebih banyak yang bergaji lebih dari 50K, sedangkan penduduk bukan asli US dengan range jam kerja yang sama sebagian besar bergaji kurang dari atau sama dengan 50K



Modelling Experiments

Algoritma yang sudah dicoba adalah Support Vector Machine (SVM), Decision Tree, Logistic Regression, K-nearest neighbors (KNN), Ridge Classifier, SGDClassifier, Gaussian process classification (GPC), Naïve bayes, dan Random Forest

Di antara algoritma tersebut algoritma yang mendapatkan nilai accuracy tertinggi dengan default setting adalah Logistic Regression dengan nilai sebesar 82%

Dengan hanya menggunakan feature age, education.num, marital.status, relationship, capital.gain, capital.loss, dan hours.per.week saja, accuracy Logistic Regression masih bernilai 82%

Modelling Experiments (1-3 slide)

Kami menentukan model yang dipakai dengan melihat nilai evaluasi model klasifikasi seperti classification_report dan confusion_matrix.

Secara data balance, data kami sudah diundersampling yang mana membuat data kami memiliki class yang balance.

Maka dari itu, nilai accuracy akan dipakai sebagai nilai utama yang akan menentukan model mana yang akan dipakai. Setelah itu model akan dipilih berdasarkan nilai recall dan precision.

Executive Summary & Recommendation

- Semakin tinggi jenjang pendidikan semakin tinggi pula penduduk bisa mendapatkan gaji di atas 50>
- Semakin tinggi jenjang pendidikan semakin tinggi pula penduduk bisa mendapatkan pekerjaan yang bagus, yang mana pekerjaan itu menghasilkan pendapatan di atas 50K

Recommendation:

- Untuk meninggkatkan GDP Amerika Serikat, kita dapat melakukan program beasiswa untuk meningkatkan jumlah penduduk berpendidikan tinggi, agar mereka bisa mendapatkan pekerjaan yang bergaji tinggi.

Pembagian Tugas

Penulisan laporan:

- Annisa Ulfa
- Edwin Reyhan D.R.R.
- Muhammad Fikri
- Muhammad Rizky

Pembuatan slide:

Konten: Annisa, Rizky, Edwin

Template: Muhammad Fikri

Presentasi: Annisa, Edwin, Rizky