Komparasi Model Algoritma *Logistic Regression, k-Nearest Neighbor, dan Decision Tree* Terhadap Nilai *Accuracy* Pada Data "Income Classification"*

Rizky Maulana. Department of Engineering Technikal and UPN Veteran Jakarta, rizkimaulana9145@gmail.com

Mohammad Arif Mustofa, Department of Physics, Universitas Sebelas Maret, moh.arif11022000@gmail.com

Setya Nugraha, Department of Engineering Informatics, Dian Nuswantoro University setyannugrahaa@gmail.com

Abstract*— Dengan menggunakan klasifikasi pada data income, penelitian ini mengeksplorasi apakah pendapatan pekerejaan di suatu negara merupakan penentu kelas pendapatan suatu negara. Studi ini juga menguji adanya hubungan variable seperti umur, workclass, tingkat pendidikan, status, ras, jenis kelamin, capital gain, jam kerja dalam sepekan, dan kewarganegaraan. Pemodelan klasifikasi juga dilakukan dengan model Decision Tree, KNN, dan Logistic Regression. Diperoleh hasil akuasi setelah dilakukan standardisasi pada evaluasi model akurasi algoritma decisiontree 0.814 menjadi 0.814, algoritma KNN 0.769 menjadi 0.834, dan logisticregression 0.792 menjadi 0.851.

I. PENDAHULUAN

Klasifikasi merupakan penyusunan bersistem dalam kelompok atau golongan menurut kaidah yang ditetapkan. Klasifikasi digunakan dalam berbagai sektor kehidupan. Salah satunya ekonomi. Program Bantuan untuk masyarakat kalangan bawah merupakan program yang membantu masyarakat kalangan bawah untuk tetap memenuhi kebutuhan hariannya. Untuk mengelompokkan golongan pendapatan masyarakat tersebut maka digunakan 7 kriteria, yaitu *age, finalweight, education number, capital gain, capital loss, hours per week.*. oleh karena itu dibutuhkan suatu alat analisis yang mampu menganalisis dengan baik data yang sangat besar tersebut.

Terdapat beberapa langkah dalam pengolahan data sebelum melakukan data mining, yakni membersihkan data dari noise dan data yang tidak konsisten, mengkombinasikan kembali data-data yang telah bersih, maka kita akan memiliki database yang baru, selanjutnya data dilihat kembali apakah membutuhkan suatu transformasi ataukah tidak, barulah setelah itu data dapat diolah. Klasifikasi merupakan pengelompokkan yang sistematias pada sejumlah objek, gagasan, buku atau benda-benda lain ke dalam kelas atau golongan tertentu berdasarkan ciri-ciri yang sama.

Permasalahan utama dalam upaya pengurangan kemiskinan saat ini terkait dengan adanya fakta bahwa pertumbuhan ekonomi tidak tersebar secara merata di seluruh wilayah Indonesia, ini dibuktikan dengan tingginya perbedaan pendapatan antar daerah. Selain itu kemiskinan juga merupakan sebuah hubungan sebab akibat (kausalitas melingkar) artinya tingkat kemiskinan yang tinggi terjadi karena rendahnya pendapatan perkapita, pendapatan perkapita yang rendah terjadi karena investasi perkapita yang juga rendah [2].

Diharapkan dari penelitian yang dilakukan terhadap sampel data penduduk miskin tersebut dapat diperoleh suatu informasi yang bisa membantu pihak kecamatan untuk merancang strategi dalam meningkatkan kesejahteraan masyarakat.

Dataset income classiffication merupakan data yang diambil dari masyarakat suatu negara. Data ini dibuat untuk melihat kesejahteraan masyaraktnya yang dilihat dari penghasilan perbulannya, dimana kategori masyarakat yang menengah atas memiliki penghasilan diatas 50k dollar perbulan

II. LANDASAN TEORI

A. *Implementasi dalam "Income Classification"*

Menurut Purwanto dan Sulistyastuti, implementasi adalah kegiatan untuk mendistribusikan keluaran kebijakan (to deliver policy output) yang dilakukan oleh para implementer

*Research supported by ABC Foundation.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

kepada kelompok sasaran(target group) sebagai upaya untuk mewujudkan tujuan kebijakan. Tujuan kebijaka diharapkan akan muncul manakala policy output dapat diterima dan dimanfaatkan dengan baik oleh kelompok sasaran sehingga dalam jangka panjang hasil kebijakan akan mampu diwujudkan. Implementasi merupakan salah satu tahapan dari serangkai proses atau siklus suatu kebijakan.

