ARS 2018 The 4th Annual Research Seminar
Prosiding Annual Research Seminar 2018
Computer Science and ICT
ISBN : **978-979-587-813-1**
Vol.4 No.1

Prediksi Harga Ponsel Menggunakan Metode Random Forest

Vanissa Wanika Siburian[1]
Jurusan Sistem Komputer
Universitas Sriwijaya
Palembang, Indonesia
[1]vanissasiburian@gmail.com

Ika Elvina Mulyana[2]
Jurusan Sistem Komputer
Universitas Sriwijaya
Palembang, Indonesia
[2]ikaelvinamulyana@gmail.com

***Abstrak*—Harga ponsel dapat dipengaruhi oleh spesifikasi yang dimiliki ponsel tersebut. Dengan spesifikasi harga dapat ditentukan. Pada penelitian ini dilakukan klasifikasi untuk memprediksi harga ponsel dengan spesifikasi yang diberikan dengan metode Random Forest. Pengklasifikasian pada penelitian menggunakan tujuh variabel prediksi dan satu variabel respon menghasilkan akurasi sebesar 81%. Kemudian tingkat akurasi tertinggi pada variabel respon terdapat pada kategori harga murah.**

Kata Kunci—Klasifikasi, Random Forest, Kurva ROC

I. PENDAHULUAN

Harga adalah suatu nilai tukar, dimana digunakan dalam pemasaran dan bisnis yang paling efektif. Harga menjadikan patokan pertama dalam pembelian dan penjualan. Setiap pelanggan pada saat ingin membeli sebuah ponsel pasti memikirkan spesifikasi ponsel tersebut yang akan dibelinya sesuai dengan estimasi harga yang disiapkan.

Artificial Intelligence (kecerdasan buatan) yang membuat mesin mampu menjawab pertanyaan secara cerdas sekarang ini dalam bidang rekayasa yang sangat luas. *Machine Learning* memberi kita teknik terbaik untuk kecerdasan buatan seperti klasifikasi, regresi, pembelajaran terawasi dan pembelajaran tanpa pengawasan dan banyak lagi. Kita dapat menggunakan pengklasifikasi apapun seperti *Decision Tree, Naïve Bayes*, dan banyak lagi. Berbagai jenis algoritma pemilihan fitur tersedia untuk memilih fitur yang terbaik dan meminimalkan kumpulan data. Karena ini adalah masalah pengoptimalan maka banyak teknik yang digunakan untuk mengoptimalkan atau mengurangi dimensi dari dataset. Pada hal ini dibahas banyak fitur yang sangat penting dipertimbangkan untuk memperkirakan harga ponsel. Misalnya prosesor dari ponsel, ketahanan baterai, ukuran dan ketebalan ponsel juga merupakan faktor penting. Memori internal, pixel kamera, dan kualitas video harus dipertimbangkan. Browsing internet juga merupakan salah satu kendala paling penting di era teknologi saat ini.[1]

Menggunakan data sebelumnya untuk memprediksi harga produk yang tersedia dan peluncuran baru merupakan latar belakang penelitian yang menarik bagi para peneliti *machine learning*. Sameerchand-Pudaruth [2] memprediksi harga mobil bekas di Mauritius. Dia menerapkan banyak teknik seperti regresi linier berganda, *k-nearest neighbors (KNN), Decision Tree, dan Naïve Bayes* untuk memprediksi harga. Sameerchand-Pudaruth mendapat hasil yang sebanding dari semua teknik ini [2] Selama penelitian ditemukan bahwa sebagian besar algoritma populer yaitu *Decision Tree dan Naïve Bayes* tidak dapat menangani, mengklasifikasikan, dan memprediksi nilai-nilai numerik. Jumlah contoh untuk penelitiannya hanya 97 (47 Toyota + 38 Nissan + 12 Honda). Karena lebih sedikit jumlah instansi yang digunakan, akurasi prediksi yang sangat buruk dicatat.

