Jurnal Sistem Komputer dan Kecerdasan Buatan
Vol II No.2 Maret Tahun 2019

Implementasi Data Mining untuk Prediksi Standar Hidup Layak Berdasarkan Tingkat Kesehatan dan Pendidikan Masyarakat

Bofandra Muhammad[1]

[1] School of Engineering and Technology, Tanri Abeng University
[1]bofandra@tau.ac.id,

Diterima : 01 Maret 2019
Disetujui : 20 April 2019

***Abstract*— Indeks Pembangunan Manusia (IPM) adalah sebuah indikator yang menunjukkan kemampuan manusia di suatu tempat untuk dapat meningkatkan kualitas hidupnya. Di dalamnya terdapat komponen tingkat kesehatan, tingkat pendidikan, dan tingkat kesejahteraan (standar hidup layak). Pada penelitian ini, dilakukan implementasi Data Mining untuk memprediksi standar hidup layak berdasarkan tingkat kesehatan dan tingkat pendidikan. Sebagaimana metode Badan Pusat Statistik (BPS) dalam megukur Indeks Pembangunan Manusia, indikator tingkat kesehatan menggunakan Angka Harapan Hidup (AHH), tingkat pendidikan menggunakan Rata-Rata Lama Sekolah (RLS) dan Harapan Lama Sekolah (HLS), sementara standar hidup layak menggunakan Pengeluaran Per Kapita (PPK). Data BPS yang digunakan dalam penelitian ini mencakup seluruh Kota dan Kabupaten di Indonesia selama kurun waktu 2010 – 2016. Data tersebut kemudian dikelompokkan berdasarkan provinsi. Pengujian dilakukan pula untuk menguji ketepatan prediksi PPK untuk di tahun yang sama, serta 1, 2, 3, 4, dan 5 tahun setelahnya. Hasil pengujian menunjukkan bahwa telah dapat dilakukan prediksi nilai PPK berdasarkan AHH, RLS, dan HLS dengan 28 dari 34 (82,32%) provinsi memperoleh nilai *coefficient of determination* di atas 0,7. Jadi, lebih dari 70% variasi nilai PPK dapat diprediksi berdasarkan data AHH, RLS, dan HLS.**

***Index Terms*—Indeks Pembangunan Manusia, Data Mining, *Coefficient of Determination*, Angka Harapan Hidup, Rata-Rata Lama Sekolah, Harapan Lama Sekolah, dan Pengeluaran per Kapita.**

I. PENDAHULUAN

Indeks Pembangunan Manusia (IPM) adalah sebuah standard internasional yang menunjukkan kemampuan manusia di suatu tempat untuk dapat meningkatkan kualitas hidupnya. Di dalamnya terdapat komponen tingkat kesehatan, tingkat pendidikan, dan tingkat kesejahteraan. Indikator tingkat kesehatan adalah Angka Harapan Hidup (AHH), indikator tingkat pendidikan adalah Rata-Rata Lama Sekolah (RLS) dan Harapan Lama Sekolah (HLS), sementara indikator standar hidup layak adalah Pengeluaran per Kapita (PPK). IPM menekankan pada pentingnya meningkatkan semua indikator, bukan hanya indikator terkait perekonomian saja [1,2].

Meskipun bukan sebagai satu-satunya indikator, namun dalam persepsi secara umum indikator perekonomian (standar hidup layak) masih seringkali dijadikan sebagai acuan untuk mengatakan sukses atau tidaknya pembangunan di suatu wilayah. Oleh karenanya, penelitian ini dimaksudkan untuk mendorong peningkatan indikator-indikator lainnya (kesehatan dan pendidikan) dengan menunjukkan pengaruhnya terhadap peningkatan indikator perekonomian dalam satuan yang terukur untuk seluruh kota dan kabupaten di Indonesia.

Pada penelitian ini, akan digunakan 2 metode Data Mining, yaitu Regresi Linier dan Multilayer Perceptron. Regresi Linier mewakili metode linier, sementara Multilayer Perceptron mewakili metode non-linier. Pengujian dengan kedua metode tersebut untuk menunjukkan apakah metode linier atau non-linier yang dapat menghasilkan tingkat hubungan antara *independent variable* (AHH, RLS, dan HLS) dan *dependent variable* (PPK) yang lebih tinggi

berdasarkan perbandingan nilai coefficient of determination (r^2).