B. Data Mining

Data mining dalam istilah sederhana adalah penemuan pola yang berguna dalam pengolahan data, data mining juga disebut sebagai ilmu pengetahuan, machine learning, dan analis prediksi[5]

B. Logistic Regression

Logistic regression memperluas gagasan pada beberapa linear regression untuk situasi dimana variabel saling ketergantungan , y adalah diskrit. Pada logistic regression (Homer & Lemeshow 2000) tidak ada asumsi membuat tentang distribusi pada variabel yang independen. Pemberian set pada sampel N (xi, yi) dengan xi∈Rd, dimana d adalah nomor dimensi dan label kelas yang sesuai yi∈ {1, 2, … , K}. kemudian , logistic regression mencoba untuk memperkirakan probabilitas posterior pada sampel x baru seperti [5]:

$$p(y=k \mid x)=\frac{\exp(-(w_{k_0}+{w_k}^{\mathrm{T}}x))}{1+\sum_{l=1}^{K-1}\exp(-(w_{l_0}+{w_l}^{\mathrm{T}}x))},\ k=1,\ldots,K-1,$$

C. Decission Tree

Mempelajari pohon keputusan dari record pada kelas yang diberi label. Decision tree adalah sebuah flowchart yang seperti struktur pohon, dimana setiap node internal (node tidak berdaun) menandakan sebuah tes pada atribut, setiap branch merepresentasikan hasil dati tes tersebut, dan setiap leaf nide (atau node terminal) memegang label kelas. Node yang paling atas di pohon disebut node akar[5].

D. k-Nearest Neighbor

Klasifikasi Algoritma k-nearestneighbor(k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut, Ketepatan algoritma k-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana memilih dan memberi bobot terhadap fitur, agar performa klasifikasi menjadi lebih baik,menurut (Wu, 2009)KNN juga merupakan contoh teknik lazy learning, yaitu teknik yang menunggu sampai pertanyaan (query) datang agar sama dengan data training[7].

III. METODE

Metode yang di usulkan pada penelitian ini bertujuan untuk melakukan klasifikasi terhadap data pendapatan dan melakukan komparasi tiga algoritma klasifikasi dengan metode evaluasi algoritma Decision tree, K-nn dan Logistic Regression.

Tahapan pertama untuk melakukan komparasi algoritma adalah menentukan objek data yang akan di olah, tahap kedua dilakukan pemisahan data otomatis training dan testing. Tahap ketiga dilakukan proses ekstraksi data mining terhadap data set yang telah di siapkan sebelumnya dengan tiga algoritma Decision tree, K-nn, Logistic Regression. Tahap ke empat melakukan komparasi hasil akurasi.

Data set yang digunakan pada penelitian ini adalah data set hasil kuesioner dari bebagai kalangan masyarakat mengenai pendapatannya.

IV. ANALISA DAN PEMBAHASAN

dataset dari hasil survei dari 32561 responden, ada beberapa variabel dalam survei ini yaitu *age, finalweight, education number, capital gain, capital loss, hours per week.* serta hasil secara fakta pendapatan bulanan responden diatas 50.000$ atau dibawah 50.000$ berdasarkan variabel-variabel yang diketahui.

