Pengembangan Aplikasi Machine Learning Menggunakan Algoritma Support Vector Regression Dan Statistical-Based Feature Selection Untuk Memprediksi Kemiskinan

Development On Machine Learning Application Using Support Vector Regression and Statistical-Based Feature Selection For Predicting Poverty

R. Ilham Fadhilah I.[1], Dr. Dedy Rahman Wijaya, S.T. [2], M.T., Elis Hernawati, S.T., M.Kom. [3]
[1,2,3]*Program Studi D3 Sistem Informasi, Fakultas Ilmu Terapan, Universitas Telkom*
[1]ilhamfadhilah@student.telkomuniversity.ac.id, [2]dedyrw@tass.telkomuniversity.ac.id,
[3]elishernawati@tass.telkomuniversity.ac.id

Abstrak

Masalah yang terjadi di setiap negara adalah kemiskinan yang bahkan terjadi di Indonesia. Setiap negara telah mengembangkan teknologi untuk memprediksi tingkat kemiskinan di setiap daerahnya. Salah satu cara yang dipakai untuk memprediksi kemiskinan di Indonesia adalah dengan melakukan survey. Akan tetapi proses survey memakan waktu yang cukup lama dan akan menggunakan sumber daya manusia yang cukup banyak. Maka dikembangkanlah aplikasi prediksi kemiskinan menggunakan data salah satu e-commerce yang ada di Indonesia dengan memanfaatkan data transaksi yang terjadi di dalamnya. Dengan bantuan machine learning aplikasi ini akan dapat memprediksi tingkat kemiskinan dari data e-commerce. Algoritma machine learning yang digunakan adalah Support Vector Machine dan seleksi fitur yang digunakan untuk memilih fitur yang relevan untuk di prediksi adalah dengan menggunakan seleksi fitur Statistical-Based Aplikasi ini akan membantu melengkapi data survey yang dilakukan oleh Badan Pusat Statistik dalam menangani survey angka kemiskinan di Indonesia dengan cara yang lebih efisien memiliki keakuratan cukup akurat.

Kata Kunci: Machine Learning, Kemiskinan, E-Commerce, Prediksi

Abstract

The problem that occurs in every country is poverty which even occurs in Indonesia. Each country has developed technology to predict poverty levels in each region. One of the methods used to predict poverty in Indonesia is conduct a survey. However, the survey process takes a long time and use a lot of human resources. Then a poverty prediction application were developed using data from one of the e-commerce sites in Indonesia by utilizing transaction data that occurs in it. By using one of machine learning algorithm this application will be able to predict poverty levels from e-commerce data. The chosen machine learning algorithm is Support Vector Machine and feature selection that used to select relevant features for predictions is Statistical-Based feature selection. This application will help complete the survey data conducted by the Badan Pusat Statistik in handling poverty survey in Indonesia in a more efficient way with more accurate result.

Keywords: Machine Learning, Poverty, E-Commerce, Prediction.

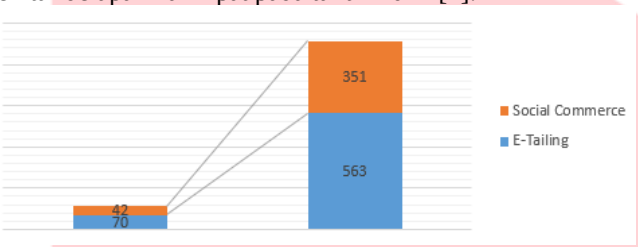
I. PENDAHULUAN

A. Latar Belakang

Masalah yang sering timbul di berbagai negara termasuk di Indonesia salah satunya adalah kemiskinan. Banyak faktor yang menyebabkan kemiskinan itu timbul di negara Indonesia. Pada Bulan Maret 2018, Badan Pusat Statistik mencatat bahwa jumlah penduduk miskin (penduduk dengan pengeluaran per kapita per bulan di bawah Garis Kemiskinan) di Indonesia mencapai 25,95 juta orang (9,82 persen), berkurang sebesar 633,2 ribu orang dibandingkan dengan kondisi September 2017 yang sebesar 26,58 juta orang (10,12 persen) [1].

Di sisi lain, pasar e-commerce di Indonesia menjadi pasar e-commerce terbesar di Asia Tenggara dengan kontribusi hingga lima puluh persen dari seluruh transaksi di wilayah Asia Tenggara. Peningkatan penetrasi e-commerce di Indonesia seiring dengan pertambahan penduduk yang menggunakan internet, sehingga kontribusi terhadap ekonomi Negara Indonesia pada sektor ini berpotensi terus meningkat. Menurut studi McKinsey perusahaan jasa seperti

GO-JEK, traveloka, dan platform B2B seperti IndoTrading berada di luar lingkup riset ini. Namun, bahkan tanpa memperhitungkan sektor jasa B2B, gross merchandise value pasar e-commerce di Indonesia diproyeksikan akan tumbuh sekitar delapan kali lipat pada tahun 2022 [2].



Gambar 1 Gambar Proyeksi Nilai Pasar E-commerce Indonesia (Dalam Triliun Rupiah) [1]

Selama ini data yang didapatkan oleh BPS adalah dengan menyelenggarakan Survey Sosial Ekonomi Nasional yang biasa disingkat Susenas. Susenas adalah survey berbasis rumah tangga yang mengumpulkan informasi mengenai karakteristik sosial ekonomi seperti: pendidikan, kesehatan, keluarga berencana, keterangan bepergian, kriminalitas, perumahan, perlindungan sosial serta konsumsi dan pengeluaran rumah tangga. Susenas pertama dimulai sejak tahun 1963 dan selalu diselenggarakan setiap tahun [3]. Survey yang dilaksanakan oleh BPS yaitu Susenas, dapat diperkirakan akan memakan waktu yang cukup lama dan akan ada rentang waktu antara survey satu dengan yang lainnya, bahkan jika dihitung biaya survey untuk pendataan data masyarakat kurang mampu, dapat dipastikan akan memakan biaya yang cukup besar. Dengan adanya teknologi informasi khususnya teknologi Machine Learning, cara survey untuk pendataan kemiskinan di suatu daerah akan lebih mudah dan akurat. Bahkan dengan pengembangan lebih lanjut, teknologi tersebut akan dapat memantau naik turunnya kemiskinan di Indonesia tanpa waktu lama. Dari permasalahan diatas, dapat ditarik solusi untuk menyelesaikan permasalahan waktu dan biaya survey Susenas yaitu dengan mengembangkan sebuah perangkat lunak dengan teknologi Machine Learning. Machine Learning yang digunakan akan melengkapi survey kemiskinan pada suatu daerah dengan data dari aplikasi e-commerce yang sedang marak digunakan untuk masyarakat Indonesia. Dengan adanya aplikasi tersebut, survey mengenai kemiskinan di suatu daerah di Indonesia akan sangat terbantu dan tidak akan memakan waktu yang lama dalam mendapatkan data karena akan menggunakan data e-commerce.

