APLIKASI UNTUK MEMPREDIKSI KEMISKINAN BERBASIS DATA E-COMMERCE MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION DAN SPARSE LEARNING BASED FEATURE SELECTION

APPLICATION TO PREDICT POVERTY BASED ON E-COMMERCE DATA USING LOGISTIC REGRESSION ALGORITHM AND SPARSE LEARNING BASED FEATURE SELECTION

Sherli Yualinda[1], Elis Hernawati, S.T., M.Kom.[2], Dr. Dedy Rahman Wijaya, S.T., M.T.[3] [123]Program Studi D3 Sistem Informasi, Fakultas Ilmu Terapan Universitas Telkom
sherli@student.telkomuniversity.ac.id[1], elishernawati@tass.telkomuniversity.ac.id[2], dedyrw@tass.telkomuniversity.ac.id[3]

Abstrak

Kemiskinan adalah keadaan dimana terdapat ketidakmampuan untuk memenuhi suatu kebutuhan serta keperluan mulai dari kebutuhan makan, pakaian, tempat tinggal, pendidikan, kesehatan serta yang lainnya, suatu tingkat kemiskinan dapat diukur oleh BPS. Secara konsep, untuk mengukur tingkat kemiskinan yaitu dengan cara perhitungan kemampuan seseorang untuk memenuhi kebutuhan dasar atau *basic needs approach* yang diukur dari sisi pengeluaran oleh BPS. Metode lain untuk melengkapi hasil survei dan sensus yang diusulkan peneliti untuk memprediksi kemiskinan adalah dengan mengunakan *machine learning logistic regression* dengan metode *sparse learning based feature selection* baerbasis data *e-commerce*. Dari hasil percobaan ini, menghasilkan nilai yang cukup relevan antara nilai prediksi jumlah fitur dengan nilai asli, tetapi sejumlah kecil fitur tidak selalu menunjukkan hasil yang buruk dan sebaliknya, di mana penggunaan yang besar jumlah fitur tidak selalu mendapatkan hasil yang baik.

Kata Kunci: Kemiskinan, BPS, *machine learning logistic regression* , *sparse learning based feature selection*, data *e-commerce*.

Abstract

Poverty is a condition where there is an inability to meet a need and needs ranging from food, clothing, shelter, education, health and other needs, a poverty level can be measured by BPS. Conceptually, to measure poverty, that is by calculating a person's ability to meet basic needs or basic needs approach, which is measured in terms of expenditure by BPS. Another method to supplement survey and census results proposed by researchers to predict poverty is to use machine learning logistic regression with sparse learning based features selection methods based on e- commerce data. From the results of these experiments, produce a value that is quite relevant between the predicted value of the number of features with the original value, but the small number of features does not always show poor results and vice versa, where the use of a large number of features does not always get good results.

Keywords: Poverty, BPS, machine learning logistic regression, sparse learning based on feature selection, e-commerce data.

I. PENDAHULUAN

Sebagai salah satu negara berkembang, Indonesia juga sama mengalami permasalahan yang terkait dengan kemiskinan. Badan Pusat Statistik (BPS) pada maret 2018 mencatat peresentase jumlah penduduk miskin secara nasional sebesar 9,82% yang artinya sekitar 25,95% juta orang dari total penduduk Indonesia berada pada kategori miskin. Jumlah ini mengalami penurunan sebesar 633,2 ribu orang jika dibandingkan dengan September 2017 yang sebesar 26,58 juta orang atau sekitar 10,12%. Salah satu yang menjadi penyumbang terbesar jumlah penduduk miskin berada di kawasan pedesaan yaitu sebesar 15,81 juta orang, jumlah ini menurun 505 ribu orang dibandingkan pada periode September 2017[1].

Data yang diperoleh dari Badan Pusat Statistik (BPS) berasal dari Survei Sosial Ekonomi Nasional atau yang disingkat SUSENAS, kegiatan ini dilaksanakan setiap semester pada bulan Maret dan September. Secara konsep, untuk mengukur tingkat kemiskinan yaitu dengan cara perhitungan kemampuan seseorang untuk memenuhi kebutuhan dasar atau *basic needs approach* yang diukur dari sisi pengeluaran oleh BPS. Metode ini biasanya digunakan untuk menghitung jumlah penduduk yang hidup dibawah Garis Kemiskinan (GK)[1].

Karakteristik setiap wilayah yang berbeda membuat garis kemiskinan (GK) juga dibagi menurut kondisi geografis baik secara *administratif* maupun menurut jenis wilayahnya. Secara *administratif* garis kemiskinan di setiap daerah sebenarnya berbeda hal ini disesuaikan dengan karakteristik daerah tersebut. Hal lainnya yang dapat dilihat adalah *share* komoditi terhadap garis kemiskinan. Secara nasional, Garis Kemiskinan Makanan (GKM) memberikan kontribusi yang paling besar terhadap total garis kemiskinan yaitu sebesar 73,48%. Beras serta rokok masih menjadi penyumbang terbesar terhadap garis kemiskinan untuk di daerah perkotaan maupun di pedesaan[1].

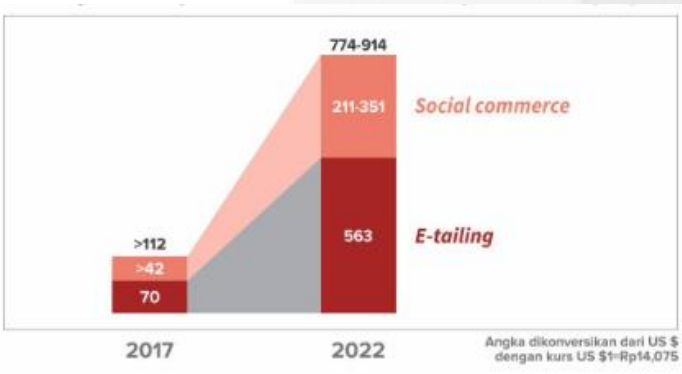
Survei Sosial Ekonomi Nasional (Susenas) yang diselenggarakan oleh BPS atau Badan Pusat Statistik merupakan salah satu sumber informasi mengenai gambaran kondisi sosial ekonomi masyarakat. Data ini digunakan untuk memperoleh indikator pencapaian kesejahteraan rakyat, indikator tersebut meliputi angka partisipasi sekolah dan angka melek huruf untuk bidang pendidikan, pemberian ASI pada baduta serta imunisasi untuk bidang kesehatan, dll[2].

Metode pengumpulan data yang dilakukan oleh BPS atau Badan Pusat Statistik pada tahun 2018 yaitu dengan menggunakan

pengumpulan data Susenas Kor yang dilaksanakan pada bulan Maret 2018. Jumlah total sampel Susenas Kor sebanyak 300.000 rumah tangga dalam 34 provinsi seluruh Indonesia dengan cara melakukan wawancara tatap muka antara pencacah dengan responden, keterangan tentang rumah tangga dikumpulkan melalui wawancara dengan kepala rumah tangga, suami/istri kepala rumah tangga, atau anggota rumah tangga lain yang mengetahui kerakteristik yang ditanyakan[2].