Konsep *Support Vector Machine (*SVM) digunakan oleh peneliti lain Mariana Listiani [3] untuk pekerjaan yang sama. Listiani memprediksi harga mobil sewaan menggunakan teknik yang disebutkan di atas. Ditemukan dalam penelitian ini bahwa teknik SVM jauh lebih baik dan akurat untuk prediksi harga dibandingkan dengan yang lain seperti regresi linier berganda ketika satu set data yang sangat besar tersedia. Peneliti juga menunjukkan bahwa SVM juga menangani data dimensi tinggi lebih baik dan menghindari masalah yang kurang pas dan pas. Untuk menemukan fitur-fitur penting untuk SVM Listiani digunakan Algoritma Genetika. Namun, teknik gagal menunjukkan dalam hal varian dan standar deviasi berarti mengapa SVM lebih baik daripada regresi berganda sederhana[3].

Terdapat beberapa metode dalam melakukan pengklasifikasian salah satunya adalah metode *Random Forest*. Dimana *Random Forest* dapat meningkatkan akurasi karena adanya pemilihan secara acak dalam membangkitkan simpul anak untuk setiap node (simpul diatasnya) dan diakumulasikan hasil klasifikasi dari setiap pohon *(tree),* kemudian dipilih hasil klasifikasi yang paling banyak muncul.[4]

Pada paper ini digunakan metode *Random Forest* dalam memperkirakan harga ponsel. Hasil diharapkan dapat menghasilkan tingkat akurasi yang lebih tinggi dari penelitian sebelumnya.

II. METODOLOGI

A. DATASET

Datasetnya yang digunakan dalam penelitian ini adalah *Mobile Price Prediction* yang didapat dari web Kaggle [1]. Dataset ini berisi spesifikasi dari ponsel. Pada dataset ini terdapat dua jenis variabel yang digunakan yaitu tujuh variabel prediksi dan satu variabel respon.

Adapun tujuh variabel prediksi adalah *battery power* (mAh), *dual sim*, *four g* (4G), internal memori (GB), RAM

ISBN : 978-979-587-813-1
Vol.4 No.1

(Mb), *touch screen*, dan wifi. Semua variabel memiliki nilai yang berbeda seperti ditunjukkan pada tabel 1.

TABEL I. PERBEDAAN NILAI VARIABEL

Varible	*Count*	*Min*	*Max*	*Mean*
Battery power (mAh)	2000	501	1998	1238,51
Dual sim	2000	0	1	0,5
Four g (4G)	2000	0	1	0,5
Internal Memori (GB)	2000	2	64	32,04
RAM (Mb)	2000	256	3998	2124,21
Touch Screen	2000	0	1	0,5
Wifi	2000	0	1	0,5

Pada tabel 1 Terdapat nilai *count, min, max* dan *mean*. *Count* adalah jumlah seluruh data dari variabel, *min* adalah nilai terkecil pada variabel, *max* adalah nilai terbesar dari variabel dan *mean* adalah nilai rata-rata dari seluruh data per variabel.

Untuk variabel respon (label) menggunakan 1 variabel dijelaskan pada tabel 2.

TABEL II. VARIABEL RESPON

Variabel	Keterangan
0	Murah
1	Standar
2	Mahal
3	Sangat mahal

B. METODE RANDOM FOREST

Metode *random forest* (RF) merupakan metode yang dapat meningkatkan hasil akurasi, karena dalam membangkitkan simpul anak untuk setiap node dilakukan secara acak. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node, internal node,* dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. *Root node* merupakan simpul yang terletak paling atas, atau biasa disebut sebagai akar dari pohon keputusan. Internal node adalah simpul percabangan, dimana node ini mempunyai output minimal dua dan hanya ada satu input. Sedangkan *leaf node* atau terminal node merupakan simpul terakhir yang hanya memiliki satu input dan tidak mempunyai output. Pohon keputusan dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai information gain. Untuk menghitung nilai *entropy* digunakan rumus seperti pada persamaan 1, sedangkan nilai information gain menggunakan persamaan 2[5].