E-commerce yang sering dipakai oleh masyarakat Indonesia dapat dipakai untuk menjadi media prediksi kemiskinan di Indonesia. Menggunakan perhitungan jumlah pembelian dari suatu daerah dapat diketahui di daerah tersebut apakah daerah tersebut termasuk daerah yang Makmur atau sebaliknya. Kemudian dari perhitungan tersebut dapat dijadikan proses prediksi kemiskinan menggunakan aplikasi dengan basis machine learning.

B. Rumusan Masalah

Rumusan masalah yang mendasari pengembangan Aplikasi Machine Learning Menggunakan Algoritma Support vector machine dan Statistical-based Feature Selection Untuk Memprediksi Kemiskinan adalah sebagai berikut:

1. Bagaimana cara melengkapi hasil survey dan sensus tingkat kemiskinan di suatu daerah di Indonesia berdasarkan data E-commerce?
2. Bagaimana cara mengidentifikasi item yang berpengaruh dalam prediksi kemiskinan?
3. Bagaimana cara penyajian data hasil dari prediksi kemiskinan?

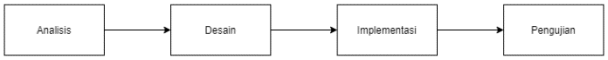
C. Tujuan

Tujuan yang diharapkan menyelesaikan rumusan masalah pada pembuatan proyek akhir ini adalah dengan sebagai berikut:

1. Untuk melengkapi hasil survey dan sensus kemiskinan adalah dengan menggunakan algoritman machine learning support vector machine dan statistical-based feature selection.
2. Untuk mengidentifikasi item yang berpengaruh dari data e-commerce adalah dengan menggunakan algoritma seleksi fitur yaitu statistical-based feature selection.
3. Untuk menyajikan data hasil prediksi kemiskinan adalah dengan menggunakan platform web yang di dalamnya akan menjalankan algoritma support vector regression.

D. Metode Pengerjaan

Metode pengembangan perangkat lunak yang digunakan pada pengerjaan proyek akhir ini adalah dengan menggunakan SDLC Waterfall [2]. Berikut gambaran tahapan – tahapan SDLC sequential linear:



Gambar 2 Tahapan SDLC Waterfall [4]

Dari gambaran diatas dapat dijabarkan mengenai metode pengerjaan yang digunakan saat pengerjaan proyek akhir. SDLC sequential linear digunakan karena pada pengembangan aplikasi ini menggunakan waktu yang relatif pendek. Lama pengerjaan akan di jelaskan secara detail di bagian jadwal pengerjaan. Berikut adalah penjabaran dari tahapan SDLC sequential linear:

1. Analisis

Pada tahap ini dilakukan penganalisaan kebutuhan aplikasi dari mulai penganalisaan dataset yang akan digunakan. Selain kebutuhan dari sisi aplikasi, pada tahap ini dianalisis kebutuhan library yang akan digunakan terutama pada bagian permodelan machine learning.

2. Desain

Pada tahap desain, dilakukan pembuatan desain kebutuhan aplikasi seperti kebutuhan database, alur aplikasi, permodelan alur data machine learning, dan desain antarmuka pengguna.

3. Implementasi

Proses implementasi pada proyek akhir ini akan menggunakan platform web sebagai antamuka pengguna dengan sistem machine learning yang digunakan untuk memprediksi kemiskinan. Algoritma machine learning yang diimplementasikan pada aplikasi ini adalah algoritma support vector regression. Karena dataset yang digunakan adalah dataset yang berukuran cukup besar, maka proses permodelan machine learning akan dibantu dengan metode seleksi fitur dengan algoritma statistical-based feature selection.

4. Pengujian

Aplikasi proyek akhir ini akan diuji dengan melakukan pengujian metode black-box pada bagian antarmuka pengguna untuk menguji kesesuaian respon aplikasi dengan aktifitas user.

II. TINJAUAN PUSTAKA

A. *Kemiskinan*

Untuk mengukur kemiskinan, BPS menggunakan konsep kemampuan memenuhi kebutuhan dasar (basic needs approach). Dengan pendekatan ini, kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran. Jadi Penduduk Miskin adalah penduduk yang memiliki rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan [5]. Sumber data utama yang digunakan adalah data Survei Sosial Ekonomi modul konsumsi dan pengeluaran. Dalam menghitung garis kemiskinan adalah dengan menggunakan rumus perhitungan garis kemiskinan (GK). Ketentuan penentuan garis kemiskinan adalah dengan menggunakan konsep berikut [5]:

1. Garis Kemiskinan (GK) merupakan penjumlahan dari Garis Kemiskinan makanan dengan Garis Kemiskinan Non Makanan. Penduduk yang memiliki rata – rata pengeluaran perkapita perbulan dibawah Garis kemiskinan dikategorikan sebagai penduduk miskin.

2. Garis Kemiskinan Makanan (GKM) merupakan nilai pengeluaran kebutuhan minimum makanan yang disetarakan dengan 2100 kilokalori perkapita perhari. Paket komoditi kebutuhan dasar makanan diwakili oleh 52 jenis komoditi seperti padi-padian, umbi-umbian, ikan, daging, telur dan susu, sayuran, kacang-kacangan, buah-buahan, minyak dan lemak, dll.