II. LANDASAN TEORI

F. *Regresi Linier*

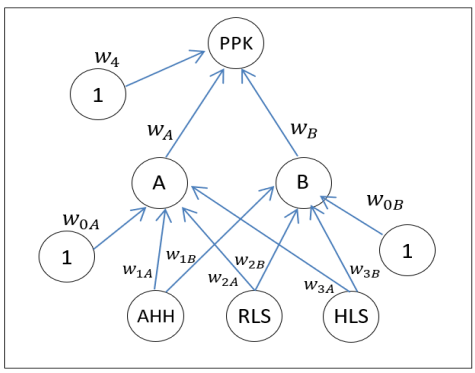
Regresi linier menggunakan fungsi matematis untuk memodelkan variabel kelas (target yang dituju) sebagai kombinasi linear variabel-variabel atribut (yang mempengaruhi variabel kelas). Dalam penelitian ini, variabel kelas adalah PPK, sementara variabel-variabel atributnya adalah AHH, RLS, dan HLS.

$$w_0 + w_1.AHH + w_2.RLS + w_3.HLS = PPK \quad (1)$$

Pada rumus (1), w_1 , w_2 , dan w_3 merupakan bobot (di dalam kombinasi linier) untuk masing-masing variabel atribut. Sementara w_0 adalah nilai bias [3].

G. *Multilayer Perceptron*

Tidak semua hubungan antara variabel kelas dan variabel atribut dapat dimodelkan dalam bentuk persamaan linier. Oleh karenanya dibutuhkan cara lain untuk modelkannya. Di antara metode yang ditemukan adalah menggunakan *Multilayer Perceptron*. Pada metode ini, variabel-variabel atribut tidak langsung mempengaruhi variabel kelas, akan tetapi melalui variabel perantara terlebih dahulu (lingkaran A dan B) sebagaimana dapat dilihat pada Gambar 1 [3].



Gambar 1. Multilayer Perceptron

H. *Coefficient of Determination*

Di antara cara untuk mengevaluasi hasil metode numerik, seperti Regresi Linier dan *Multilayer Perceptron*, adalah dengan menghitung nilai *root mean-squared error* (RMSE), *mean absolute error* (MAE), atau *coefficient of determination* (r^2). Pada penelitian ini, digunakan r^2 sebagai nilai evaluasi hasil prediksi. r^2 adalah perbandingan nilai varians *predicted* (hasil prediksi) dan *actual* (data yang dimiliki) pada variabel kelas. Nilai r^2 berupa bilangan desimal antara 0 s/d 1. r^2 juga dapat dihitung dengan mengkuadratkan nilai *correlation coefficient* (r) [3, 4].

$$r = \frac{S_{PA}}{\sqrt{S_p S_A}} \quad (2)$$

Pada rumus (2), S_{PA} adalah nilai kovarian dari *predicted* dan *actual* pada variabel kelas, S_P adalah standar deviasi nilai *predicted* pada variabel kelas, dan S_A adalah standar deviasi nilai *actual* pada variabel kelas [3]. r^2 dipilih untuk digunakan karena tidak terpengaruh oleh rentang nilai pada variabel kelas. Contoh perbedaan skala nilai kelas pada penelitian ini dapat dilihat dari perbedaan rentang nilai PPK antara Kota Lombok Timur di Nusa Tenggara Barat dan Kota Jakarta Selatan di DKI Jakarta.

I. Penelitian Sebelumnya

Sebelumnya, Kamila Latif telah menggunakan Regresi Linier untuk membuktikan adanya pengaruh tingkat pendidikan terhadap pendapatan rumah tangga daerah perkotaan di Kabupaten Tanah Datar, Sumatera Barat. Selain itu, Herry Faisal juga membuktikan adanya pengaruh pendidikan terhadap jumlah penduduk miskin di Provinsi Kalimantan Barat dengan menggunakan metode Least Square Dummy Variable (LSDV) [5, 6].

