Volume 6 No 1 Juni (2023)

e-ISSN: **2615-3270**

p-ISSN: **2615-3599**

Eigen Mathematics Journal

Homepage jurnal: http://eigen.unram.ac.id

Comparison of Data Mining Methods in Classifying Village Status of Purwakarta and West Bandung Regencies (Podes 2021)

Munifah Zuhra Almasah[a]*, Arie Wahyu Wijayanto[b]

[a]Politeknik Statistika STIS, Jl. Otto Iskandardinata 64c, Jakarta 11480, Indonesia. Email: 211911036@stis.ac.id
[b]Politeknik Statistika STIS, Jl. Otto Iskandardinata 64c, Jakarta 11480, Indonesia. Email: ariewahyu@stis.ac.id

A B S T R A C T

Each village has different characteristics and is constantly changing along with the level of development in a village. These changes in conditions are used as indicators to classify villages into urban or rural village status. In this study, researchers will compare or evaluate of several data mining methods, namely decision trees, support vector machines, naïve bayes, and random forests to find the best algorithm in classifying urban villages and rural villages in Purwakarta and West Bandung Regencies. The data used in this study were 357 records and 8 attributes sourced from village potential data (Podes 2021). Furthermore, it was obtained that the best method in classifying urban villages and rural villages is to use random forests with accuracy value and F- score of 0,9. BPS can consider the random forest method in classifying village status as the basis for development policies by local governments.

Keywords: Villages Status, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest

Diserahkan: 03-12-2022; Diterima: 26-05-2023; Doi: https://doi.org/10.29303/emj.v6i1.156

1. Pendahuluan

Secara administratif, wilayah Indonesia dibagi menjadi tingkat provinsi, kabupaten/kota, kecamatan, dan desa. Desa merupakan tingkat wilayah administratif terkecil di Indonesia yang sering kali dijadikan sebagai unit observasi. Setiap desa memiliki karakteristik yang berbeda-beda dan terus berubah seiring dengan tingkat pembangunan di suatu desa. Perubahan kondisi tersebut dijadikan sebagai indikator untuk mengklasifikasikan desa ke dalam status desa perkotaan atau desa perdesaan. BPS mengklasifikasikan status wilayah desa dengan melibatkan beberapa variabel yang telah ditetapkan berdasarkan Peraturan Kepala Badan Pusat Statistik Nomor 120 Tahun 2020 tentang Klasifikasi Desa Perkotaan dan Perdesaan di Indonesia. Kriteria desa perkotaan tahun 2020 merupakan penyempurnaan kriteria desa perkotaan tahun 2000 dengan tetap menggunakan tiga indikator sebagai ukurannya, yaitu kepadatan penduduk, persentase keluarga pertanian, dan akses untuk mencapai fasilitas perkotaan. Perbedaan dengan kriteria sebelumnya adalah tidak digunakan lagi variabel bioskop, perubahan kriteria pada rumah tangga telepon dan rumah tangga pengguna listrik (BPS, 2020).

Klasifikasi merupakan salah satu topik utama dalam *data mining*. Klasifikasi adalah menganalisis data menggunakan model yang menggambarkan kelas data (Aryani & Wijayanto, 2021). Klasifikasi desa perkotaan dan perdesaan dapat digunakan untuk perencanaan pembangunan wilayah yang mencakup

*Corresponding author.
Alamat e-mail: *211911036@stis.ac.id*

berbagai aspek dengan mempertimbangkan peran keterkaitan antara desa dan kota. Hal ini sejalan dengan rencana pembangunan pemerintah untuk membangun Indonesia dengan memperkuat daerah dan desa yang menyebar ke seluruh pelosok negeri (Tarigan, 2003 dalam Apriliansyah, et al., 2021). Dengan demikian, pemerintah perlu mengetahui status desa perkotaan dan desa perdesaan dalam merencanakan pembangunan daerah di wilayah desa. Selain itu, pengembangan desa secara berkelanjutan dapat mendukung pencapaian target dalam kurun waktu 2020-2024, yaitu terwujudnya desa berkembang dan mandiri, serta kolaborasi perdesaan dengan perkotaan (Kemendes, 2021).

