Journal of System and Computer Engineering (JSCE) ISSN (online): 2723-1240
Vol. 2 No. 1 Januari 2022

Klasifikasi Status NEET pada Penduduk Usia Muda di Indonesia dengan SVM dan *Random Forest*

Herdina Dwi Ramadhanti
[1]Politeknik Statistika STIS

Abstrak

Not in Education, Employment, or Training (NEET) adalah suatu indikator untuk mengetahui tingkat kerentanan penduduk usia muda dalam pengangguran, putus sekolah, serta keputusasaan terhadap pasar tenaga kerja. Menurut ILO, Indonesia merupakan salah satu negara dengan tingkat NEET tertinggi di Asia sehingga menjadi suatu masalah yang perlu untuk segera diatasi. Salah satu alternatif yang dapat dilakukan untuk mengatasi fenomena tersebut adalah dengan deteksi dini terhadap penduduk yang berisiko menjadi NEET yang dapat dilakukan dengan menggunakan indikator-indikator yang telah melekat dalam individu seperti jenis kelamin, status perkawinan, dan disabilitas. Penelitian ini bertujuan untuk melakukan klasifikasi terhadap status NEET pada penduduk usia muda agar dapat digunakan untuk memprediksi apakah individu termasuk ke dalam NEET dengan menggunakan metode klasifikasi yang meliputi *Support Vector Machine* (SVM) dan *random forest*. Metode SVM dipilih untuk mewakili *non-ensemble method* sedangkan *random forest* dipilih untuk mewakili *ensemble method*. Data yang digunakan merupakan data sekunder yang diperoleh dari *raw data* SAKERNAS periode Agustus 2018. Hasil penelitian menunjukkan bahwa metode *random forest* memberikan hasil akurasi yang lebih tinggi sehingga memiliki kemampuan yang lebih baik dalam mengklasifikasikan penduduk muda menurut status NEET yaitu dengan akurasi sebesar 82,94 persen. Oleh karena itu, metode ini dapat digunakan untuk memprediksi status NEET dalam rangka menunjang pengurangan persentase NEET di Indonesia.

Kata Kunci: *Data mining, Klasifikasi, NEET, Pengangguran muda, Random forest, SVM.*

Abstract

Not in Education, Employment, or Training (NEET) refers to a person who is unemployed and not receiving an education or vocational training. According to ILO, the proportion of NEET in Indonesia is quite high and in fact is one of the highest rates in Asia. Thus, NEET in Indonesia is a problem that needs to be tackled immediately. One alternative that can be done to overcome this issue is by conducting early detection of young people who are at risk of becoming NEET. This detection can be done by using indicators that are inherent in the individual such as gender, marital status, and disability. This study aims to classify the NEET status of the young population so that it can be used to predict whether an individual is included in the NEET by using classification methods that include Support Vector Machine (SVM) and random forest. The SVM method is applied to represent the non-ensemble method while random forest is applied to represent the ensemble method.The data used in this study are secondary data obtained from SAKERNAS for the period of August 2018. The results of the study indicate that the random forest method provides higher accuracy results so that it has a better ability to classify young people according to NEET status with an accuracy of

82.94 percent. Therefore, this method can be used to predict the status of NEET in order to support the reduction in the percentage of NEET in Indonesia.

Keywords: *Classification, Data mining, NEET, Random forest, SVM, Youth unemployment.*

1. Pendahuluan

Menurut Bappenas (2017), Indonesia akan mengalami bonus demografi pada tahun 2030, yakni kondisi ketika jumlah penduduk umur produktif lebih besar dibandingkan penduduk umur tidak produktif. Hal ini dipandang sebagai keuntungan ekonomis karena peran mereka sebagai tenaga kerja merupakan salah satu faktor produksi yang pada akhirnya dapat memberikan andil dalam memicu pertumbuhan ekonomi. Oleh karena itu, Indonesia juga diharapkan mampu memanfaatkan peluang ini dengan maksimal.

Kendati demikian, fenomena bonus demografi juga menciptakan tantangan tersendiri lantaran kegagalan suatu negara dalam memanfaatkan bonus demografi justru dapat berubah menjadi beban perekonomian dan menciptakan gelombang pengangguran. Masalah pengangguran tersebut dapat terjadi ketika tingginya pertumbuhan angkatan kerja tidak diiringi dengan ketersediaan lapangan pekerjaan. Sementara itu, Menteri Ketenagakerjaan, Ida Fauziyah, menyatakan bahwa salah satu tantangan terbesar kondisi ketenagakerjaan di Indonesia adalah masalah pengangguran muda. Hal ini dikarenakan penyumbang terbesar pengangguran di Indonesia merupakan angkatan kerja muda, yaitu pada kelompok umur 15 hingga 24 tahun. Faktanya, bersumber dari data BPS, apabila ditinjau data tingkat pengangguran terbuka (TPT) menurut kelompok umur, dapat diketahui bahwa TPT pada kelompok umur ini selalu memiliki persentase tertinggi. Dari tahun ke tahun, persentase tersebut berada di kisaran 10-30 persen. Hal ini sangat kontras dengan TPT pada kelompok umur lainnnya yang cenderung memiliki persentase yang jauh lebih rendah. Tingginya tingkat pengangguran muda di Indonesia merupakan suatu masalah karena pengangguran yang berkelanjutan akan membuat transisi dari usia muda menuju dewasa menjadi sulit, meningkatkan risiko kemiskinan di masa depan, meningkatkan peluang terlibat dalam perilaku yang bermasalah, dan dapat mengurangi keterlibatan usia muda dalam politik dan sosial (Bay dan Blekeseanu, 2002).