$$\boldsymbol{Entropy}\ (\boldsymbol{Y}) = -\sum_{\mathbf{i}} \mathbf{p}\ (\mathbf{c}|\mathbf{Y})\ \log_{\mathbf{2}}\ \mathbf{p}\ (\mathbf{c}|\mathbf{Y}) \qquad (1)$$

Dimana Y adalah himpunan kasus dan p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$\boldsymbol{Information\ Gain}\ (\boldsymbol{Y}, \boldsymbol{a})$$

$$= \boldsymbol{Entropy}\ (\boldsymbol{Y}) - \sum_{\boldsymbol{veValues}} \frac{|\mathbf{Yv}|}{|\mathbf{Ya}|} \boldsymbol{Entropy}\ (\boldsymbol{Yv}) \quad (2)$$

Dimana *Values* (a) merupakan semua nilai yang mungkin dalam himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a.

C. PENERAPAN ALGORITMA GAIN RATIO

Pemilihan atribut sebagai simpul, baik akar (*root*) atau simpul internal didasarkan pada nilai *information gain* tertinggi dari atribut-atribut yang ada. Nilai *gain ratio* diperoleh dari hasil perhitungan information gain yang dibagi dengan *split information*. Nilai *split information* dapat dilihat pada persamaan 3[6]. Sedangkan nilai *gain ratio* seperti pada persamaan 4[7].

$$\boldsymbol{Split\ Information}\ (\boldsymbol{S}, \boldsymbol{A}) = \sum_{\mathbf{i}}^{\mathbf{c}} \frac{|\mathbf{Si}|}{|\mathbf{S}|}\ \log_{\mathbf{2}} \frac{|\mathbf{Si}|}{|\mathbf{S}|} \qquad (3)$$

Dimana *split information* (S, A) adalah nilai estimasi *entropy* dari variabel *input* S yang memiliki kelas c dan |Si|/|S| merupakan probabilitas kelas i dalam atribut.

$$\boldsymbol{Gain\ Ratio}\ (\boldsymbol{S}, \boldsymbol{A}) = \frac{\boldsymbol{Information\ Gain}\ (\boldsymbol{S},\boldsymbol{A})}{\boldsymbol{Split\ Information}\ (\boldsymbol{S},\boldsymbol{A})} \qquad (4)$$

D. CONFUSION MATRIX MULTICLASS

Matriks *Confusion Multiclass* adalah tabel yang sering digunakan untuk menggambarkan kinerja model klasifikasi pada sutau set data *testing* yang nilai-nilai yang sebenarnya sudah diketahui. Dapat dilihat persamaan Matriks *Confusion Multiclass* pada tabel 3 di bawah ini.

TABEL III. MATRIKS *CONFUSION MULTICLASS*

		PREDICTED			
		A	B	C	D
ACTUAL	A	TP_A	E_{AB}	E_{AC}	E_{AD}
	B	E_{BA}	TP_B	E_{BC}	E_{BD}
	C	E_{CA}	E_{CB}	TP_B	E_{CD}
	D	E_{EA}	E_{DB}	E_{DC}	TP_C

Tabel diatas di asumsikan ada 4 kelas prediksi dengan variabel A,B,C, dan D. TP adalah singkatan dari *True Positive* yang merupakan kasus di mana kita memprediksi ya dan nilai aktualnya benar.

Catatan bahwa matriks *confusion multiclass* adalah perkembangan dari matiks *confusion binary* dimana sebelumnya terdapat FN (*False Negative*), FP (*False Positive*), dan TN (*True Negative*). Kemudian pada matriks *confusion multiclass* hanya tertera TP karena untuk penentuan FN adalah dari seluruh jumlah baris per variabel sedangkan untuk penentuan FP adalah dari seluruh jumlah kolom per variable dan TN adalah kasus-kasus di mana kita memprediksi tidak ada dan nilai aktualnya salah.