Pengolahan datanya meliputi tahap perekaman data, pemerikasaan konsistensi antar isisan dalam kuesioner sampai dengan tahap tabulasi, sepenuhnya dilakukan dengan menggunakan komputer. Sebelum tahap ini dimulai, terlebih dahulu dilakukan pengecekan awal atas kelengkapan isisan daftar pertanyaan, penyuntingan terhadap isisan yang tidak wajar, termasuk hubungan keterkaitan (konsistensi) antara satu jawaban dengan jawaban yang lainnya. Proses perekaman data dilakukan di BPS kota/kabupaten[2].

Pada paparan diatas maka dapat disimpulkan bahwa masalah kemiskinan melibatkan banyak hal, sehingga membutuhkan berbagai sisi untuk dapat melihat kemiskinan secara lebih mendalam dan akurat. Dalam hal ini juga pada proses pengumpulan data menggunakan metode survei yang dilakukan oleh Badan Pusat Statistik (BPS) membutuhkan waktu yang cukup lama serta tahapan yang dilakukan untuk mendapatkan data yang *valid* terbilang rumit, serta biasanya kepala rumah menghindar saat diwawancarai dikarenakan merasa takut akan penipuan sehingga metode lain untuk melengkapi hasil survei dan sensus yang diusulkan peneliti untuk memprediksi kemiskinan adalah dengan mengunakan *machine learning logistic regression* dengan metode *sparse learning based feature selection* baerbasis data *e-commerce* serta alasan mengapa penulis memilih data e-commerce sebagai data yang akan diolah untuk memprediksi tingkat kemiskinan di suatu daerah karena Indonesia merupakan negara yang memiliki pasar *e-commerce* terbesar di Asia Tenggara dengan kontribusi sekitar 50% dari seluruh transaksi. Pada kontribusi ini berpotensi terus meningkat lantaran penduduk Indonesia yang sering menggunakan internet. Pada tahun 2018, firma konsultan manajemen McKinsey dan Company merilis hasil riset mengenai status industri *e-commerce* Indonesa terkini dan proyeksi perkembangannya pada beberapa tahun kedepan. Pada Temuan – temuan mereka meliputi pertumbuhan nilai pasar e-commerce Indonesia hingga tahun 2022 dan potensi dampak pertumbuhan terhadap ekonomi dan sosial Indonesia. Berikut merupakan gambaran prediksi peningkatan data *e-commerce* di Indonesia[3].



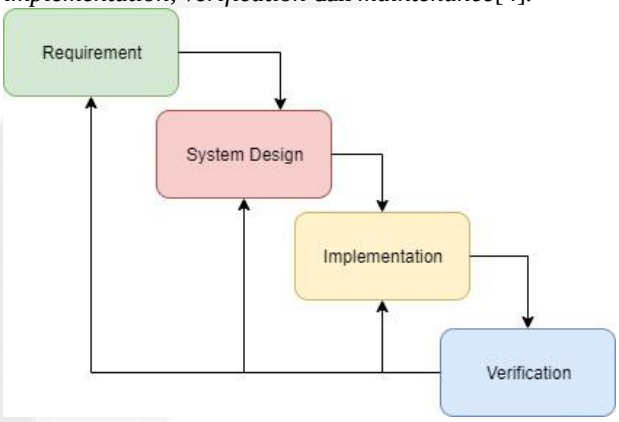
Gambar I-1 Prediksi Peningkatan *E-Commerce*

E-commerce diperkirakan tumbuh hingga US$65 miliar (Rp910 triliun), studi McKinsey mendefinisikan *e-commerce* sebagai proses jual beli barang fisik secara *online* yang dibagi menjadi 2 kategori, yaitu yang pertama adalah *E-tailing* dimana jual beli formal melalui *platform online* yang didesain untuk memfasilitasi transaksi seperti Bukalapak dan Tokopedia, kemudian yang kedua adalah *Social Commerce* yang diamana pemasaran barang melalui media sosial seperti Facebook atau Instagram dengan pembayaran dan pengiriman dilaksanakan di *platform* lain. Menurut McKinsey, pasar *e-commerce* Indonesia diprediksikan meningkat sekitar delapan kali lipat pada tahun 2022[3].

II. METODE PENELITIAN

Berikut merupakan metode penelitian dari dari aplikasi untuk memprediksi kemiskinan berbasis data *e-commerce* menggunakan algoritma *logistic regression* dan *sparse learning bases feature selection* :

1. Penentuan Topik
   Pada tahap ini penulis pertama-tama menentukan topik yang nantinya akan dibuat sebuah aplikasi untuk menyelesaikan proyek akhir di semester 6 mendatang dengan mengangkat judul Aplikasi untuk meprediksi kemiskinan berbasis *e-commerce* menggunakan *machine learning logistic regression* dengan metode *sparse learning based feature selection* yang nantinya akan menghasilkan proyeksi prediksi kemiskinan sesuai dengan data yang dimasukkan.
2. Identifikasi Masalah
   Pada tahap ini penulis mengidentifikasi masalah – masalah yang dihadapi oleh pemerintah Indonesia mengenai prediksi tingkat kemiskinan.
3. Studi Literatur
   Pada pembuatan laporan proyek akhir ini untuk menentukan fakta-fakta apa saja yang terkait dengan metode-metode yang digunakan untuk dapat memprediksi tingkat kemiskinan di Indonesia menggunakan data *e-commerce*.
4. Perancangan Sistem
   Pada proses perancangan penulis memilih menggunakan metode *waterfall* yang dimana metode ini akan mengambarkan pengembangan perangkat lunak yang dimulai dari tahap *requirement*, *design*, *implementation*, *verification* dan *maintenance*[4].



Gambar II-1 Metode Waterfall[4]

4.1 Requirement Analysis

Tahap requirement bertujuan untuk memahami perangkat lunak yang akan dibuat sesuai dengan yang diharapkan oleh pengguna, penulis dan batasan perangkat lunak tersebut. Adapun cara yang ditempuh penulis untuk mendapatkan informasi pengguna adalah dengan melakukan wawancara terhadap calon pengguna serta tinjauan pustaka dengan mencari referensi buku, web, jurnal yang berhubungan

dengan perangkat lunak yang akan dibangun.

4.2 System Design

Pada tahap ini, penulis mengunakan bahasa pemrograman *python* untuk membuat aplikasi, merancang sistem menggunakan tools BPMN untuk menggambarkan proses bisnis yang ada saat ini (*AS IS*) serta proses bisnis yang akan datang (*TO BE*), serta merancang pembentukan database yang menggunakan ERD untuk menentukan entitas serta *atribut* yang digunakan.