Penelitian yang dilakukan oleh Sari, et al., (2014) bertujuan untuk mengklasifikasikan status desa perkotaan dan perdesaan dengan metode *Support Vector Machine* (SVM). Klasifikasi dengan SVM menghasilkan akurasi dengan nilai terbaik sebesar 90% menggunakan fungsi kernel *Radial Basis Function* (RBF) dengan parameter C = 100 dan $\gamma = 2^{-5}$. Kemudian penelitian dari Supartini, et al., (2017) menggunakan metode *k-fold cross validation* untuk menghitung akurasi dari fungsi diskriminan kuadratik dalam mengklasifikasikan desa di Kabupaten Tabanan. Selanjutnya, Apriliansyah, et al., (2021) menunjukkan bahwa dari 438 desa/kelurahan di Provinsi D.I. Yogyakarta, model *decision tree* mampu mengklasifikasi secara benar 392 desa sesuai dengan status desa sebelumnya dengan presisi sebesar 87,5% dan F1-*score* sebesar 87,95%. Sementara itu, pada penelitian ini akan dilakukan klasifikasi terhadap wilayah desa perkotaan dan desa perdesaan menggunakan beberapa metode klasifikasi *data mining* di Kabupaten Purwakarta dan Bandung Barat.

2. Metodologi

2.1. Metode Pengumpulan Data

Penelitian ini menggunakan data sekunder dari Pendataan Potensi Desa Provinsi Jawa Barat tahun 2021 yang dilakukan oleh Badan Pusat Statistik. Adapun jumlah unit analisis (desa/kelurahan) yang akan digunakan adalah sebanyak 357 desa yang tersebar di Kabupaten Purwakarta dan Kabupaten Bandung Barat. Variabel yang akan digunakan pada penelitian ini adalah sebagai berikut.

Tabel 1 – Variabel yang digunakan.

Variabel	Deskripsi Variabel	Skala Pengukuran
Status Desa	1 : Desa Perkotaan 0 : Desa Perdesaan	Nominal
TK	1 : Ada atau ≤ 2,5 km 0 : > 2,5 km	Nominal
SMP	1 : Ada atau ≤ 2,5 km 0 : > 2,5 km	Nominal
SMA	1 : Ada atau ≤ 2,5 km 0 : > 2,5 km	Nominal
Pasar Desa	1 : Ada atau ≤ 2 km 0 : > 2 km	Nominal
Pertokoan	1 : Ada atau ≤ 2 km 0 : > 2 km	Nominal
Rumah Sakit	1 : Ada atau ≤ 5 km 0 : > 5 km	Nominal
Hotel	1 : Ada 0 : Tidak Ada	Nominal
Pub	1 : Ada 0 : Tidak Ada	Nominal
Salon	1 : Ada 0 : Tidak Ada	Nominal
Telepon kabel	Persentase keluarga pengguna telepon kabel	Rasio
Listrik PLN	Persentase keluarga pengguna listrik PLN	Rasio

2.2. Metode Analisis

Tahapan analisis yang dilakukan pada penelitian mengikuti siklus bisnis *Cross-Industry Standard Process for Data Mining* (CRISP-DM) berikut ini .

Gambar. 1 – Alur CRISP-DM *Data Mining*

a. Business understanding

Memahami proses data secara komprehensif. Selanjutnya mengubah pengetahuan menjadi definisi permasalahan *data mining* dan rencana awal proyek yang dirancang untuk mencapai tujuan (Wirth & Hipp, 2000).

b. Data understanding

Tahapan ini berfokus pada pemahaman awal mengenai data yang dibutuhkan untuk memecahkan permasalahan bisnis yang diberikan atau mendeteksi himpunan bagian yang menarik untuk membentuk hipotesis bagi informasi tersembunyi (Jackson, 2002).