3. Garis Kemiskinan Non Makanan (GKNM) adalah kebutuhan minimum untuk perumahan, sandang, pendidikan dan kesehatan. Paket komoditi kebutuhan dasar non makanan diwakili oleh 51 jenis komoditi di perkotaan dan 47 jenis komoditi di pedesaan.

B. *Sistem Informasi*

Menurut Fat penertian sistem adalah sebagai beriku: “Sistem adalah suatu himpunan suatu ‘benda’ nyata atau abstrak yang terdiri dari bagian – bagian atau komponen – komponen yang saling berkaitan, berhubungan, berketergantungan, saling mendukung, yang secara keseluruhan bersatu dalam satu kesatuan untuk mencapai tujuan tertentu secara efisien dan efektif [6].” Kemudian kata sistem mengandung arti “kumpulan dari komponen – komponen yang memiliki unsur keterkaitan antara satu dan lainnya [7].”

C. *Machine Learning*

Machine Learning adalah cabang aplikasi dari *Artificial Intelligence* (Kecerdasan Buatan) yang fokus pada pengembangan sebuah sistem yang mampu belajar "sendiri" tanpa harus berulang kali di program oleh manusia. Aplikasi *Machine Learning* membutuhkan data sebagai bahan belajar (*training*) sebelum mengeluarkan *output*. Aplikasi sejenis ini juga biasanya berada dalam domain spesifik alias tidak bisa diterapkan secara general untuk semua permasalahan [8]. *Machine Learning* berfokus pada pengembangan program – program komputer yang dapat mengajarkan dirinya sendiri untuk tumbuh dan berubah bila diberikan data baru. Sebagai contoh, sistem *Machine Learning* dapat dilatih pada pesan email untuk belajar membedakan antara pesan spam dan pesan non-spam. Setelah pembelajaran, dapat digunakan untuk mengklasifikasikan pesan email baru ke kategori folder spam dan non-spam. Pada kedokteran, *Machine Learning* dapat digunakan untuk mendiagnosa penyakit, pada telekomunikasi bentuk panggilan dianalisis untuk optimasi jaringan dan memaksimalkan kualitas layanan [9]

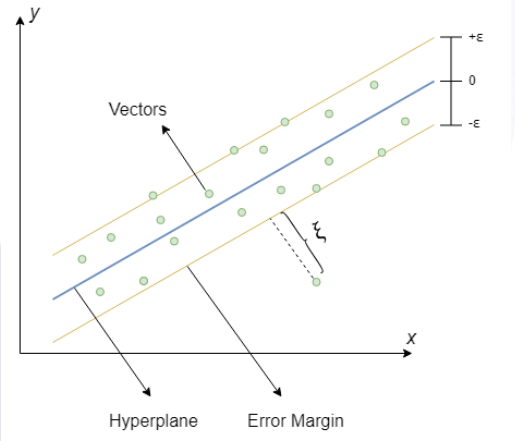
D. *Library Python Scikit-Learn*

Sklearn adalah library open-source yang digunakan untuk Machine Learning di bahasa pemrograman Python. Sklearn juga library yang simple dan efisien untuk melakukan data mining dan analisis data. Sklearn mempunyai fitur regresi, klasifikasi, dan pengelompokan algoritma seperti support vector machine, nearest neighbors, naïve bayes dan sebagainya. Library Scikit-Learn dibangun menggunakan library lain seperti NumPy, SciPy, dan matpolib [10].

E. *Support Vector Regression (SVR)*

SVR merupakan pengembangan dari *support vector machines* (SVM) untuk kasus regresi. SVM merupakan algoritma *machine learning* yang menangani kasus prediksi dengan label berupa nilai bulat atau berkelas – kelas, seperti memprediksi kualitas daging sapi dengan beberapa kategori [11], memprediksi objek 3 dimensi [12], atau memprediksi mood ketika didengarkan audio [13]. Kasus – kasus tersebut menggunakan *dataset* dengan label yang bernilai bulat atau berkelas – kelas sehingga algoritma support vector yang digunakan adalah SVM.

Support vector regression pertama diusulkan oleh Vapnik, Steven Golowich, and Alex Smola pada tahun 1997 [14]. Algoritma *support vector regression* bekerja mirip seperti support vector machines (SVM) dengan menemukan sebuah fungsi *f*(x) regresi agar terbentuk sebuah *hyperplane* yang sesuai dengan input data dengan sebuah *error* (ε) dan membuat *error* (ε) dengan sangat minimal [15].



Gambar 3 Ilustrasi dan Penjelasan SVR

Gambar 3 memperlihatkan sebaran data prediksi dari label (*y*) dari sebuah variabel independen (*x*). Ilustrasi tersebut menampilkan data yang bersifat 1 dimensi atau *one-dimensional space* sehingga hyperplane yang terbentuk adalah garis. Karena SVR menangani data output yang bersifat kontinus, yang berarti nilai dari output tersebut memiliki kemungkinan yang tidak terbatas, maka hal itu

ditangan SVR dengan menggunakan *tolerance margin* dengan nilai ε. Data yang berada di dalam *tolerance margin* tidak dihitung sebagai *error*.