4.3 Implementation

Pada perancangan ini, penulis menggunakan metode model persamaan *logistic regression* dimana model persamaan aljabar layaknya OLS (*Ordinary Least Squares*) yang biasa digunakan adalah sebagai berikut: Y = B0 + B1X + e. Dimana e adalah error varians atau residual. Dengan model regresi ini, tidak menggunakan interpretasi yang sama seperti halnya persamaan regresi OLS. Model Persamaan yang terbentuk berbeda dengan persamaan OLS.

Berikut persamaannya Regresi Logistik[5]:

$$In\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X$$

Keterangan :
In = logaritma natural
B_0= konstanta
B_1= koefisien masing-masing variable
X = variable independen
$\hat{p}$ = probabilitas logistik yang dirumuskan sebagai berikut :

$$\hat{p} = \frac{\exp(B_0+B_1X)}{1+\exp(B_0+B_1X)} = \frac{e^{B_0+B_1X}}{1+e^{B_0+B_1X}}$$

Keterangan :
exp atau e : fungsi exponent.

4.4 Verification

Pada tahap ini seluruh unit yang dikembangkan pada tahap implementasi kemudian masuk ke dalam tahap pengujian dengan menggunakan metode *blackbox testing* yang dimana untuk memastikan apakah sitem telah berjalan sesuai dengan yang diinginkan ataukah sistem mengalami *eror*. Selain menggunakan metode *blackbox testing* penulis juga menggunakan UAT (*User Acceptance Testing*) yang digunakan untuk memastikan apakah aplikasi telah sesuai dengan keinginan user.

III. TINJAUAN PUSTAKA

Berikut merupakan beberapa teori pokok pembahasan yang sesuai dengan aplikasi yang dibangun dalam proyek akhir ini.

A. Machine Learning

Sebuah aplikasi yang menggunakan *Artificial Intelligence* (AI) yang menyediakan sistem kinerja secara otomatis serta belajar memperbaiki dari pengalaman tanpa diprogram secara eksplisit. Pembelajaran mesin ini berfokus pada pengembangan program computer yang dapat mengakses data serta menggunakannya untuk belajar sendiri. Sistem pembelajaran mesin terdiri dari tiga bagian utama, yaitu :

a) Model : sistem yang membentuk prediksi atau identifikasi
b) Parameter : sinyal atau faktor yang digunakan oleh model untuk membentuk keputusan
c) Pembelajaran : sistem yang menyesuaikan parameter dan model dalam prediksi versus hasil aktual[6].

Pada definisi lainnya disebutkan *machine learning* adalah cabang aplikasi AI (*Artificial Intelligence*) atau biasanya disebut sebagai suatu kecerdasan buatan. *Machine learning* dalam belajar harus membutuhkan data sebagai acuan untuk belajar terlebih dahulu dalam mengeluarkan sebuah *output*, tanpa data *machine learning* tidak dapat bekerja atau berfungsi dengan baik, terdapat tiga model *machine learning* diantaranya adalah sebagai berikut[7]:

1. Model *Supervised Learning*
   Model ini biasanya disebut sebagai model terarah karena umumnya diberikan instruksi yang jelas seperti apa saja yang perlu dipelajari serta bagaimana cara tersebut dapat dipelajari, biasanya pada model ini digunakan untuk memprediksi masa depan berdasarkan historis. *Supervised learning* dibagi menjadi dua yaitu :
   a) Classification
   Pada metode ini biasanya paling umum digunakan pada data mining, yang dimana setiap *atribut* atau fitur dalam metode ini harus diberikan label supaya komputer dapat mengetahui atau mengklasifikasikan sebuah objek dengan menggunakan label tersebut.
   b) Regresion
   Dalam metode ini tidak jauh beda dengan metode *classification* namun pada metode ini digunakan untuk membuat sebuah pola pada setiap atributnya yang bertujuan untuk mencari sebuah pola serta menentukan sebuah nilai numerik.
2. Model *Unsupervised Learning*

   Pada metode ini tidak deiberi label akan tetapi secara otomatis dibagi berdasarkan kemiripan serta struktur lain dari data tersebut.

B. *Feature Selection*

Feature selection adalah suatu kegiatan yang bisa dilakukan secara preprocessing yang bertujuan untuk memilih *feature* yang berpanguh maupun yang tidak berpengaruh dalam suatu penganalisaan data, terdapat dua kelompok yang ada pada *feature selection* diantaranya yaitu sebagai berikut[8]:

1. Rangking Selection

Pada *rangking selection* bertujuan untuk memberikan rangking pada setiap fitur yang ada serta tidak memperdulikan fitur yang tidak memenuhi standar serta tidak dirasa penting.

2. Subset Selection

Pada *subset selection* digunakan untuk mencari tahu sesuatu dari fitur yang dianggap sebagai optimal fitur. Terdapat tiga jenis metode pada *subset selection* diantaranya adalah sebagai berikut[8]:

a) *Feature selection* tipe *wrapper*

Pada tipe ini melakukan fitur seleksi dengan tahapan pemilihan secara bersamaan dengan cara

pelaksanaan pemodelan.

b) *Feature selection* tipe *filter*

Pada tipe ini dialakukan dengan cara memanfaatkan salah satu dari beberapa jenis filter yang ada.

c) Feature selection embedded

Pada tipe ini memanfaatkan suatu *machine learning* dalam proses seleksi fitur dan dalam sistem ini apabila mesin pembelajaran menganggap fitur tersebut tidak berpengaruh maka secara otomatis dihilangkan.

C. *Logistic Regression*

Sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear atau yang biasa disebut dengan istilah *Ordinary Least Squares (OLS) regression*. Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel terikat yang berskala dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah.

Model persamaan aljabar layaknya OLS yang biasa digunakan adalah sebagai berikut: Y = B0 + B1X + e. Dimana e adalah *error varians* atau *residual*. Dengan model regresi ini, tidak menggunakan interpretasi yang sama seperti halnya persamaan regresi OLS. Model Persamaan yang terbentuk berbeda dengan persamaan OLS[5].

Berikut persamaannya Regresi Logistik :

$$In\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X$$

Keterangan :

In = logaritma natural

B_0 = konstanta

B_1 = koefisien masing-masing variable

X = variable independen

$\hat{p}$ = probabilitas logistik yang dirumuskan sebagai berikut :

$$\hat{p} = \frac{\exp(B_0+B_1X)}{1+\exp(B_0+B_1X)} = \frac{e^{B_0+B_1X}}{1+e^{B_0+B_1X}}$$

Keterangan :

exp atau e : fungsi exponen.