A	C	E	K	L	M	O
age	fnlwgt	education	capital-ga	capital-los	hours-per	income
39	77516	13	2174	0	40	<=50K
50	83311	13	0	0	13	<=50K
38	215646	9	0	0	40	<=50K
53	234721	7	0	0	40	<=50K
28	338409	13	0	0	40	<=50K
37	284582	14	0	0	40	<=50K
49	160187	5	0	0	16	<=50K
52	209642	9	0	0	45	>50K
31	45781	14	14084	0	50	>50K
42	159449	13	5178	0	40	>50K
37	280464	10	0	0	80	>50K
30	141297	13	0	0	40	>50K
23	122272	13	0	0	30	<=50K
32	205019	12	0	0	50	<=50K
40	121772	11	0	0	40	>50K
34	245487	4	0	0	45	<=50K
25	176756	9	0	0	35	<=50K
32	186824	9	0	0	40	<=50K
38	28887	7	0	0	50	<=50K
43	292175	14	0	0	45	>50K
40	193524	16	0	0	60	>50K
54	302146	9	0	0	20	<=50K
35	76845	5	0	0	40	<=50K
43	117037	7	0	2042	40	<=50K
59	109015	9	0	0	40	<=50K
56	216851	13	0	0	40	>50K
19	168294	9	0	0	40	<=50K
54	180211	10	0	0	60	>50K
39	367260	9	0	0	80	<=50K
49	193366	9	0	0	40	<=50K

Dari dataset ini dapat langsung dilakukan pengolahan menggunakan *Jupiter Notebook*, berikut hasil dari perbandingan dari 3 metode algoritma:

A. Hasil Accuracy Menggunakan Algoritma Decission Trees

Dalam tuning model algoritma Decission Trees akan dicoba memaksimalkan nilai akurasi dengan mengubah

parameter max_depth, min_samples_leaf serta memeriksa keakuratan dua kriteria: gini dan entropi.

```
algorithms = [DecisionTreeClassifier(criterion='gini', max_depth=9),
              DecisionTreeClassifier(criterion='gini', min_samples_leaf=66),
              DecisionTreeClassifier(criterion='gini', max_depth=16, min_samples_leaf=65)]
df_DT = pd.DataFrame(algoritm_score_list(), columns=['algorithm', 'accuracy', 'standardized'])
df_DT
```

```
all predictions finished
```

	algorithm	accuracy	standardized
0	DecisionTreeClassifier(max_depth=9)	0.852	False
1	DecisionTreeClassifier(min_samples_leaf=66)	0.854	False
2	DecisionTreeClassifier(max_depth=16, min_sampl...	0.854	False

Setelah dilakukan tuning pada model DT dihasilkan data akurasi model dari beberapa metode tidak terlalu meiliki perbedaan yang besar. Sehingga diambil nilai tertinggi dengan parameter(criterion='gini',max_depth=16,min_samples_leaf= 65) bernilai 0.854.

B. Hasil Accuracy Menggunakan Algoritma Logistic Regression

Dalam tuning model Logistic Regression akan dicoba memaksimalkan nilai akurasi menggunakan parameter penalty dan C.

```
algorithms = []
for c in [100, 10, 1.0, 0.1, 0.01]:
    algorithms.append(LogisticRegression(solver='liblinear', penalty='l1', C=c))

df_LR = pd.DataFrame(algoritm_score_list(standardized=True), columns=['algorithm', 'accuracy', 'standardized'])
df_LR
```

```
all predictions finished
```

	algorithm	accuracy	standardized
0	LogisticRegression(C=100, penalty='l1', solver...	0.852	True
1	LogisticRegression(C=10, penalty='l1', solver=...	0.853	True
2	LogisticRegression(penalty='l1', solver='libli...	0.853	True
3	LogisticRegression(C=0.1, penalty='l1', solver...	0.851	True
4	LogisticRegression(C=0.01, penalty='l1', solve...	0.826	True

Setelah dilakukan tuning pada model Logistic Regression dihasilkan data akurasi model tertinggi sebesar 0.853 dengan parameter (C=10, penalty='l1', solver='liblinear').