Nilai yang jatuh diluar *tolerance margin* dinyatakan deviasinya dengan *slack variable* (ξ). SVR memiliki kriteria seperti berikut ini:

$$\min_{w,b,\xi,\xi^*} \quad \frac{1}{2}w^t w + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^*$$

$$\text{subject to} \quad w^t\emptyset(x_i) + b - z_i \le \varepsilon + \xi_i,$$
$$z_i - w^t\emptyset(x_i) - b \le \varepsilon + \xi_i^*,$$
$$\xi_i, \xi_i^* \ge 0, i = 1, \dots, l.$$

Dimana $w, C, \xi, \varepsilon,$ dan b masing - masing adalah slope matrix, parameter regularisasi, slack variable, nilai tolerance margin, dan bias.

$$\min_{\propto,\propto^*} \frac{1}{2}(\alpha - \alpha^*)^t Q(\alpha - \alpha^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*)$$

$$\text{subject to} \quad e^t(\alpha - \alpha^*) = 0,$$
$$0 \le \alpha_i, \alpha_i^* \le C, i = 1, \dots, l,$$

Dimana $\alpha - \alpha^*$ menunjukkan penggandaan lagrangian. $Q_{ij} = K(x_i, x_j) \equiv \emptyset(x_i)^t\emptyset(x_j)$. Fungsi perkiraan yang didapatkan setelah menyelesaikan permasalahan diatas adalah

$$\alpha^* - \alpha = \sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(x_i, x) + b$$

F. Seleksi Fitur Statistical-based

Algoritma seleksi fitur didasarkan pada ukuran statistik yang berbeda – beda, karena mereka mengandalkan ukuran statistik yang bervariasi. Banyak dari mereka yang menggunakan metode filter-based. Selain itu sebagian besar algoritma berbasis statistik (statistical-based) menganalisis fitur secara individual [16]. Berikut beberapa algoritma seleksi fitur yang representative pada kategori algoritma berbasis statistik (statistical-based):

1. Chi-Square (chi2)

Chi-square menghitung relevansi antara fitur dengan label [17]. Semakin tinggi nilai chi-square maka semakin banyak informasi dari label yang dibawa oleh fitur tersebut dan semakin relevan fitur tersebut dengan label. Persamaan chi-square dapat dilihat seperi berikut ini [18]:

$$x^2 = \sum_{j=1}^{2}\sum_{s=1}^{c}\frac{(n_{js} - \mu_{js})^2}{\mu_{js}}$$

dimana:
c = jumlah kelas,
n_{js} = jumlah pola pada interval ke -j dan kelas ke -s,
R_i = jumlah pola pada interval ke -j = $\sum_{s=1}^{c} n_{js}$,
K_s = jumlah pola pada kelas ke -s = $\sum_{j=1}^{2} n_{js}$,
N = jumlah semua pola = $\sum_{j=1}^{2} R_i$,
μ_{js} = frekuensi yang diperkirakan dari $n_{js} = R_i \times K_s/N$, jika R_i dan K_s adalah 0, maka μ_{js} adalah 0.1.

Skor Chi-Square yang lebih tinggi, menunjukkan fitur tersebut lebih penting.

2. Chi-Square (chi2)

Ide utama dari CFS adalah untuk menggunakan heuristic berbasis korelasi untuk mengevaluasi nilai subset fitur [16]. Berikut persamaan dari CFS:

$$CFS_score(S) = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{rff}}},$$

dimana CFS score akan menunjukkan heuristis "merit" dari fitur subset S dengan k fitur. $\overline{r_{cf}}$ adalah rata rata dari fitur korelasi kelas dan $\overline{rff}$ adalah rata – rata dari korelasi fitur. Pada persamaan diatas, pembilang menunjukkan kekuatan prediktif dari sebuah set fitur, sedangkan penyebut menunjukkan seberapa banyak redudansi yang dimiliki set fitur. Ide dasar dari CFS adalah fitur yang baik akan memiliki korelasi yang kuat dengan kelas label dan memiliki korelasi yang lemah dengan antar fitur.

3. F-Score

F-score [19] diketahui sebagai teknik sederhana untuk menghitung diskriminasi dari dua set bilangan real dengan training vector adalah xk, k=1, 2, … , n. Jika jumlah instance positif dan negatif masing – masing n+ and n–, maka persamaan f-score dari fitur ke – I didefinisikan sebagai berikut:

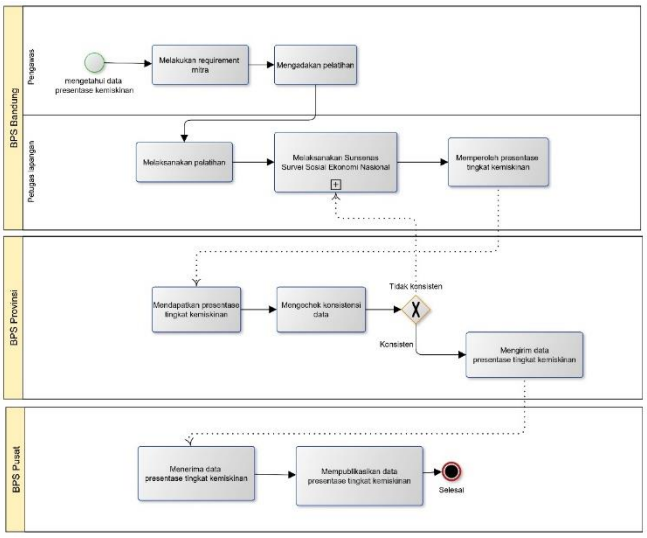
$$F_i = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}\left(x_{k,i}^{(+)} - \bar{x}_i^{+}\right)^2 + \frac{1}{n_- - 1}\sum_{k,i}^{n-}\left(x_{k,i}^{(-)} - \bar{x}_i^{-}\right)^2}$$

Dimana masing – masing dari $\bar{x}_i$, $\bar{x}_i^{(+)}$, dan $\bar{x}$ adalah rata – rata dari fitur ke – i dari keseluruha dataset, rata - rata positif dan rata - rata negatif. Pembilang menunjukkan diskriminasi antara set positif dan negatif, sedangkan penyebut menunjukkan masing – masing dalam 2 set. Semakin besar nilai f-score, maka fitur tersebut akan semakin disikriminatif.