D. *Data Mining*

Secara teknis, *data mining* adalah suatu proses yang memanfaatkan teknik-teknik sattistik, matematika serta kecerdasan buatan yang digunakan untuk mengekstrak dan mengidentifikasi informasi dan *knowledge* (pola-pola yang *valid*,baru,memiliki potensi dan mudah dipahami) yang berasal dari sekumplan data yang besar, di dalam proses *data mining* terdiri dari banyak langkah perulangan yang rumit, artinya bahwa banyak suatu dugaan/kesimpulan atau pencarian yang berbasis eksperimentasi yang dilibatkan[9].

E. *Library Python Scikit-learn*

Scikit-learn adalah suatu *library* untuk *machine learning* yang merupakan *free software* dan memungkinkan melakukan beragam pekerjaan dalam *data science*, seperti regresi (*regression*), kalsifikasi (*classification*), pengelompokan atau penggugusan (*clustering*), data *preprocessing*, *dimentionality reduction*, dan model *selection* (pembandingan, validasi, serta pemilihan parameter maupun model)[10].

F. *Library Pyhon Scikit-Feature*

Scikit-feature adalah suatu repositori pemilihan fitur *open-source* dengan *python* yang berisi sekitar 40 algoritma pemilihan fitur, termasuk algoritma pemilihan fitur tradisional dan beberapa algoritma pemilihan fitur *structural* serta *streaming*. *Scikit-feature* ini berfungsi sebagai *platform* untuk memfasilitasi apliakasi pemilihan sebuah fitur, penelitian serta studi banding. Saat ini *scikit-feature* terdiri dari beberapa algoritma seperti *similarity based feature selection*, *information theoretical based feature selection*, *sparse learning based feature selection*, *statistical based feature selection*, *wrapper based feature selection*, *structural feature selection*, *streaming feature selection*[11]**.**

G. *Normalisasi*

Normaliasai dalam proses *data mining* adalah suatu proses penskalaan nilai atupun *atribut* fitur dari sebuah data yang akan dinormalisasi sehingga data tersebut dapat memiliki skala atau *range* yang telah ditetapkan sebelumnya. Terdapat beberapa metode dalam proses normalisasi diantaranya adalah min-max, z-score, *decimal scaling*, sigmoidal, dll. Metode *min-max* digunakan untuk transformasi linier terhdap data aslinya. Metode *z-score* adalah normalisasi yang berdasarkan nilai rata-rata atau sering disebut dengan *mean* dan *standart deviation* (deviasi standar) dari data. Metode *decimal scaling* adalah suatu metode normalisasi yang menggerakkan nilai desimal dari suatu data. Metode *sigmoidal* adalah suatu metode normalisasi yang secara nonlinier kedalam *range* -1-1 dengan menggunakan fungsi *sigmoid*, metode ini berguna untuk data yang melibatkan *data outlier* (data yang jauh dari jangkauan data lainnya)[12].

H. Bahasa Pemrograman *Python*

Salah satu bahasa pemrograman yang populer di dunia kerja Indonesia, *phyton* secara default telah terpasang di beberapa sistem operasi berbasis Linux seperti Ubuntu, Linux Mint, Fedora. Untuk sistem operasi lain, sudah tersedia *Installer* yang disediakan untuk sistem operasi tersebut. Selain itu *phyton* memiliki sebuah *package manager* yang popular dan unggul bernama PIP. Dengan PIP pengguna dapat menghapus atau memasang pustaka *phyton*[13].

I. RMSE (*Root Mean Square Error*)

RMSE adalah sebuah metode yang digunakan untuk mengukur tingkat akurasi hasil prakiraan suatu model. RMSE digunakan untuk menilai rata-rata dari jumlah kuadrat kesalahan dan juga dapat menunjukkan ukuran besarnya kesalahan dari prakiraan yang dihasilkan oleh suatu model[14].

J. R Squared

R2 adalah ukuran statik diantara 0 – 1 yang berfungsi untuk menghitung seberapa mirip hasil prediksi dengan data aslinya yang digambarkan dengan titik-titik yang mendekati atau mengikuti garis regresi[15].

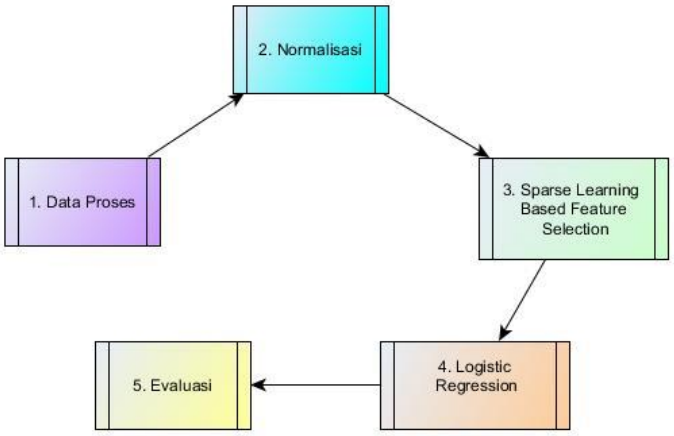
IV. ANALISIS DAN PERANCANGAN

A. Gambaran Sistem Usulan

Berikut merupakan gambaran sistem usulan dari *preprocessing* hingga sistem usulan aplikasi.

1. Gambaran Sistem Usulan Alur Pengembangan Model

Berikut merupakan gambaran usulan alur pengembangan model.



Gambar IV-1 Sistem Usulan alur pengembangan model

Pada paparan gambar diatas terdapat beberapa proses sebelum data *e-commerce* diolah di dalam aplikasi untuk memprediksi kemiskinan berbasis data *e-commerce* menggunakan algoritma *logistic regression dan sparse learning based feature selection*.

a. Proses Data
Pada proses ini pertama data akan di *cleaning* atau *preprocessing* data, proses ini sangat penting karena data yang belum lengkap atau *missing value* akan mengakibatkan data tidak siap untuk diproses. Pada data *e-commerce* yang didapatkan dari salah satu perusahaan e-commerce di Indonesia ternyata terdapat *missing value* dimana data masih ada yang *null*, sehingga data yang *null* diganti dengan 0, setelah itu data akan ditentukan mana yang merupakan fitur serta mana yang label.

b. Normalisasi
Setelah dilakukannya *data process*, maka data akan masuk ke dalam proses normalisasi, hal ini dilakukan untuk menskalakan nilai data dari 0-10. Alasan mensklalakan nilai data dari 0-10 karena jika range nya 0-1, maka nilai dibelakang koma semakin panjang dan machine learningnya akan semakin sulit untuk di training karena nilainya terlalu kecil. Pada proses nomalisasi ini penulis menggunakan metode *Rescaling (min-max normalization)*. Berikut merupakan rumus dasarnya.

$$MinMax = \frac{x - \min(x)}{\max(x) - \min(x)} * 10$$

Dengan keterangan sebagai berikut :
x = Nilai dari masing-masing fitur.
min(x) = Nilai terendah dari setiap fitur.
max(x) = Nilai tertinggi dari setiap fitur.