c. Data preparation

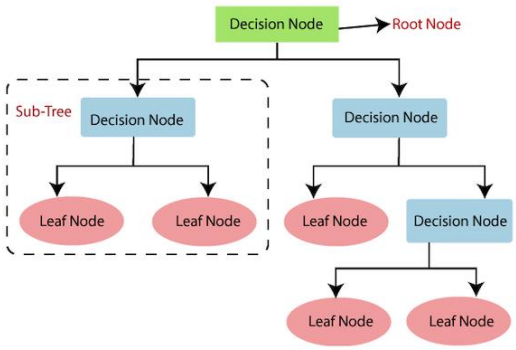
Proses pengumpulan, penggabungan, penataan, dan pengorganisasian data. Data dibersihkan terlebih dahulu melalui serangkaian proses yang disebut dengan *preprocessing*. *Preprocessing* terdiri dari pengecekan *missing value* dan redundansi (Hutahaean & Wijayanto, 2022). Hal ini bertujuan agar data yang digunakan pada pemodelan selanjutnya telah bebas dari *noise* sehingga model yang dihasilkan kualitasnya semakin baik (Iman & Wijayanto, 2021).

d. Modelling

Menggunakan *k-fold cross validation* dan membuat model klasifikasi dengan mengaplikasikan berbagai metode *data mining* sebagai berikut.

- *Decision Tree*

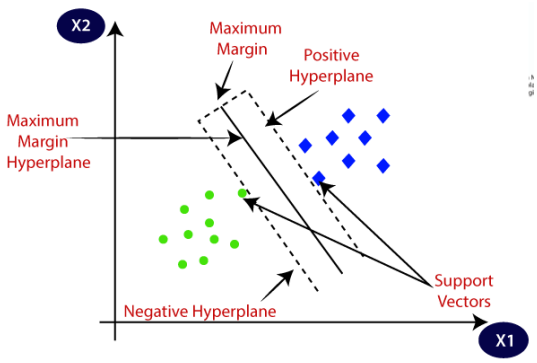
  Decision tree merupakan metode yang didasarkan untuk membuat model aturan berupa pohon dari *data training* yang ada. Kemudian model yang terbentuk dapat digunakan untuk mengklasifikasikan objek yang baru. Pohon keputusan digunakan ketika hasil prediksi merupakan keanggotaan dari suatu kelas/kelompok. Modal yang dihasilkan oleh *decision tree* ini memiliki kelebihan dan kesederhanaannya dan kemudahannya untuk diinterpretasikan. Algoritma *decision tree* bekerja secara *top-down* dengan memilih atribut yang merupakan prediktor terbaik sebagai akar. Kemudian atribut berikutnya yang merupakan atribut *splitting* terbaik menjadi cabang dari pohon yang terbaik hingga atribut-atribut yang ada semuanya menjadi cabang dari pohon (Pramana, 2018).



Gambar. 2 – Decision Tree

- *Support Vector Machine* (SVM)

  Support vector machine (SVM) merupakan salah satu algoritma *machine learning* dengan pendekatan *supervised learning* untuk klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan margin terbesar (*maximum margin*). Metode SVM memiliki konsep sentral dalam mengklasifikasikan data, yaitu mencari *hyperplane* terbaik untuk memisahkan antara dua kelas yang telah ditentukan (Putri & Wijayanto, 2022). Tujuan SVM untuk menemukan bidang yang memiliki margin maksimum yang merupakan jarak terbesar antara titik data dari kedua kelas.