III. ANALISIS DAN PERANCANGAN

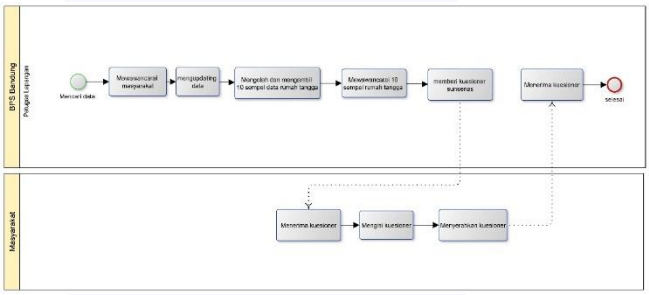
A. Gambaran Sistem Saat Ini

Untuk melakukan sensus kemiskinan, BPS akan melakukan survey secara langsung terhadap masyarakat untuk mengetahui angka kemiskinan pada suatu daerah. Hal tersebut akan memakan waktu yang cukup lama sehingga data yang dicari akan didapatkan hingga berbulan – bulan dan akan memakan biaya yang cukup besar dalam melakukan survey tersebut. Berikut gambaran proses bisnis ketika BPS melakukan survey secara langsung kepada masyarakat:



Gambar 4 BPMN Gambaran Saat Ini

BPS akan mengambil data dari susenas untuk diolah. Proses pengolahan data meliputi pencacahan data dan pengklarifikasian data jika ada data yang dinilai kurang tepat. Setelah proses pengolahan data, maka data akan dikirimkan ke BPS tingkat provinsi dan dicek kembali untuk memastikan data sudah benar. Jika data sudah benar, maka data akan dikirim ke BPS pusat untuk di publikasikan prediksi kemiskinannya ke masyarakat. Berikut detail proses bisnis survey data yang dilakukan oleh Susenas:

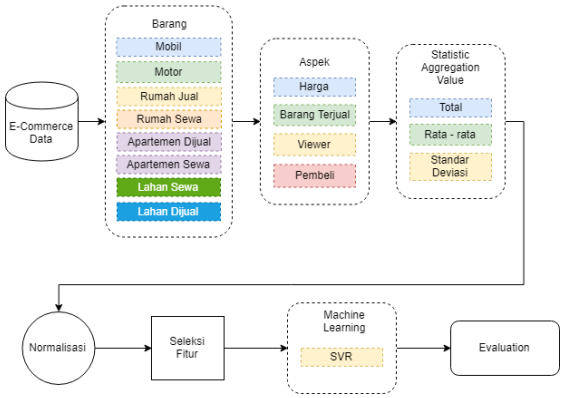


Gambar 5 Gambar Proses Bisnis Survey Susenas

Petugas lapangan akan melakukan survey kepada 10 rumah masyarakat di suatu kecamatan secara acak untuk dimintai data pemasukan dan pengeluaran rumah tangga.

B. *Gambaran Metode Usulan*

Aplikasi *Machine Learning* Menggunakan Algoritma *Support vector machine* Dan *Statistical-based Feature Selection* Untuk Memprediksi Kemiskinan akan mempermudah prediksi tingkat kemiskinan pada suatu daerah. Dengan menggunakan data *e commerce* sebagai *dataset* untuk data *training* pada *Machine Learning*, maka sistem akan mengeluarkan tingkat prediksi kemiskinan di suatu daerah. Pada proses pengolahan data menggunakan *Machine Learning* mulai dari data mentah hingga terbentuk model *regression* terdapat detail proses sebagai berikut:



Gambar 6 Gambar Alur Pengolahan Data Menggunakan Machine Learning

Gambar diatas terdapat beberapa proses yang terjadi sebelum *Machine Learning* dapat membentuk model regresi. Proses pertama adalah proses penentuan data mana saja yang akan menjadi masukkan (input) untuk *Machine Learning* dan pengelompokkan data sesuai dengan aspek yang dilihat dan sesuai dengan fitur ekstraksi statistika.

C. *Abbreviations and Acronyms*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

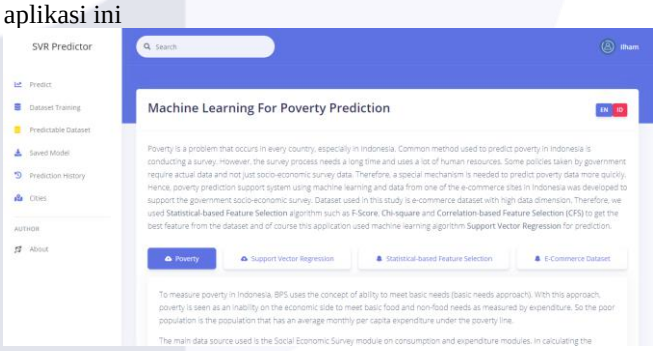
IV. IMPLEMENTASI

A. *Implementasi*

Setelah melalui tahap analisis dan perancangan, maka akan dilakukan tahap implementasi atau pengkodean pada aplikasi. Tampilan website digunakan untuk membuat user interface penggunaan aplikasi yang mudah digunakan oleh pengguna.

1. Halaman Awal / Home

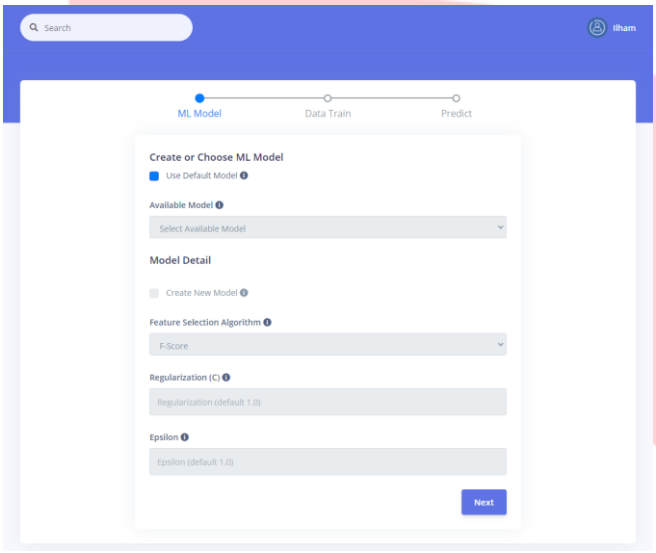
Pada halaman awal, pengguna akan ditampilkan penjelasan singkat mengenai dasar aplikasi ini, dan beberapa dasar yang digunakan untuk memprediksi hingga penjelasan *dataset* yang digunakan. Berikut gambar halaman awal pada aplikasi ini