c. *Sparse Learning Based Feature Selection*
Pada tahap ini setelah proses normalisasi, data akan masuk ke dalam proses seleksi fitur yaitu *sparse learning based feature selection* dimana ini bertujuan untuk meminimalkan kesalahan yang terjadi, *sparse regularizer* memaksa kosfisien fitur menjadi kecil atau persis nol serta kemudian fitur yang sesuai dapat dihilangkan dengan mudah. Terdapat beberapa metode pemilihan fitur baik perspektif yang diawasi maupun yang tidak diawasi diantaranya yaitu ll_l21, ls_l21, REFS. Pada pemilihan fitur REFS (*Efficient and Robust Feature Selection*) ini berfungsi untuk klasifikasi multi-kelas, berikut merupakan rumusnya.

$$\min_{\mathbf{w}} \| \mathbf{XW} - \mathbf{Y}||_{2,1} + \mathbf{a} \| \mathbf{W}||_{2,1}$$

Pada seleksi fitur ll_l21 dan ls_l21 sama-sama berasal dari rumus seleksi fitur *p, q-Norm Regularizer*, dab berikut merupakan rumus dari ll_l21 (*logistic loss* (logistic l21)).

$$\min_{W,c} \sum_{i=1}^{t} \sum_{j=1}^{n_i} log(1 + exp(-Y_{ij}(W_j^T X_{ij} + c_i))) + p_1 \| W \|_{2,1} + p_{l2} \| W \|_F^2$$

Dimana $X(i,j)$ menunjukan sampel dari j ke i, $Y(i,j)$ menunjukan label yang sesuai, Wi dan ci untuk model i, parameter regularisasi atau pi digunakan untuk mengontrol sparsitas grup. Sedangkan rumus untuk ls_l21 (least-*squares loss* (least l21)) adalah

$$\min_{W} \sum_{i=1}^{t} \| W_i^T X_i - Y_i \|_F^2 + p_1 \| W \|_{2,1+pL_2} \| W \|_F^2$$

Dimana x_i menunjukkan matriks input dari i, y_i menunjukkan label yang sesuai, *wi* adalah model untuk i.

d. *Logistic Regression*
Pada tahap ini data yang telah keluar dari proses *sparse learning based feature selection* maka akan masuk ke dalam proses selankutnya yaitu proses algoritma *logistic regression* yang dimana data akan diolah dengan sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linier atau yang sering disebut dengan istilah OLS (*Ordinary Least Squares*) *regression*. Dengan model persamaan *logistic regression* sebagai berikut :

$$In\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X$$

Keterangan :
In = logaritma natural
B_0 = konstanta
B_1 = koefisien masing-masing variable
X = variable independen
$\hat{p}$ = probabilitas logistik yang dirumuskan sebagai berikut :

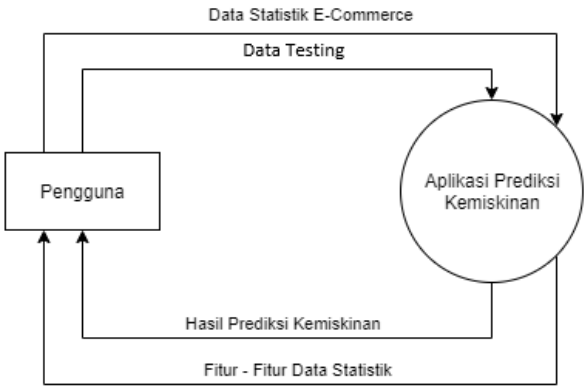
$$\hat{p} = \frac{\exp(B_0 + B_1 X)}{1 + \exp(B_0 + B_1 X)} = \frac{e^{B_0 + B_1 X}}{1 + e^{B_0 + B_1 X}}$$

Keterangan :
exp atau e : fungsi exponent.

e. Evaluasi
Pada tahap ini terdapat matrik kinerja untuk regresi. Terdapat 2 matrik yakni R^2 (*R-squared*) dan RMSE dimana R^2 mewakili bagian dari *varians vector* yang nantinya dapat diprediksi oleh model regresi, jika R^2 =1 maka model regresi dapat dikatakan benar dalam memprediksi nilai, dan jika sebaliknya R^2 = negatif maka regresi dapat dikatakan salah dalam hasil memprediksi sebuah nilai. Sedangkan RMSE digunakan untuk mengukur kesalahan dan atau perbedaan antara *vector actual* dengan prediksi. Jika nilai RMSE lebih tinggi maka lebih banyak perbedaan antara nilai *actual* dengan nilai prediksi.

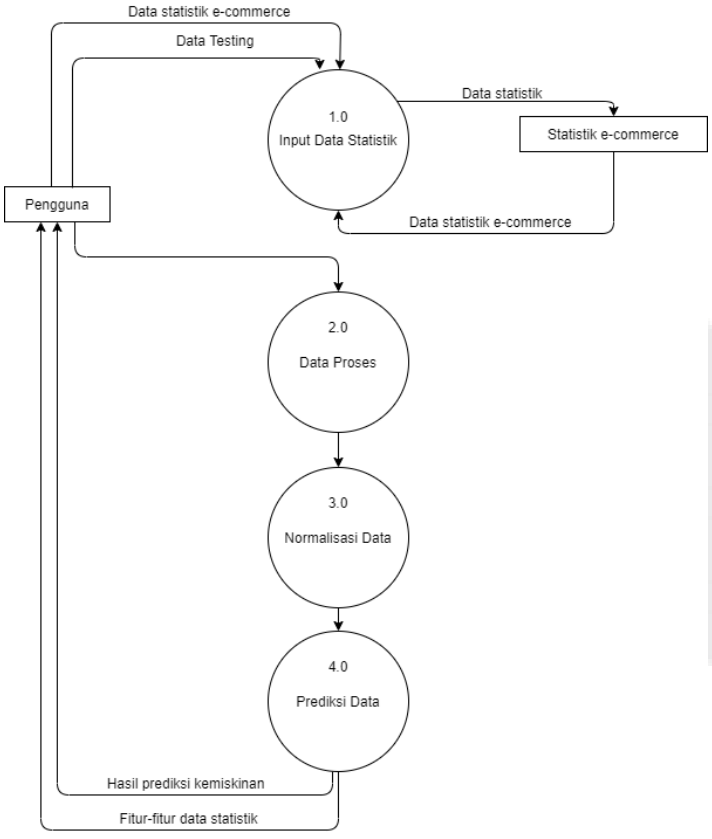
2. Gambaran Sistem Usulan Aplikasi
Berikut merupakan gambaran sistem usulan pada aplikasi

untuk memprediksi kemiskinan berbasis data *e-commerce* menggunakan algoritma *logistic regression* dan *sparse learning bases feature selection* yang digambarkan dengan menggunakan *data flow diagram*. alasan mengapa penulis menggunakan DFD karena pada aplikasi ini termasuk ke dalam *structural programing* yang dimana pemrograman bertumpu pada pemanggilan *library* yang telah didefinisikan sebelumnya serta DFD dapat menggambarkan fungsi, proses, penangkapan data, memanipulasi, menyimpan serta mendistribusikan data antara sauatu sistem pada lingkungannya serta antara komponen-komponen suatu sistem.