Perbedaan penelitian ini, dibandingkan penelitian-penelitian sebelumnya tersebut, adalah digunakannya juga metode analisa data non-linear, yaitu Multilayer Perceptron, untuk membandingkannya dengan metode Analisa linier, yaitu Regresi Linier. Selain itu, ruang lingkup data yang digunakan mencakup kota/ kabupaten di seluruh provinsi di Indonesia.

III. METODOLOGI

Penelitian ini menggunakan metodologi CRISP-DM yang sudah dikenal luas sebagai standar langkah-langkah yang perlu dilakukan dalam proses Data Mining. Tahapan-tahapan dalam CRISP-DM: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment*.

Pada tahap Business Understanding, ditentukan tujuan dari proses Data Mining yang akan dilakukan serta manfaatnya bagi pihak-pihak terkait.

Pada tahap Data Understanding, dilakukan pengumpulan data awal dan eksplorasi data. Setelahnya, pada tahap Data Preparation, data awal tersebut dipilih, dibersihkan, dan disesuaikan bentuknya untuk siap diproses di tahap selanjutnya.

Pada tahap Modelling, dilakukan pembuatan model pengetahuan dan pengujian. Kemudian dilakukan Evaluation terhadap hasil pengujian, untuk menentukan sukses atau tidaknya proses Data Mining yang telah dilakukan.

Pada tahap terakhir, dilakukan proses Deployment, untuk memastikan hasil dari proses Data Mining dapat digunakan oleh target pengguna dengan sebaik-baiknya [7].

IV. PEMBAHASAN

Business Understanding

Tujuan dari proses *Data Mining* dalam penelitian ini adalah untuk mendapatkan metode terbaik, sekaligus membuktikan, bahwa tingkat standar hidup layak dapat diprediksi berdasarkan tingkat pendidikan dan tingkat kesehatan masyarakat. Hal ini diharapkan dapat menjadi acuan untuk peneliti selanjutnya dalam mengembangkan metode yang lebih baik, ataupun mengimplementasikannya di berbagai kota/ kabupaten di Indonesia.

Data Understanding

Sumber data penelitian ini adalah halaman Indeks Pembangunan Manusia pada website resmi BPS (https://www.bps.go.id/subject/26/indeks-pembangunan-manusia.html). Data yang tersedia adalah AHH, RLS, HLS, dan PKK untuk seluruh kota/ kabupaten di Indonesia.

Data Preparation

Pada penelitian ini digunakan data AHH, RLS, HLS, dan PKK untuk seluruh kota/ kabupaten di Indonesia pada tahun 2010 – 2016. Data tersebut kemudian dipisahkan per provinsi, sehingga diperoleh 34 kelompok data. Penulis kemudian membagi kembali masing-masing kelompok data berdasarkan waktu untuk *training* (pembelajaran) dan testing (pengujian). Sehingga diperoleh 204 file untuk data latih (train) dan 204 file untuk data uji (test). Penamaannya mengikuti ketentuan <train/test>_<provinsi>_<kode-tahun>.csv.

Contohnya, file data latih *train_jak_x1.csv* berisi data AHH, RLS, dan HLS tahun 2014, serta PPK tahun 2015 untuk seluruh kota/ kabupaten di Provinsi DKI Jakarta. Adapun contoh untuk data uji adalah *test_sumbar_x.csv* yang berisi data AHH, RLS, HLS, dan PPK tahun 2016 untuk seluruh kota/ kabupaten di Provinsi Sumatera Barat.