ISBN : 978-979-587-813-1
Vol.4 No.1

E. PERHITUNGAN AKURASI

Perhitungan akurasi dilakukan setelah proses klasifikasi selesai dilakukan. Perhitungan ini berfungsi menunjukkan tingkat kebenaran pengklasifikasian data terhadap data yang sebenarnya. Perhitungan akurasi dilakukan dengan menggunakan rumus sebagai berikut pada persamaan 5.

$$\textbf{Akurasi} = \frac{\sum \textbf{data uji benar klasifikasi}}{\sum \textbf{jumlah total data uji}} \textbf{ X 100} \quad (5)$$

Setelah menghitung nilai akurasi dari setiap spesifikasi maka dilakukan penghitungan nilai *precision* atau nilai presisi. Rumus untuk menghitung nilai presisi sebagai berikut pada persamaan 6.

$$\textbf{precision} = \frac{\textbf{tp}}{\textbf{tp+fp}} \textbf{ X 100\%} \quad (6)$$

Dimana tp adalah nilai *true positive* dan fp adalah nilai *false negative*. Nilai tp merupakan nilai yang sama antara data latih (*predictive*) dengan data uji (*reference*). Nilai tp + fp pada rumus 4 merupakan jumlah keseluruhan data uji.[8]

F. KURVA ROC (Receiver Operating Characteristic)

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positive* sebagai garis vertikal[9].

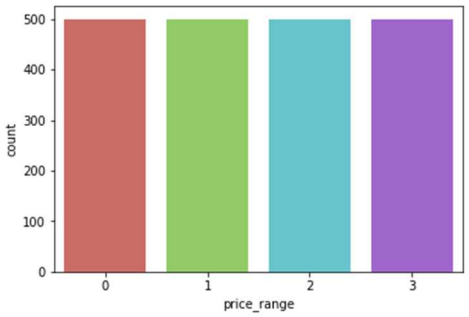
Nilai yang di plot pada kurva adalah nilai TPR (*True Positive Rate*) dan FPR (*False Positive Rate*) dimana masing-masing nilai dapat dihitung dengan menggunakan persamaan sebagai berikut:

$$\textbf{TPR} = \frac{\textbf{TP}}{\textbf{TP+FP}} \quad (7)$$

$$\textbf{FPR} = \frac{\textbf{FP}}{\textbf{TP+FP}} \quad (8)$$

III. ANALISA DAN PEMBAHASAN

Pada bab ini membahas hasil dari penelitian tentang prediksi harga ponsel. Sebelumnya telah dikumpulkan data spesifikasi sebanyak 2000 data. Setelah di identifikasi di dapatkan hasil seperti pada gambar 1.



Gambar 1. Perhitungan jumlah data tiap price_range

Dari 2000 data spesifikasi dikumpulkan menghasilkan jumlah banyak data yang sama yakni 500 data dengan *price range* 0,1,2,3.

Kemudian, untuk pembagian data menjadi 70:30, dimana 70% adalah data *training* dan 30% adalah data *testing*. Setelah pembagian data *training* dan *testing* tersebut selanjutnya ke tahap penelitian prediksi dengan menggunakan data yang telah dibagi dan menghasilkan nilai sebagai berikut.

TABEL IV. HASIL KELUARAN *PRECISION, RECALL, F1-SCORE*

	precision	***recall***	***f1-score***
0	0.91	0.89	0.90
1	0.74	0.76	0.75
2	0.75	0.68	0.72
3	0.84	0.90	0.87
avg / total	0.81	0.81	0.81

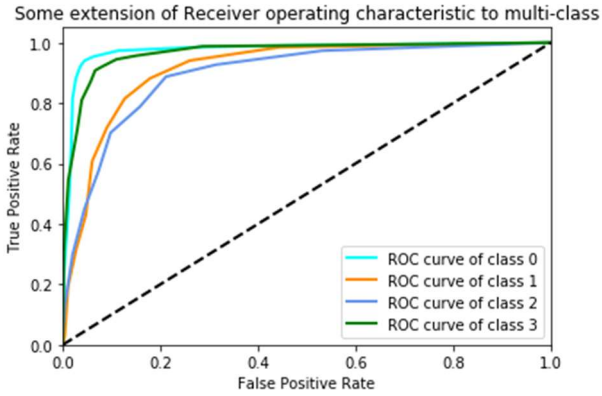
Nilai *precision, recall, f1-score* yang dihasilkan dari setiap variabel bernilai sama sebesar 81%.