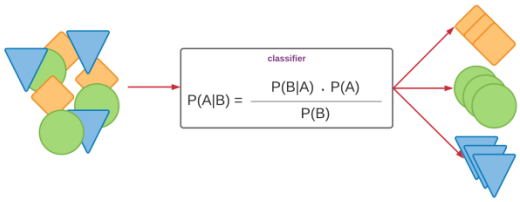


Gambar. 3 – Support Vector Machine (SVM)

- *Naïve Bayes*

  Naïve bayes merupakan metode klasifikasi berbasis peluang menggunakan teorema Bayes dengan asumsi semua atribut sama pentingnya dan bersifat independen secara statistik. Dalam metode *naïve bayes* diperlukan data latih dan data uji yang ingin diklasifikasikan, dalam *naïve bayes*, semaskin banyak data latih yang yang dilibatkan, semakin baik hasil yang prediksi yang diberikan (Fadlan, et al., 2018).

$$P(C|F_1 \dots F_n) = \frac{P(C).P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$

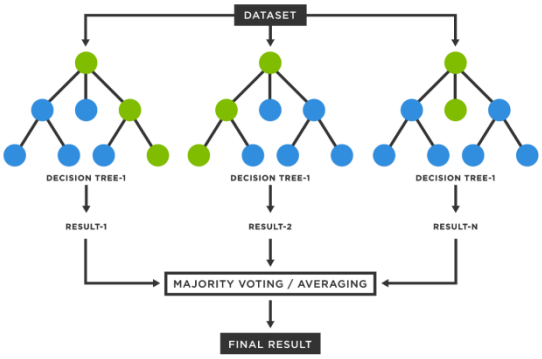


Gambar. 4 – Naïve Bayes

- *Random Forest*

  Random forest merupakan metode klasifikasi berbasis pohon keputusan yang dilakukan dengan

membuat banyak pohon dari sampel-sampel *training set*. *Training set* dibuat sejumlah k yang disampel menggunakan *metode random sampling with replacement* (WR) atau dikenal dengan istilah "*bagging*". Penggunaan pohon yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik (Tahyudin et al., 2021 dalam Syaidatussalihah, 2022). Metode *random forest* memiliki keunggulan proses penyelesaian yang relatif cepat, tidak terjadi *overfit* seiring dengan penambahan jumlah pohon, dan memiliki akurasi yang lebih baik dari *decision tree* (Breiman, 2001 dalam Nurkhaliza & Wijayanto, 2022).



Gambar. 5 – Random Forest

e. Evaluation

Melakukan interpretasi terhadap hasil klasifikasi *data mining* dan mengevaluasi kinerja dari model-model yang sudah dibangun. Evaluasi terhadap model tersebut dilihat dari nilai akurasi dan F1-*score*. F1-*score* dapat menjadi pertimbangan pada data yang *imbalance* (Nurpiana & Wijayanto, 2022).

f. Deployment

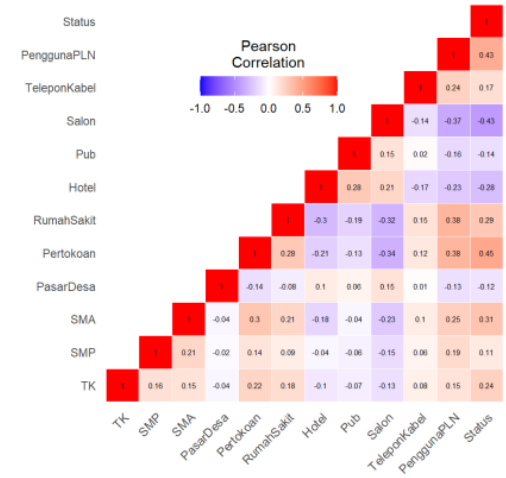
Rencana penerapan atau penggunaan model. *Tools data mining* dioptimalkan untuk pengembangan model dan biasanya tidak menyediakan antarmuka khusus untuk saat menerapkan model (Bellazi & Zupan, 2008).

2.3. Klasifikasi Desa Perkotaan dan Perdesaan

Klasifikasi adalah proses dari pembangunan terhadap suatu model yang mengklasifikan suatu objek sesuai dengan atribut-atributnya (Susilowati, 2015 dalam Kemala, 2021). Suatu wilayah desa/kelurahan dikatakan perdesaan adalah wilayah yang belum memenuhi kriteria klasifikasi desa perkotaan. Penilaian suatu wilayah didasarkan oleh tiga komponen utama penilaian yaitu kepadatan penduduk, persentase keluarga pertanian, dan keberadaan atau akses ke fasilitas perkotaan. Jika wilayah desa atau kelurahan masing-masing komponennya memiliki skor 9 (sembilan) atau lebih maka wilayah tersebut diklasifikasikan sebagai desa perkotaan (BPS, 2020).