Gambar IV-2 Diagram Konteks

Pada gambaran aplikasi yang diusulkan di atas, pengguna menggunakan data statistik e-commerce untuk memprediksi kemiskinan, kemudian data diolah oleh aplikasi prediksi kemiskinan dan kemudian pengguna akan mendapatkan hasil dari pengolahan data berupa prediksi kemiskinan. Berikut merupakan gambaran alur jalannya aplikasi prediksi kemiskinan.



Gambar IV-3 DFD Level 1 Proses Prediksi Kemiskinan

Pada proses alur aplikasi prediksi kemiskinan di atas pengguna menggunakan data statistik yang nantinya data tersebut akan masuk ke dalam proses input data statistik, setelah di olah data akan keluar sebagai *data statistic* yang nantinya akan disimpan sebagai statistik *e-commerce* yang kemudian data statistik masuk kembali ke dalam proses input data statistik yang akan diolah dan dihasilkan sebuah dataset yang nantinya akan digunakan pengguna untuk masuk ke dalam proses selanjutnya yaitu proses normalisasi data dan kemudian data akan masuk ke dalam proses prediksi data , di dalam proses prediksi data, data akan diolah dan menghasilkan hasil prediksinya

V. IMPLEMENTASI DAN PENGUJIAN

A. Implementasi

Selah melalui tahap Analisis dan Perancangan, berikut adalah tahap implementasi dari aplikasi untuk memprediksi kemiskinan berbasis data *e-commerce* menggunakan algoritma *logistic regression* dan *sparse learning bases feature selection*.

a) Halaman Registrasi.

Gambar V-1 Halaman Registrasi

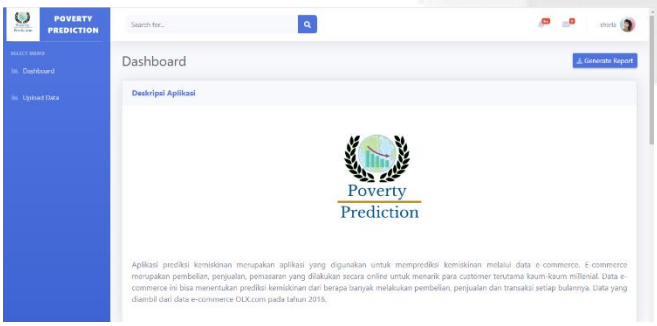
Pada halaman registrasi digunakan untuk membuat akun jika user belum memiliki akun sebelumnya. Di dalam halaman registrasi user harus menginputkan *first name, last name, email, username, password, repeat password*.

b) Halaman Login

Gambar V-2 Halaman Login

Setelah user telah memiliki akun, maka user tersebuat dapat masuk ke dalam halam utama dengan memasukkan username serta password sebelumnya di halaman login ini. Jika user belum memiliki akun, maka dapat menekan tombol *create account* yang digunakan untuk mendaftarkan diri.

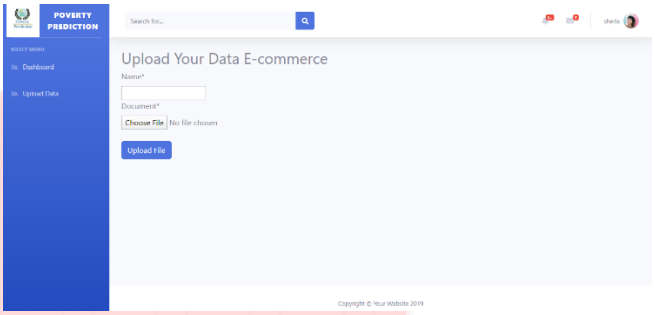
c) Halaman Dashboard



Gambar V-3 Halaman Dashboard

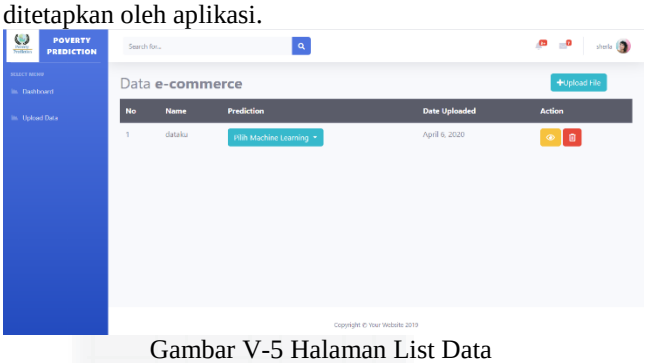
Pada tampilan menu utama ini user akan disediakan 2 pilihan, yang pertama adalah *dashboard* yang digunakan untuk melihat deskripsi yang berhubungan dengan aplikasi serta jenis-jenis *machine learning* yang akan digunakan untuk memprediksi suatu kemiskinan di suatu daerah.

d) Halaman Upload Data dan Halaman List Data



Gambar V-4 Halaman Upload Data

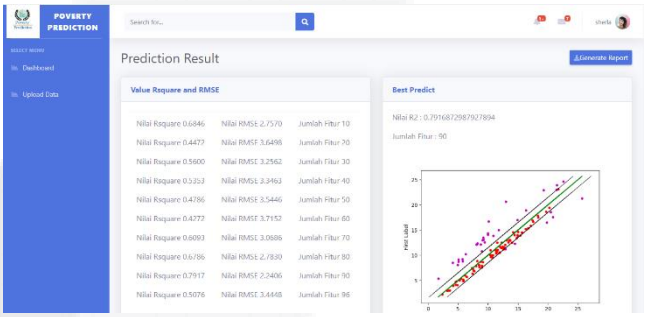
Pada halaman upload, user dapat meng-*upload* data *e-commerce* yang dimiliki sesuai dengan *template* yang telah ditetapkan oleh aplikasi.



Gambar V-5 Halaman List Data

Setalah user meng-*upload* data *e-commerce* maka akan tampil list data yang telah diinputkan sebelumnya.

e) Halaman Hasil Prediksi



Gambar V-6 Halaman Hasil Prediksi

Setelah user memilih machine leraning yang diinginkan, maka akan muncul halaman hasil prediksi yang berisi tentang hasil prediksi.

B. Pengujian

Setelah melaui tahap implementasi maka tahap berikutnya adalah pengujian dari aplikasi. Berikut merupakan pengijian menggunakan metode *black box testing*.

1) Pengujian Login

Pengujian dilakukan untuk menguji kesesuaian antara Fungsionalitas login dengan spesifikasi kebutuhan pengguna. Berikut merupakan *scope of testing login*.