C. Hasil Accuracy Menggunakan Algoritma k-Nearest Neighbor

Selanjutnya dilakukan tuning pada model KNN dengan mencari akurasi tertinggi menggunakan variasi parameter n_

```
fig = plt.figure(figsize=(7, 7))
plt.grid(b=True)
sns.lineplot(x=range(1, 50, 5), y=df_KN_n_neighbors['accuracy'])
plt.plot(df_KN_n_neighbors[df_KN_n_neighbors['accuracy'] == max(df_KN_n_neighbors['accuracy'])].index.values[0]*5+1,
         max(df_KN_n_neighbors['accuracy']),
         "or")
```

Setelah dilakukan tuning didapatkan akurasi tertinggi ketika n=46

C. Perbandingan Hasil Accuracy Dari Beberapa Model Yang Telah Di Tuning

```
algorithms = [DecisionTreeClassifier(),
              DecisionTreeClassifier(criterion='gini', max_depth=16, min_samples_leaf=65),
              LogisticRegression(solver='liblinear'),
              LogisticRegression(C=10, penalty='l1', solver='liblinear'),
              KNeighborsClassifier(),
              KNeighborsClassifier(n_neighbors=46)]
final_list = final_list + algoritm_score_list(standardized=True)
```

```
all predictions finished
```

```
final_df = pd.DataFrame(final_list, columns=['algorithm', 'accuracy', 'standardized'])
final_df.head(12).sort_values(by='accuracy', ascending=False)
```

	algorithm	accuracy	standardized
1	DecisionTreeClassifier(max_depth=16, min_sampl...	0.854	False
7	DecisionTreeClassifier(max_depth=16, min_sampl...	0.854	True
3	LogisticRegression(C=10, penalty='l1', solver=...	0.853	False
9	LogisticRegression(C=10, penalty='l1', solver=...	0.853	True
8	LogisticRegression(solver='liblinear')	0.851	True
11	KNeighborsClassifier(n_neighbors=46)	0.836	True
10	KNeighborsClassifier()	0.834	True
6	DecisionTreeClassifier()	0.816	True
0	DecisionTreeClassifier()	0.815	False
2	LogisticRegression(solver='liblinear')	0.792	False
5	KNeighborsClassifier(n_neighbors=46)	0.790	False
4	KNeighborsClassifier()	0.769	False

Didapatkan dari perbandingan beberapa algoritma yang diterapkan, nilai *accuracy* tertinggi sebesar 0.854 menggunakan algoritma DecisionTreeClassifier dengan parameter(criterion='gini',max_depth=16,min_samples_leaf= 65)

V. KESIMPULAN

Dari pengujian akurasi dataset oleh masing – masing algoritma tersebut dapat disajikan pada tabel. Berdasarkan nilai *accuracy* algoritma yang lebih akurat adalah DecisionTree dan LogisticRegression disusul oleh k-Nearest Neighbor.

Selain itu, Algoritma DecisionTree dan Logistic Regression tidak memerlukan proses data yang di standarisasi terlebih dahulu agar dapat bekerja dengan baik. Sedangkan untuk KNearestNeighbors, dengan dilakukan proses standarisasi data meningkatkan akurasi sebesar 4,6%.

REFERENCES

[1] Sumanta, Jaka. 2005. Fenomena lingkaran kemiskinan di Indonesia : Analisis ekonometri regional

[2] Suryawati. 2004. Teori Ekonomi Mikro. UPP. AMP YKPN. Yogyakarta

[3] Annur. 2018. Klasifikasi Masyarakat Miskin Menggunakan Metode Naïve Bayes. Universitas Ichsan Gorontalo. Gorontalo

[4] Wijaya et al. 2018. Implementasi Algoritma C5.0 dalam klasifikasi pendapatan Masyarakat (Studi Kasus: Kelurahan Mesjid Kecamatan Medan Kota). STMIK Budi Darma. Medan

[5] Amirullah and Taufieurrochman. 2017. Komparasi Model Klasifikasi Algoritma Keterlambatan Siswa Mauk Sekolah. SMTIK Nusa Mandiri, Jakarta

[6] Wu, Xindong & Kumar, Vipin. 2009. The Top Ten Algorithms in Data Mining. Boca Raton: CRC Press

[7] Dewi. 2016. Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan. Manajemen Informatika AMIK BSI Pontianak, Pontianak.