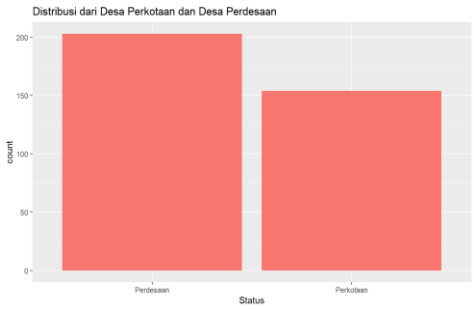
3. Hasil dan Pembahasan

Data yang akan diolah terdiri dari dua kategori (kelas), yaitu desa perkotaan dan desa perdesaan. Pada penelitian ini, tidak ada *missing value* pada data sehingga tidak perlu dilakukan imputasi. Pada gambar 6 terlihat bahwa tidak terdapat warna yang pekat pada plot korelasi. Hal ini menunjukkan setiap variabel independen tidak memiliki korelasi yang tinggi sehingga tidak terdapat redundansi pada data dan semua atribut dapat digunakan untuk melakukan klasifikasi.



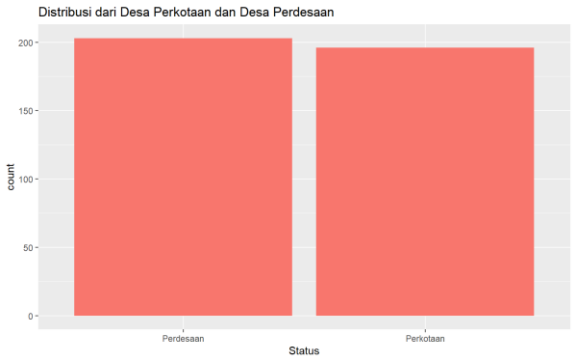
Gambar. 6 – Plot Korelasi Antar Fitur

Selanjutnya menyiapkan data dengan melihat distribusi kelas atau kategori dan *imbalance ratio* dari data asli. Pada gambar 7 terlihat bahwa *imbalance ratio* sebesar 0,7586 yang berarti perbandingan data antara dua kategori tersebut *imbalance* atau belum begitu seimbang sehingga perlu dilakukan sedikit penyesuaian dengan metode *resampling*.



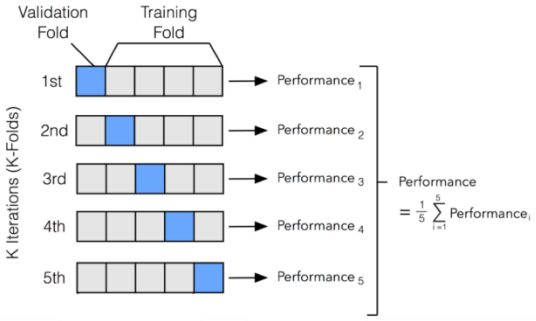
Gambar. 7 – Distribusi Klasifikasi Desa

Untuk menyeimbangkan distribusi data, penulis menggunakan algoritma resampling *Adaptive Synthetic* (ADASYN) sehingga *diperoleh imbalance ratio* sebesar 0,9655 (dapat dilihat pada gambar 8).



Gambar. 8 – Hasil Resampling ADASYN

Record yang digunakan sebanyak 357 yang kemudian dilakukan pembagian data dengan *10-fold cross validation* seperti yang dijelaskan pada gambar 9 menggunakan *software* R Studio.