Tabel 1. Data Latih (*train*) dan Data Uji (*test*)

Tahun Data		***AHH***	***RLS***	***HLS***	***PPK***
x	*train*	2015	2015	2015	2015
	test	**2016**	**2016**	**2016**	**2016**
x1	*train*	2014	2014	2014	2015
	test	**2015**	**2015**	**2015**	**2016**
x2	*train*	2013	2013	2013	2015
	test	**2014**	**2014**	**2014**	**2016**
x3	*train*	2012	2012	2012	2015
	test	**2013**	**2013**	**2013**	**2016**
x4	*train*	2011	2011	2011	2015
	test	**2012**	**2012**	**2012**	**2016**
x5	*train*	2010	2010	2010	2015
	test	**2011**	**2011**	**2011**	**2016**

Modelling

Setiap data latih (*train*) dilakukan pembelajaran menggunakan Regresi Linier dan *Multilayer Perceptron*. Prosesnya dilakukan menggunakan software WEKA, dengan parameter-parameter default. Pada Regresi Linier, digunakan metode M5 untuk proses pencarian koefisien persamaan yang optimal. Sementara untuk Multilayer Perceptron, digunakan 3 *input nodes*, 1 *output node*, 2 *hidden layers*, 500 *epochs*, dan nilai *learning rate* 0,3.

Hasil dari pelatihan, diperoleh 204 model pengetahuan untuk Regresi Linier, dan 204 model pengetahuan untuk Multilayer Perceptron.

Kemudian setiap model pengetahuan diuji menggunakan data uji (*test*) yang bersesuaian. Contohnya, model pengetahuan yang diperoleh dari hasil pembelajaran terhadap data *train_jak_x1.csv*, diujikan terhadap file data *test_jak_x1.csv.*

Evaluation

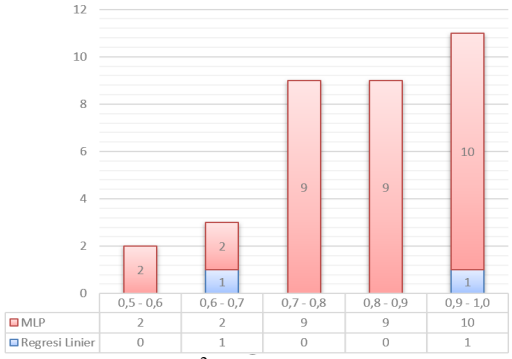
Sebagaimana disebutkan dalam Bab II, pada penelitian ini digunakan nilai *coefficient of determination* (r^2) untuk mengevaluasi hasil dari prediksi masing-masing model pengetahuan yang telah dibuat.

Pada Tabel 2, pada 2 kolom terakhir (Metode Terbaik) dapat dilihat bahwa *Multilayer Perceptron* (M) lebih banyak menghasilkan nilai *coefficient of determination* yang lebih tinggi dibandingkan Regresi Linier (L) untuk masing-masing provinsi, yaitu sebanyak 32 dari 34 provinsi. Hanya 2 provinsi (Kep. Riau dan DKI Jakarta) yang memperoleh nilai r^2 terbaik menggunakan Regresi Linier.