2. Halaman Prediksi

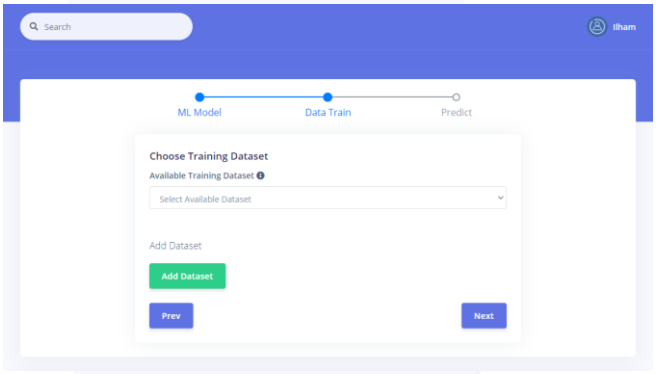
Pada halaman prediksi pengguna akan ditampilkan tahapan - tahapan untuk memprediksi dengan aplikasi ini.

Berikut halaman awal tampilan website yang digunakan sebagai antarmuka untuk memprediksi kemiskinan.



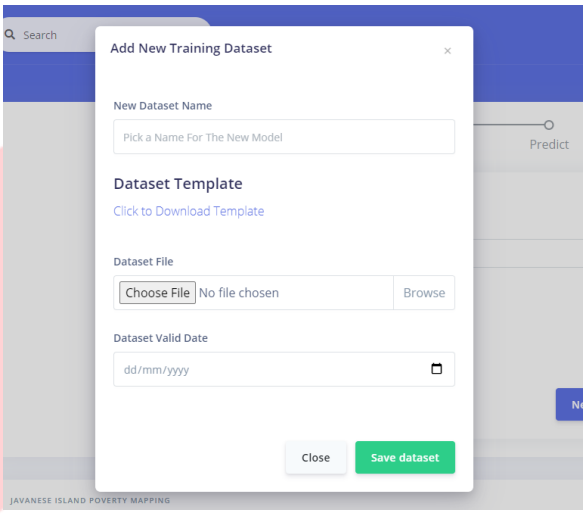
Gambar 7 Tampilan Awal Prediksi Bagian 1

Pada Gambar 7 ditampilkan form tahapan pertama untuk menggunakan model machine learning. Tahapan pertamanya adalah memilih model machine learning yang tersimpan di database, membuat model machine learning yang baru, atau secara default aplikasi akan memilihkan model machine learning yang paling baik yang sudah dipasangkan pada aplikasi.



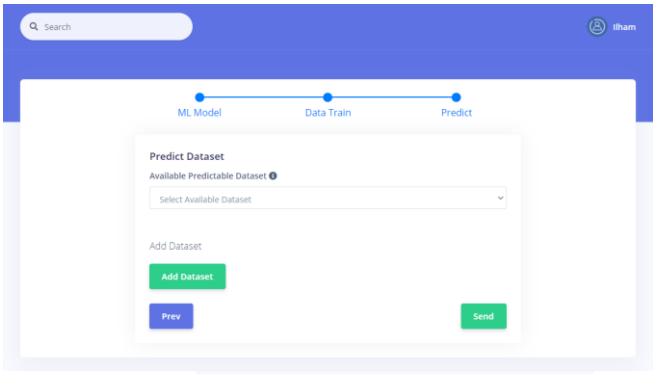
Gambar 8 Tampilan Awal Prediksi Bagian 2

Pada Gambar 8 ditampilkan form tahapan kedua. Berlaku jika pengguna ingin membuat model machine learning yang baru atau jika pengguna ingin bereksperimen dengan hyperparameter pada SVR seperti pada tahap pertama, yaitu melakukan proses training machine learning dengan dataset yang ada di dalam aplikasi.



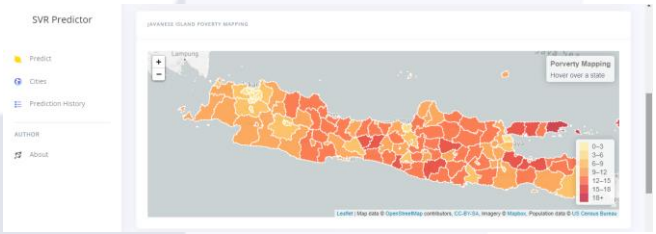
Gambar 9 Form Upload Dataset Training

Jika pengguna ingin mengunggah dataset maka pengguna cukup tekan tombol “Add Dataset” untuk mengunggah dataset ke dalam aplikasi seperti pada Gambar 9.



Gambar 10 Tampilan Awal Prediksi Bagian 3

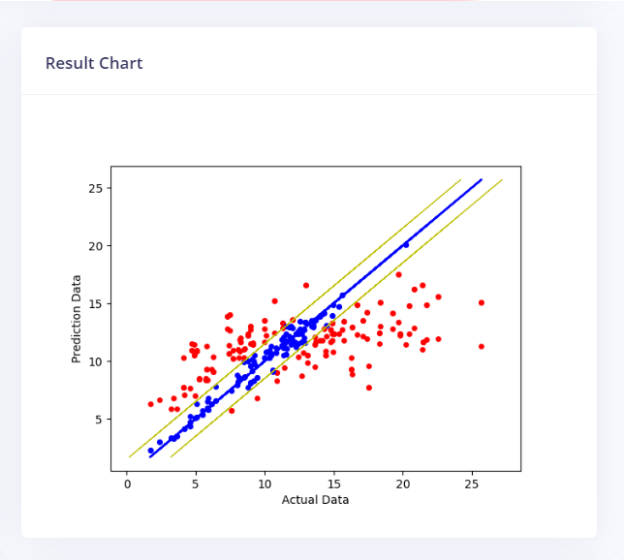
Pada Gambar 10 ditampilkan form tahapan ketiga yaitu tahap prediksi dataset yang dimiliki oleh pengguna menggunakan model yang dipilih pengguna. Seperti pada tahap kedua, aplikasi hanya menerima dataset dengan kolom yang sudah disesuaikan dengan template. Pengguna juga dapat mengunduh template dataset dari link yang terdapat pada Gambar 10