Dan kemudian dilakukan perhitungan akurasi dengan menggunakan *confusion matrix* multiclass seperti pada tabel V.

TABEL V. MATRIKS CONFUSION *MULTICLASS* DATA

		PREDICTED			
		0	1	2	3
ACTUAL	0	134	17	0	0
	1	14	102	19	0
	2	0	18	106	27
	3	0	0	17	146

Dengan menggunakan persamaan 5 didapat hasil akurasi sebesar 81%. Adapun dengan menggunakan kurva ROC dapat prediksi dari model yang dihasilkan cukup memuaskan.



Gambar 2. Kurva ROC dari dataset

Gambar 2 terdapat lima garis dimana garis berwarna biru muda menunjukan kurva ROC dalam kelas 0 (Murah)

ISBN : 978-979-587-813-1
Vol.4 No.1

yang berarti tingkat prediksinya tertinggi, garis berwarna orange menunjukan kurva ROC dalam kelas 1 (Standar) yang berarti tingkat prediksinya cukup tinggi, garis berwarna biru tua menunjukan kurva ROC dalam kelas 2 (Mahal) yang berarti tingkat prediksinya cukup rendah, garis berwarna hijau menunjukan kurva ROC dalam kelas 3 (Sangat Mahal) yang berarti tingkat prediksinya terendah dan garis berwarna hitam putus-putus adalah baseline (acuan) dimana jika semakin luas jarak antara garis-garis tersebut dengan garis baseline maka semakin bagus tingkat prediksi.

IV. KESIMPULAN

Dari hasil percobaan yang dilakukan didapatlah tingkat akurasi prediksi dengan menggunakan metode *Random Forest* sebesar 81%. Selain akurasi didapatkan pula nilai *precision*, nilai *recall*, dan nilai dari *f1-score* yang sama sebesar 81%. Kemudian tingkat akurasi tertinggi pada variabel respon terdapat pada kategori harga murah yang di tunjukan pada gambar 2.

Dapat disimpulkan bahwa dari hasil percobaan ini dengan menggunakan metode *random forest* lebih baik dari percobaan sebelumnya.

REFERENSI

[1] M. Asim and Z. Khan, “Mobile Price Class prediction using Machine Learning Techniques.”

[2] S. Pudaruth, “Predicting the price of used cars using machine learning techniques,” *Int. J. Inf. Comput. Technol*, vol. 4, no. 7, pp. 753–764, 2014.

[3] M. Listiani, “Support vector regression analysis for price prediction in a car leasing application,” Citeseer, 2009.

[4] F. A. Kurniawan and A. P. Kurniati, “Analisis Dan Implementasi Random Forest dan Classification dan Regression Tree (CART) untuk Klasifikasi pada Misuse Intrusion Detection System.” IT Telkom, Program Studi Teknik Informatika, Skripsi. Bandung: IT Telkom, 2011.

[5] K. Schouten, F. Frasincar, and R. Dekker, “An information gain-driven feature study for aspect-based sentiment analysis,” in *International Conference on Applications of Natural Language to Information Systems*, 2016, pp. 48–59.

[6] R. C. Barros, A. C. De Carvalho, A. A. Freitas, and others, *Automatic design of decision-tree induction algorithms*. Springer, 2015.

[7] S. B. Kotsiantis, “Decision trees: a recent overview,” *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013.

[8] R. Rukmigayatri and others, “Klasifikasi Kemunculan Titik Panas Pada Lahan Gambut Di Sumatera Dan Kalimantan Menggunakan Algoritme Random Forest,” 2015.

[9] C. Vercellis, *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons, 2011.