Tabel V-1 Scope of Testing Login

Perangkat Lunak	Aplikasi Untuk Memprediksi Kemiskinan Berbasis Data E-Commerce Menggunakan Algoritma Logistic Regression Dan Sparse Learning Based Feature Selection.
Deskripsi	Aplikasi ini dapat memprediksi tingkat kemiskinan disuatu daerah.
Fungsi	Login
Aturan	**1. Username dan password harus diisi sesuai dengan data registrasi** 2. Username dan password tidak diisi sesuai dengan data registrasi 3. Username dan password tidak diisi

Tabel V-2 Test Case Matrix Function Login

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution/ Kesimpulan
1.	Login	1.	Entry data login dengan mengikuti aturan 1 : 1. Username 2. Password	Username dan Password sesuai dengan data registrasi	Aplikasi menampilkan halaman *dashboard*	Aplikasi menampilkan halaman *dashboard*	Valid
		2.	Entry data login dengan mengikuti aturan 2 : 1. Username 2. Password	Username dan Password tidak sesuai dengan data registrasi	Aplikasi menampilkan Pesan error message "invalid creditials"	Aplikasi menampilkan Pesan error message "invalid creditials"	Valid
		3.	Entry data login dengan mengikuti aturan 3 : 1. Username 2. Password	Username dan Password dikosongkan	Aplikasi menampilkan Pesan error message "invalid creditials"	Aplikasi menampilkan Pesan error message "invalid creditials"	Valid

2) Pengujian Registrasi
Pengujian dilakukan untuk menguji kesesuaian fungsionalitas registrasi dengan spesifikasi kebutuhan pengguna. Berikut merupakan *scope of testing registrasi.*

Tabel V-3 Scope of Testing Registrasi

Perangkat Lunak	Aplikasi Untuk Memprediksi Kemiskinan Berbasis Data E-Commerce Menggunakan Algoritma Logistic Regression Dan Sparse Learning Based Feature Selection.
Deskripsi	Aplikasi yang digunakan untuk memprediksi tingkat kemiskinan disuatu daerah.
Fungsi	Registrasi

Aturan	1. First name, last name, email, username, password dan repeat password tidak diisi 2. **First name, last name, email, username, password dan repeat password harus diisi**		3. **Password dan repassword sama** 4. Password dan repassword tidak sama

Tabel V-4 Test Case Matrix Function Registrasi

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution/ Kesimpulan
1.	Registrasi	1.	Entry data login dengan mengikuti aturan 1 : 1. First Name 2. Last Name 3. Username 4. Email 5. Password 6. Repeat Password	Semua form diisi	Aplikasi menampilkan halaman login	Aplikasi menampilkan halaman login	Valid
		2.	Entry data login dengan mengikuti aturan 2 : 1. First Name 2. Last Name 3. Username 4. Email 5. Password 6. Repeat Password	Semua form tidak diisi	Aplikasi menampilkan Pesan error message "all form must be set"	Aplikasi menampilkan Pesan error message "all form must be set"	Valid
		3.	Entry data login dengan mengikuti aturan 3 : 1. First Name 2. Last Name 3. Username 4. Email 5. Password 6. Repeat Password	Password : Sherla123 Repeat Password : Sherla123	Aplikasi menampilkan halaman login	Aplikasi menampilkan halaman login	Valid
		4.	Entry data login dengan mengikuti aturan 4 : 1. First Name	Password : yualinda Repeat Password	Aplikasi menampilkan Pesan error	Aplikasi menampilkan Pesan error	Valid

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution/ Kesimpulan
			2. Last Name 3. Username 4. Email 5. Password 6. Repeat Password	: yualinda11	message "Password not maching"	message "Password not maching"	

3) Pengujian Upload File

Pengujian dilakukan untuk menguji kesesuaian fungsionalitas *upload* data dengan spesifikasi kebutuhan pengguna. Berikut merupakan *scope of testing* dari *upload file*.

Tabel V-5 Scope of Testing Upload File

Perangkat Lunak	Aplikasi Untuk Memprediksi Kemiskinan Berbasis Data E-Commerce Menggunakan Algoritma Logistic Regression Dan Sparse Learning Based Feature Selection.
Deskripsi	Aplikasi yang digunakan untuk memprediksi tingkat kemiskinan disuatu daerah.
Fungsi	Upload data
Aturan	1. *Name* dan *document* tidak diisi 2. *Name* dan *document* diisi

Tabel V-6 Test Case Matrix Function Upload Data

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution/ Kesimpulan
1.	Upload Data	1.	Entry data *upload data* dengan mengikuti aturan 1 : 1. Name 2. Document	Semua form diisi	Aplikasi menampilkan halaman *list document*	Aplikasi menampilkan halaman *list document*	Valid
		2.	Entry data *upload data* dengan mengikuti aturan 2 : 1. Name 2. Document	Semua form tidak diisi	Aplikasi menampilkan Pesan error message	Aplikasi menampilkan Pesan error message	Valid

4) Pengujian Forgot Password

Pengujian dilakukan untuk menguji kesesuaian fungsionalitas *forgot password* dengan spesifikasi kebutuhan pengguna. Berikut merupakan *scope of testing* dari *forgot password*.

Tabel V-7 Scope of Testing Forgot Password

Perangkat Lunak	Aplikasi Untuk Memprediksi Kemiskinan Berbasis Data E-Commerce Menggunakan Algoritma Logistic Regression Dan Sparse Learning Based Feature Selection.

Deskripsi	Aplikasi yang digunakan untuk memprediksi tingkat kemiskinan disuatu daerah.
Fungsi	Forgot Password
Aturan	1. **Email harus diisi** 2. Email dikosongkan 3. ***New Password dan new password confirmation* tidak boleh sama dengan data diri** 4. *New Password dan new password confirmation* sama dengan data diri 5. ***New Password dan new password confirmation* minimal 8 karakter** 6. *New Password dan new password confirmation* kurang dari 8 karakter 7. ***New Password dan new password confirmation* tidak boleh seluruhnya bersifat numerik** 8. *New Password dan new password confirmation* seluruhnya bersifat numerik 9. **New password dengan new password confirmation harus sama** 10. New password dengan new password confirmation tidak sama

Tabel V-8 Test Case Matrix Function Forgot Password

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution / Kesimpulan
1.	Forgot Password	1.	Entry data *forgot password* dengan mengikuti aturan 1 : 1. Email	Email diisi sesuai dengan akun registrasi, contoh : yualindasherli@gmail.com	Aplikasi menampilkan halaman *Password reset sent*	Aplikasi menampilkan halaman *Password reset sent*	Valid
		2.	Entry data *forgot password* dengan mengikuti aturan 2 : 1. Email	Email dikosongkan	Aplikasi menampilkan Pesan error message	Aplikasi menampilkan Pesan error message	Valid
		3.	Entry data *forgot password* dengan mengikuti aturan 3 : *1. New Password* 2. *New Password Confirmation*	*New Password* dan *New Password Confirmation* tidak boleh sama dengan data diri	Aplikasi menampilkan halaman *Password reset Completed*	Aplikasi menampilkan halaman *Password reset Completed*	Valid