Gambar. 9 – Cross Validation

Metode yang diusulkan pada penelitian ini yaitu dengan menggunakan algoritma klasifikasi *data mining* untuk membandingkan atau mengevaluasi model *decision tree*, *support vector machine*, *naïve bayes*, dan *random forest* diuji pada dataset potensi desa Kabupaten Purwakarta dan Bandung Barat tahun 2021. Dataset yang digunakan mempunyai 8 atribut terdiri dari data nominal dan numerik. Hasil pengukuran algoritma klasifikasi dievaluasi menggunakan nilai akurasi dan F1-*score* untuk mengetahui perbedaan atau perbandingan kinerja model klasifikasi *decision tree* (DT), *support vector machine* (SVM), *naïve bayes* (NB), dan *random forest* (RF). Nilai akurasi menunjukkan persentase model tersebut mampu memprediksi dengan tepat (Nisa & Nooraeni, 2020).

Tabel 2 – Hasil Evaluasi Model.

Evaluasi	DT	SVM	NB	RF
Akurasi	0,7805	0,8536	0,7500	0,9000
F1-*score*	0,7907	0,8571	0,8077	0,9000

Dari Tabel 2 dapat diketahui bahwa nilai maksimum persentase akurasi dari setiap model berturut-turut sebesar 78,04%, 85,36%, 85,36%, 75,00%, 90,00%. Selanjutnya, jika dilihat dari nilai F1-*score* model *decision tree*, *support vector machine*, *naïve bayes*, dan *random forest* berturut-turut persentasenya sebesar 79,07%, 85,71%, 80,77%, dan 90,00%. Algoritma *random forest* memiliki nilai akurasi dan F1-*score* tertinggi dibandingkan dengan model algoritma lainnya yaitu sebesar 0,9. Hal tersebut menunjukkan bahwa kemampuan klasifikasi yang dibentuk dengan metode *random forest* dalam mengklasifikasikan data dengan benar sebesar 90%. Akibatnya, metode *random forest* dapat dikatakan sebagai metode klasifikasi terbaik dalam pengklasifikasian desa perkotaan dan desa perdesaan di Kabupaten Purwakarta dan Kabupaten Bandung Barat berdasarkan data potensi desa tahun 2021.

4. Kesimpulan dan Saran

Penelitian ini menggunakan data potensi desa Kabupaten Purwakarta dan Bandung Barat tahun 2021 untuk membandingkan empat model klasifikasi dalam *data mining*, yaitu *decision tree*, *support vector machine*, *naïve bayes*, dan *random forest* dengan melakukan 10-*fold cross validation*. Hasil penelitian menunjukkan bahwa *decision tree*, *support vector machine*, *naïve bayes*, dan *random forest* dengan melakukan 10-*fold cross validation* dapat mengklasifikasikan desa perkotaan dan desa perdesaan dengan baik karena memiliki nilai akurasi dan F1-*score* di atas 75%. Kemudian, metode terbaik yang digunakan untuk pengklasifikasian desa perkotaan dan perdesaan yaitu menggunakan metode *random forest* karena memiliki performa kinerja sangat baik (akurasi dan F1-*score* tertinggi).

Selanjutnya BPS dapat mempertimbangkan metode *random forest* dalam mengklasifikasikan desa perkotaan dan perdesaan sebagai dasar rekomendasi kebijakan pembangunan wilayah desa bagi pemerintah daerah setempat di Kabupaten Purwakarta dan Bandung Barat. Penelitian selanjutnya dapat menggunakan indikator atau variabel lain yang sekiranya berpengaruh terhadap pengklasifikasian desa perkotaan dan desa perdesaan.

Ucapan Terima Kasih

Terima kasih disampaikan kepada Bapak Dr. Eng. Arie Wahyu Wijayanto, S.S.T., M.T., selaku dosen pengampu mata kuliah *data mining* yang telah membimbing dan memberikan masukan dalam penulisan jurnal penelitian ini.

DAFTAR PUSTAKA

Apriliansyah., Pangestika, A., Ramadhanty, A. P., Putra, G. M., Putri, G. S. N. D S., et al. (2021). Klasifikasi Status Desa/Kelurahan DIY (Yogyakarta) Menggunakan Model Decision Tree. *Journal Engineering, Mathematics and Computer Sciece*, 33–41.