Gambar 11 Pemetaan Kemiskinan Di Pulau Jawa Berdasarkan Hasil Prediksi

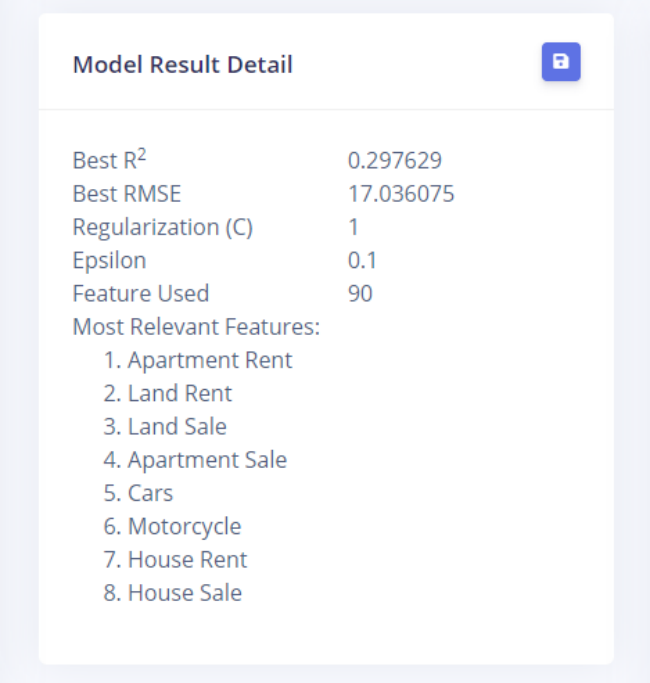
Setelah melalui tahap bagian 3, dan pengguna meng-klik tombol Send, maka aplikasi akan menampilkan hasil prediksi dari dataset yang sudah diinputkan pada tahap prediksi bagian 3. Gambar 11 menunjukkan hasil prediksi yang dipetakan di pulau Jawa. Tingkat kemiskinan yang berbeda – beda ditandai dengan pewarnaan yang berbeda pada setiap daerah. Pada Gambar 11 menunjukkan semakin merah warna suatu

daerah, maka dapat disimpulkan semakin tinggi tingkat kemiskinannya, dan begitu juga sebaliknya.



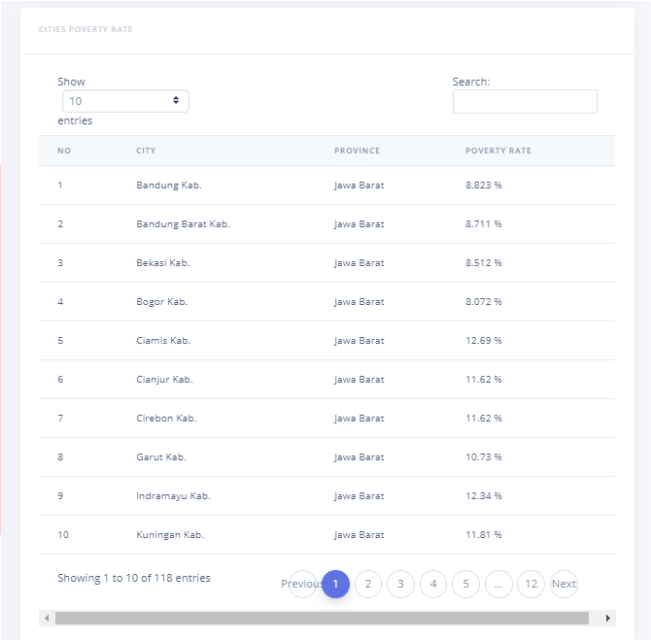
Gambar 12 Gambar Sebaran Data Hasil Prediksi Terhadap Data Aktual

Jika pengguna memilih untuk membuat model machine learning maka, aplikasi akan menunjukkan sebaran hasil prediksinya terhadap data aktual pada dataset untuk training. Dapat dilihat pada Gambar 12 terdapat sebuah garis biru yang menunjukkan hasil prediksi yang seharusnya dikeluarkan oleh model, sedangkan garis kuning adalah tolerance margin, yang berarti data yang berada diantara 2 garis kuning merupakan data yang memiliki error yang ditoleransi oleh model, selebihnya merupakan hasil prediksi yang jauh dari data aktual.



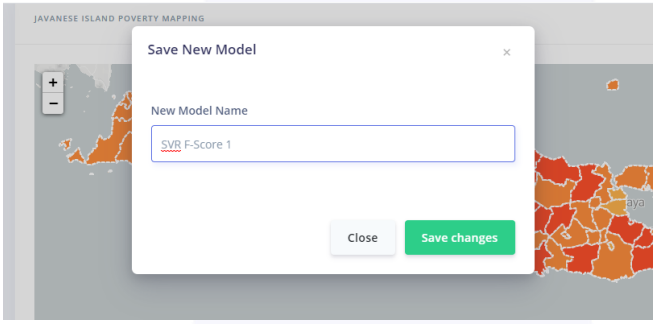
Gambar 13 Detail Hasil Model Machine Learning Baru

Pada gambar Gambar 13, aplikasi menunjukkan hasil dari model yang baru dibuat oleh pengguna, hingga fitur yang dianggap berpengaruh pada prediksi.



Gambar 14 Detail Tingkat Kemiskinan Pada Setiap Kota dan Kabupaten di Pulau Jawa

Aplikasi juga menunjukkan detail tingkat kemiskinan setiap kota dalam bentuk tabel seperti pada gambar Gambar 14. Sehingga nantinya pengguna dapat melihat secara lebih detail mengenai kota, provinsi, dan tingkat kemiskinannya.



Gambar 15 Penyimpanan Model Baru Ke Dalam Database

Pengguna dapat menyimpan model yang baru dibuat tadi ke dalam database, sehingga model tersebut bisa digunakan untuk memprediksi kembali pada waktu yang akan datang seperti padaGambar 15.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Dengan proses pengembangan aplikasi untuk memprediksi kemiskinan menggunakan algoritma machine learning, dapat disimpulkan beberapa poin yaitu:

1. Pada proyek akhir ini, dibangun algoritma machine learning support vector regression dan aplikasi berbasis web yang sudah dapat memprediksi kemiskinan dan dapat digunakan untuk melengkapi data sensus kemiskinan yang dilakukan oleh BPS.