Tabel 2. Hasil Pengujian

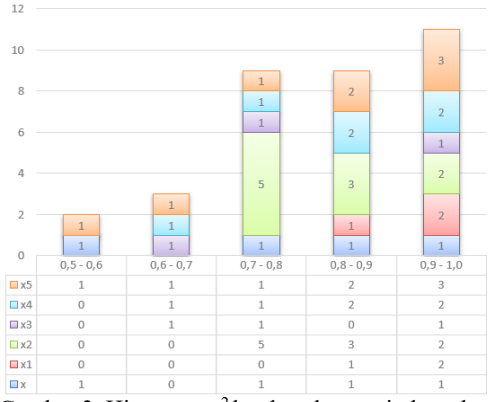
<METODE>	Regresi Linier						MLP					
<TAHUN>	x	x1	x2	x3	x4	x5	x	x1	x2	x3	x4	x5
ACEH	0,745	0,723	0,732	0,745	0,740	0,710	0,859	0,861	**0,861**	0,859	0,850	0,846
SUMATERA UTARA	0,723	0,707	0,709	0,670	0,667	0,620	0,742	0,736	**0,761**	0,656	0,675	0,733
SUMATERA BARAT	0,596	0,605	0,535	0,555	0,549	0,582	0,663	0,676	0,680	**0,684**	0,675	0,682
RIAU	0,712	0,740	0,830	0,845	0,720	0,751	0,913	0,831	0,921	**0,907**	0,925	0,670
JAMBI	0,000	0,229	0,295	0,296	0,295	0,271	0,394	0,513	0,421	0,522	0,517	**0,528**
SUMATERA SELATAN	0,684	0,693	0,673	0,664	0,699	0,686	0,843	0,827	**0,938**	0,839	0,882	0,873
BENGKULU	0,617	0,000	0,153	0,572	0,587	0,563	0,782	0,799	**0,877**	0,760	0,748	0,758
LAMPUNG	0,600	0,577	0,577	0,501	0,556	0,449	0,636	0,574	0,580	0,672	**0,697**	0,512
KEP. BANGKA BELITUNG	0,536	0,542	0,529	0,436	0,640	0,592	**0,862**	0,836	0,816	0,538	0,748	0,756
KEP. RIAU	0,405	0,408	0,421	0,421	0,952	**0,981**	0,964	0,932	0,937	0,961	0,957	0,964
DKI JAKARTA	0,574	0,580	0,387	0,487	0,600	**0,605**	0,597	0,433	0,385	0,262	0,588	0,595
JAWA BARAT	0,662	0,664	0,666	0,664	0,624	0,675	0,699	0,723	**0,788**	0,451	0,751	0,748
JAWA TENGAH	0,665	0,666	0,654	0,644	0,657	0,655	0,735	0,723	0,727	0,735	0,715	**0,744**
DI YOGYAKARTA	0,990	0,996	0,992	0,988	0,991	0,998	0,989	0,998	0,998	0,998	0,997	**1,000**
JAWA TIMUR	0,764	0,765	0,772	0,766	0,764	0,736	0,772	0,774	**0,790**	0,772	0,778	0,767
BANTEN	0,947	0,972	0,960	0,951	0,935	0,776	0,976	**0,986**	0,981	0,972	0,924	0,950
BALI	0,917	0,919	0,929	0,926	0,920	0,897	**0,934**	0,924	0,919	0,914	0,907	0,912
NUSA TENGGARA BARAT	0,665	0,545	0,499	0,453	0,640	0,795	0,757	0,734	0,688	**0,798**	0,755	0,762
NUSA TENGGARA TIMUR	0,696	0,704	0,696	0,727	0,766	0,766	0,748	0,757	0,770	0,710	**0,781**	0,757
KALIMANTAN BARAT	0,616	0,569	0,562	0,541	0,612	0,650	0,800	**0,876**	0,767	0,873	0,820	0,401
KALIMANTAN TENGAH	0,487	0,480	0,509	0,391	0,423	0,469	0,480	0,408	**0,759**	0,597	0,456	0,461
KALIMANTAN SELATAN	0,366	0,324	0,311	0,311	0,309	0,683	0,765	0,789	0,887	0,888	0,858	**0,895**
KALIMANTAN TIMUR	0,573	0,553	0,557	0,824	0,544	0,552	0,873	0,888	0,944	0,800	**0,971**	0,869
KALIMANTAN UTARA	0,832	0,825	0,909	0,961	0,997	0,977	0,996	**0,999**	0,987	0,981	0,989	0,921
SULAWESI UTARA	0,558	0,548	0,458	0,421	0,468	0,472	**0,578**	0,280	0,218	0,252	0,191	0,313
SULAWESI TENGAH	0,600	0,630	0,599	0,624	0,649	0,596	0,733	0,808	0,786	0,800	**0,816**	0,770
SULAWESI SELATAN	0,517	0,513	0,546	0,544	0,538	0,521	**0,735**	0,733	0,723	0,726	0,727	0,641
SULAWESI TENGGARA	0,646	0,653	0,648	0,477	0,691	0,691	0,774	0,791	0,838	0,279	**0,844**	0,719
GORONTALO	0,953	0,956	0,866	0,923	0,911	0,775	0,972	0,967	0,955	0,955	0,971	**0,974**
SULAWESI BARAT	0,684	0,571	0,000	0,000	0,000	0,000	0,881	0,961	0,713	**0,961**	0,734	0,500
MALUKU	0,573	0,574	0,578	0,656	0,659	0,667	0,799	0,826	**0,858**	0,791	0,760	0,615
MALUKU UTARA	0,898	0,899	0,892	0,897	0,877	0,898	0,888	0,940	**0,951**	0,924	0,940	0,946
PAPUA BARAT	0,698	0,705	0,724	0,720	0,675	0,667	0,742	0,775	**0,781**	0,604	0,644	0,714
PAPUA	0,693	0,688	0,685	0,682	0,688	0,704	0,818	0,810	0,807	0,812	0,812	**0,832**