Aryani, Y., & Wijayanto, A. W. (2021). Klasifikasi Pengembalian Radar dari Ionosfer Menggunakan SVM, Naive Bayes dan Random Forest. *Komputika: Jurnal Sistem Komputer*, *10*(2), 111-117.

BPS (2020). *Peraturan Kepala Badan Pusat Statistik Nomor 120 Tahun 2020 Tentang Klasifikasi Perkotaan dan Perdesaan di Indonesia 2020*. Jakarta: Badan Pusat Statistik.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, *77*(2), 81-97.

Chapman, Pete., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0 Step by step data mining guides.

Fadlan, C., Ningsih, S., & Windarto, A. P. (2018). Penerapan Metode Naïve Bayes Dalam Klasifikasi Kelayakan Keluarga Penerima Beras Rastra. *JUTIM (Jurnal Teknik Informatika Musirawas)*, *3*(1), 1-8.

Hutahaean, Y. M., & Wijayanto, A. W. (2022). Klasifikasi Rumah Tangga Penerima Subsidi Listrik di Provinsi Gorontalo Tahun 2019 dengan Metode K-Nearest Neighbor dan Support Vector Machine. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *10*(1), 63-68.

Kemendes PDTT (2021). *Rencana Strategis Direktorat Jenderal Pembangunan Desa dan Perdesaan 2020-2024*. Jakarta: Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi.

Iman, Q., & Wijayanto, A. W. (2021). Klasifikasi Rumah Tangga Penerima Beras Miskin (Raskin)/Beras Sejahtera (Rastra) di Provinsi Jawa Barat Tahun 2017 dengan Metode Random Forest dan Support Vector Machine. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *9*(2), 178-184.

Jackson, J. (2002). Data mining; a conceptual overview. *Communications of the Association for Information Systems*, *8*(1), 19.

Kemala, I., & Wijayanto, A. W. (2021). Perbandingan Kinerja Metode Bagging dan Non-Ensemble Machine Learning pada Klasifikasi Wilayah di Indonesia menurut Indeks Pembangunan Manusia. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *9*(2), 269-275.

Nisa, I. M. K., & Nooraeni, R. (2020). Penerapan Metode Random Forest Untuk Klasifikasi Wanita Usia Subur di Perdesaan Dalam Menggunakan Internet (SDKI 2017). *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, *8*(1), 72-76.

Nurkhaliza, A. A., & Wijayanto, A. W. (2022) Perbandingan Algoritma Klasifikasi Support Vector Machine dan Random Forest pada Prediksi Status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) Satuan Kerja BPS di Indonesia Tahun 2020. *Jurnal Informatika Universitas Pamulang*, *7*(1), 54-59.

Nurpiana, A., & Wijayanto, A. W. (2022). Comparison of Models for Classification of Learning Achievement of Middle School Students in Indonesia in 2019 using the Support Vector Machine Algorithm, Conditional Inference Trees, and Random Forest. *Jurnal Matematika, Statistika dan Komputasi*, *18*(3), 447-455.

Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika: Jurnal Sistem Komputer*, *11*(1), 59-66.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., Nooraeni, R. (2018). *Data Mining dengan R*. Bogor: In Media.

Sari, M.S., Safitri, D., Sugito. (2014). Klasifikasi Wilayah Desa-Perdesaan dan Desa-Perkotaan Wilayah Kabupaten Semarang dengan Support Machine System (SVM). *Journal Gaussian*, 751-760.

Supartini, I. A. M., Sukarsa, I. K. G., & Srinadi, I. G. A. M. (2017). Analisis Diskriminan Pada Klasifikasi Desa Di Kabupaten Tabanan Menggunakan Metode K-Fold Cross Validation. *E-Jurnal Matematika*, *6*(2), 106-115.

Syaidatussalihah, & Abdurahim. (2022). Classification of Poverty Status using the Random Forest Algorithm. *Eigen Mathematics Journal*, *5*(1), 37-44.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29-39).