2. Pada proyek akhir ini digunakan algoritma seleksi fitur statistical-based feature selection untuk memilih fitur – fitur yang digunakan untuk prediksi. Selain itu, hasil dari fitur

yang dipilih akan menentukan item – item e-commerce yang paling berpengaruh. Berdasarkan hasil percobaan item – item yang berpengaruh pada dataset e-commerce adalah Total Harga Rumah Sewa, Total Sewa Apartemen Terjual, Total Harga Sewa Apartemen, Total Harga Jual Apartemen, dan Total Pembeli Jual Apartemen.

3. Pada proyek akhir ini sudah dikembangkan aplikasi berbasis web yang menjalankan algoritma machine learning yaitu support vector regression dengan algoritma seleksi fitur statistical-based. Berdasarkan hasil pengujian, aplikasi ini bisa digunakan oleh pihak – pihak terkait.

B. Saran

Pengembangan aplikasi untuk memprediksi kemiskinan menggunakan algoritma machine learning ini masih terdapat beberapa kekurangan, oleh karena itu berikut beberapa saran untuk dapat diterapkan di penelitian lebih lanjut:

1. Meningkatkan akurasi prediksi pada machine learning sehingga hasil yang didapatkan lebih dapat memastikan tingkatan kemiskinan di suatu daerah

2. Menampilkan data prediksi dengan lebih menarik seperti mengubah warna baris detail kota di tabel sesuai dengan tingkat kemiskinan di suatu kota tersebut dan menampilkan detail item – item produk dari e-commerce tersebut yang menjadi indikator kemiskinan dari suatu daerah jika pengguna menekan suatu daerah pada peta.

REFERENCES

[1] BPS, "Profil Kemiskinan di Indonesia Maret 2018," *Badan Pus. Stat.*, no. 57, pp. 1–8, 2018, [Online]. Available: https://www.bps.go.id/pressrelease/2018/07/16/1483/persentase-penduduk-miskin-maret-2018-turun-menjadi-9-82-persen.html.

[2] D. Praditya, "Tech In Asia," *Tech In Asia*, 2019. https://id.techinasia.com/prediksi-ecommerce-indonesia (accessed Oct. 21, 2019).

[3] Admin Web kependudukankalbar.com, "Kependudukan Kalimantan Barat," *Kependudukan Kalbar*, 2019. http://www.kependudukankalbar.com/survei-sosial-ekonomi-nasional-susenas-2015.html (accessed Sep. 19, 2019).

[4] Y. Firmansyah and U. Udi, "Penerapan Metode SDLC Waterfall Dalam Pembuatan Sistem Informasi Akademik Berbasis Web Studi Kasus Pondok Pesantren Al-Habib Sholeh Kabupaten Kubu Raya, Kalimantan Barat," *J. Teknol. dan Manaj. Inform.*, vol. 4, no. 1, 2017, doi: 10.26905/jtmi.v4i1.1605.

[5] Badan Pusat Statistik, "Kemiskinan dan Ketimpangan," 2019. https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html.

[6] J. Hutahean, *Konsep Sistem Informasi*, vol. 3, no. 1. 2005.

[7] R. E. Indrajit, "Manajemen Sistem Informasi dan Teknologi Informasi."

[8] Budy, "Pengertian & Konsep Dasar Machine Learning," 2017. http://teknosains.com/others/pengertian-konsep-dasar-machine-learning (accessed Sep. 20, 2019).

[9] W. Budiharto, *Machine Learning dan Computational Intelligence*. Yogyakarta: Penerbit Andi, 2016.

[10] Scikit-Learn, "Scikit-Learn," 2010. https://scikit-learn.org/stable/index.html (accessed Oct. 02, 2019).

[11] D. R. Wijaya, R. Sarno, and E. Zulaika, "Noise filtering framework for electronic nose signals: An application for beef quality monitoring," *Comput. Electron. Agric.*, vol. 157, no. January 2018, pp. 305–321, 2019, doi: 10.1016/j.compag.2019.01.001.

[12] O. Herouane, L. Moumoun, and T. G. Mohamed Chahhou, "A hybrid boosted-SVM classifier for recognizing parts of 3D objects," *Int. J. Intell. Eng. Syst.*, vol. 11, no. 2, pp. 102–110, 2018, doi: 10.22266/IJIES2018.0430.12.

[13] R. Sarno, J. A. Ridoean, D. Sunaryono, and D. R. Wijaya, "Classification of music mood using MPEG-7 audio features and SVM with confidence interval," *Int. J. Artif. Intell. Tools*, vol. 27, no. 5, 2018, doi: 10.1142/S0218213018500161.

[14] D. Basak, S. Pal, and D. C. Patranabis, "Support Vector Regression," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10634 LNCS, no. 10, pp. 699–708, 2017, doi: 10.1007/978-3-319-70087-8_72.

[15] R. Amanda, H. Yasin, and A. Prahutama, "Analisis Support Vector Regression (Svr) Dalam Memprediksi Kurs Rupiah Terhadap Dollar Amerika Serikat," *J. Gaussian*, vol. 3, no. 4, pp. 849–857, 2014.

[16] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.

[17] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2018-Novem, pp. 160–163, 2019, doi: 10.1109/ICSESS.2018.8663882.

[18] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," *Proc. Int. Conf. Tools with Artif. Intell.*, no. December, pp. 388–391, 1995, doi: 10.1109/tai.1995.479783.

[19] P. Jaganathan, N. Rajkumar, and R. Kuppuchamy, "A Comparative Study of Improved F-Score with Support Vector Machine and RBF Network for Breast Cancer Classification," *Int. J. Mach. Learn. Comput.*, vol. 2, no. 6, pp. 741–745, 2012, doi: 10.7763/ijmlc.2012.v2.227.