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution / Kesimpulan
		4.	Entry data *forgot password* dengan mengikuti aturan 4 : 1. *New Password* 2. *New Password Confirmation*	*New Password* dan *New Password Confirmation* sama dengan data diri	Aplikasi menampilkan Pesan error message	Aplikasi menampilkan Pesan error message	Valid
		5.	Entry data *forgot password* dengan mengikuti aturan 5 : 1. *New Password* 2. *New Password Confirmation*	*New Password* dan *New Password Confirmation* minimal 8 karakter, contoh : ewerty123	Aplikasi menampilkan halaman *Password reset Completed*	Aplikasi menampilkan halaman *Password reset Completed*	Valid
		6.	Entry data *forgot password* dengan mengikuti aturan 6 : 1. *New Password* 2. *New Password Confirmation*	*New Password* dan *New Password Confirmation* kurang dari 8 karakter, contoh : Wer	Aplikasi menampilkan Pesan error message	Aplikasi menampilkan Pesan error message	Valid
		7.	Entry data *forgot password* dengan mengikuti aturan 7 : 1. *New Password* 2. *New Password Confirmation*	*New Password* dan *New Password Confirmation* tidak boleh seluruhnya bersifat numerik, contoh: ertyuiu123	Aplikasi menampilkan halaman *Password reset Completed*	Aplikasi menampilkan halaman *Password reset Completed*	Valid
		8.	Entry data *forgot password* dengan mengikuti aturan 7 : 1. *New Password*	*New Password* dan *New Password Confirmation* seluruhnya bersifat numerik, contoh: 1233455666	Aplikasi menampilkan Pesan error message	Aplikasi menampilkan Pesan error message	Valid

No.	Function/ Condition	Case No.	Test Case Description (Event)	Test Data (Input)	Expected Result	Actual Result / Comments	Conclution / Kesimpulan
			2. *New Password Confirmation*				
		9.	Entry data *forgot password* dengan mengikuti aturan 7 : *1. New Password* 2. *New Password Confirmation*	*New password* dengan *new password confirmation* harus sama, contoh : *New Password* : yualinda123 *New Password Confirmation* : yualinda123	Aplikasi menampilkan halaman *Password reset Completed*	Aplikasi menampilkan halaman *Password reset Completed*	Valid
		10.	Entry data *forgot password* dengan mengikuti aturan 7 : *1. New Password* 2. *New Password Confirmation*	*New password* dengan *new password confirmation* tidak sama, contoh : *New Password* : yualinda1234 *New Password Confirmation* : yualinda123	Aplikasi menampilkan Pesan error message	Aplikasi menampilkan Pesan error message	Valid

VI. KESIMPULAN

Dari implementasi dan pengujian yang telah di dilakukan dapat disimpulkan bahwa :

1. Aplikasi dapat menampilkan hasil prediksi kemisikinan di suatu daerah dengan menggunakan berbasis *logistic regression* dengan menggunakana algoritma *sparse learning based feature selection*.
2. Item yang berpengaruh di dalam aplikasi ini adalah mobil, apartement, dst.
3. Dapat mengembangkan aplikasi berbasis *web* yang menjalankan algoritma *machine learning* dan dilengkapi dengan grafik untuk menampilkan hasil prediksi.

REFERENSI

[1] F. M. FARUK, "Berkenalan Dengan Kemiskinan," *GEOTIMES*, 2018. [Online]. Available: https://geotimes.co.id/opini/berkenalan-dengan-kemiskinan/.

[2] B. pusat Statistik, "Statistik Kesejahteraan Rakyat 2018," *BADAN PUSAT STATISTIK*, 2018. [Online]. Available: https://www.bps.go.id/publication/2018/11/26/81ede2d56698c07d510f6983/statistik-kesejahteraan-rakyat-2018.html.

[3] D. Praditya, "Prediksi Perkembangan Industri E-commerce Indonesia pada Tahun 2022," *TECHINASIA*, 2019. [Online]. Available: https://id.techinasia.com/prediksi-ecommerce-indonesia.

[4] F. Galandi, "Metode Waterfall : Definisi, Tahapan, Kelebihan dan Kekurangan," *Pendidikan, Pengetahuan*, 2018. [Online]. Available: http://www.pengetahuandanteknologi.com/2016/09/metode-waterfall-definisi-tahapan.html.

[5] A. Hidayat, "Regresi Logistik," *statistikian*, 2015. [Online]. Available: https://www.statistikian.com/2015/02/regresi-logistik.html.

[6] PodFeeder, "Apa Itu Machine Learning," *PodFeeder*. [Online]. Available: http://www.podfeeder.com/teknologi/apa-itu-machine-learning-berikut-penjelasannya/.

[7] V. N. Drozdov, V. A. Kim, and L. B. Lazebnik, *Modern approach to the prevention and treatment of NSAID-gastropathy*, no. 2. 2011.

[8] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[9] M. K. Albert Verasius Dian Sano, S.T., "DEFINISI, KARAKTERISTIK, DAN MANFAAT DATA MINING -SERI DATA MINING FOR BUSINESS INTELLIGENCE (2)," *Binus University*, 2019. [Online]. Available: https://binus.ac.id/malang/2019/01/definisi-karakteristik-dan-manfaat-data-mining-seri-data-mining-for-business-intelligence-2/.

[10] A. Hakim, "Berkenalan dengan scikit-learn (Part 1) – Preparations," *hkaLabs*. [Online]. Available: https://hakim-azizul.com/berkenalan-dengan-scikit-learn/.

[11] J. Li, "Data Mining, Data Science, Feature Extraction, Feature Selection, Machine Learning, Python," *kdnuggets.com*, 2016. [Online]. Available: https://www.kdnuggets.com/2016/03/scikit-feature-open-source-feature-selection-python.html. [Accessed: 20-Sep-2003].

[12] M. K. NOVIANDI, "DATA MINING," 2018.

[13] R. Fajar, "Memulai Pemrograman dengan Python," *Codepolitan.com*, 2016. [Online]. Available: https://www.codepolitan.com/memulai-pemrograman-python.

[14] Kuliahkomputer, "Training dan Testing Ilmu Komputer 'Root Mean Square Error,'" *kuliahkomputer*, 2018. [Online]. Available: http://www.kuliahkomputer.com/2018/07/training-dan-testing-ilmu-komputer-root.html.

[15] A. Hershy, "Calculating R-squared from scratch (using python)," *Toward Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/r-squared-recipe-5814995fa39a.