Adapun sebaran nilai (histogram) *coefficient of determination* (r^2) terbaik berdasarkan metode Data Mining untuk ke-34 provinsi dapat dilihat pada Gambar 2. Nilai r^2 terbanyak berada pada rentang 0,9 - 1,0 yaitu sebanyak 11 dari 34 provinsi. Selain itu, 28 dari 34 (82,32%) provinsi memperoleh nilai r^2 di atas 0,7. Hal tersebut berarti lebih dari 70% variasi nilai PPK dapat diprediksi berdasarkan data AHH, RLS, dan HLS.



Gambar 2. Histogram r^2 berdasarkan metode Data Mining

Sementara untuk menentukan apakah prediksi data PPK lebih baik menggunakan data AHH, RLS, dan HLS di tahun yang sama, atau 1, 2, 3, 4, dan 5 tahun setelahnya, dapat dilihat pada Gambar 3.



Gambar 3. Histogram r^2 berdasarkan periode waktu

10 provinsi (29,41%) mendapatkan nilai r^2 terbaik dengan menggunakan periode waktu x2, yaitu prediksi PPK berdasarkan data AHH, RLS, dan HLS pada 2 tahun sebelumnya. Pada urutan selanjutnya x5: 8 provinsi, x4: 6 provinsi, x: 4 provinsi, serta x1 dan x3: 3 provinsi.

Deployment

Seluruh data terkait penelitian ini dapat diakses pada tautan https://goo.gl/3jp4Rw. Di dalamnya, terdapat pula file *knowledge-model* (per provinsi) yang dapat langsung dibuka menggunakan aplikasi WEKA untuk dapat memprediksi PPK berdasarkan AHH, RLS, dan HLS.

V. KESIMPULAN

Tujuan dari penelitian ini, prediksi standar hidup layak berdasarkan tingkat kesehatan dan tingkat pendidikan, telah tercapai dengan 82,32% (28 dari 34 provinsi) berhasil diprediksi dengan nilai *coefficient of determination* (r^2) lebih dari 70%. Selain itu, juga dapat disimpulkan bahwa *Multilayer Perceptron* dapat memprediksi lebih baik daripada Regresi Linier. Terkait periode waktu, PKK lebih baik diprediksi menggunakan AHH, RLS, dan HLS pada 2 tahun sebelumnya.

Saran bagi peneliti selanjutnya adalah untuk mencoba menggunakan metode *Data Mining* lain (selain *Multilayer Perceptron* dan Regresi Linier). Selain itu, peneliti selanjutnya dapat mencoba melakukan prediksi dengan menggunakan data per kota/ kabupaten (tanpa digabungkan per provinsi).

DAFTAR PUSTAKA

[1] "Human Development Report: 2016," UNDP, New York: 2016.
[2] "Indeks Pembangunan Manusia 2015," BPS, 2016.
[3] Bluman, Allan G, "Ellementary Statistics: A Step by Step Approach", New York: McGraw-Hill, 2009.
[4] Witten, Ian H. and F. Eibe, "Practical Machine Learning Tools and Techniques", San Fransisco: Elsevier Inc, 2005.
[5] Latif, K, "Pengaruh pendidikan terhadap tingkat pendapatan rumah tangga," KPK IPB-Unand, 1990.
[6] Faisal, H, "Pengaruh tingkat pendidikan dan kesehatan terhadap produktivitas dan jumlah penduduk miskin di Provinsi Kalimantan Barat," Fakultas Ekonomi, Universitas Tanjung Pura, 2013.
[7] Chapman, P, dkk, "CRISP-DM 1.0: Step by step data mining guide", SPSS Inc, 2000.