Lebih lanjut, pasar tenaga kerja usia muda merupakan salah satu perhatian utama di sebagian besar negara seiring dengan semakin meningkatnya pengangguran usia muda secara global (Scarpetta dkk., 2010). Dalam hal ini, indikator yang sering digunakan terkait pengangguran muda adalah Tingkat Pengangguran Terbuka (TPT). Akan tetapi, indikator tersebut hanya mencakup penduduk usia muda yang termasuk dalam angkatan kerja sehingga tidak merefleksikan situasi penduduk usia muda secara keseluruhan. Sehubungan dengan hal tersebut, *International Labor Organization* (ILO) mengembangkan indikator *Not in Employment, Education, or Training* (NEET) untuk memperluas ranah kerentanan penduduk usia muda dalam pengangguran, putus sekolah, serta keputusasaan terhadap pasar tenaga kerja (Wickremeratne dan Danusinghe, 2018).

Menurut ILO (2017), proporsi NEET di Indonesia cenderung tinggi dan merupakan salah satu negara dengan tingkat NEET tertinggi di Asia. Faktanya, tercatat hingga tahun 2018, mengacu pada data yang bersumber dari ILOSTAT, NEET di Indonesia adalah sebesar 21,7 persen, jauh lebih tinggi dibandingkan negara-negara di

sekitarnya seperti Malaysia (12,5 persen), Thailand (14,8 persen), Vietnam (8,3 persen). Dengan demikian, NEET di Indonesia merupakan suatu masalah yang perlu untuk segera diatasi. Sehubungan dengan hal tersebut, maka diperlukan kebijakan yang tepat untuk menanggulangi tingginya persentase NEET di Indonesia.

Salah satu alternatif yang dapat dilakukan untuk mengatasi fenomena NEET di Indonesia adalah dengan deteksi dini terhadap penduduk yang berisiko menjadi NEET. Pendeteksian ini dapat dilakukan dengan menggunakan indikator-indikator yang telah melekat dalam individu seperti jenis kelamin, status perkawinan, dan disabilitas. Sehingga, apabila hasil dari pendeteksian tersebut ditindaklanjuti dengan tepat, maka diharapkan dapat mencegah seseorang untuk menjadi NEET. Oleh karena itu, diperlukan metode yang tepat untuk menunjang proses penentuan NEET dengan akurat agar dapat mengakomodasi perubahan kondisi tenaga kerja di antara penduduk khususnya penduduk usia muda. Selain itu, kondisi tenaga kerja di Indonesia yang sangat dinamis juga menuntut terciptanya metode pengklasifikasian yang cepat dalam melakukan pemrosesan apabila terdapat modifikasi dan pembaharuan terhadap data.

Data mining adalah sebuah proses untuk menentukan hubungan yang terdapat di dalam data yang sebelumnya tidak diketahui oleh pengguna dimana hubungan tersebut dapat digunakan sebagai dasar untuk pengambilan keputusan (Schell dan McLeod, 2007). Secara umum, terdapat dua metode utama dalam *data mining* yang meliputi *supervised learning* dan *unsupervised learning*. Klasifikasi merupakan satu di antara algoritma dalam *supervised learning* yang paling populer. Dalam hal ini, teknik-teknik klasifikasi *data mining* dapat melakukan klasifikasi dengan menggunakan ukuran data yang besar dalam waktu yang relatif cepat.

Dengan demikian, berdasarkan pemaparan di atas, maka penelitian ini bertujuan untuk melakukan klasifikasi terhadap status NEET pada penduduk usia muda dengan menggunakan metode klasifikasi yang meliputi *Support Vector Machine* (SVM) dan *random forest*. Metode SVM dipilih untuk mewakili *non-ensemble method* sedangkan *random forest* dipilih untuk mewakili *ensemble method*. Dengan penggunaan metode-metode tersebut tersebut, diharapkan dapat diperoleh model terbaik yang mampu digunakan untuk memprediksi apakah individu termasuk ke dalam NEET.

2. Tinjauan Pustaka

2.1. *Not in, Education, or Training* (NEET)

Menurut ILO (2017), NEET adalah individu dengan usia 15 hingga 24 tahun yang tidak berada dalam pekerjaan, pendidikan, maupun pelatihan. Persentese NEET secara umum didefinisikan sebagai jumlah penduduk usia muda dikurangi dengan jumlah penduduk usia muda yang berada dalam pekerjaan, pendidikan, atau pelatihan dan kemudian dibagi dengan total penduduk usia muda dan dikalikan seratus persen. NEET terbagi menjadi *unemployed* NEET dan *inactive* NEET. *Unemployed* NEET adalah penduduk usia muda yang berstatus sebagai penganggur, yaitu penduduk yang tidak bekerja tetapi sedang mencari pekerjaan, atau mempersiapkan usaha, atau merasa tidak mungkin mendapatkan pekerjaan, atau sudah diterima bekerja tetapi belum mulai bekerja, atau sudah mempunyai usaha tetapi belum memulainya. Sementara itu, *inactive* NEET adalah penduduk usia muda yang tidak dalam pekerjaan, pendidikan, ataupun pelatihan tetapi tidak mencari pekerjaan dan tidak bersedia menerima pekerjaan.

2.2. *Supervised Learning*

Supervised learning merupakan metode pembelajaran terarah yang memiliki 'guru', dimana *dataset* memiliki label yang betul yang digunakan untuk membantu model mengenali pola yang kemudian diterapkan pada *dataset* yang lain yang tidak memiliki label (Revar dkk., 2010). Klasifikasi merupakan satu di antara algoritma dalam *supervised learning* yang paling populer. Dalam melakukan klasifikasi, data dibagi menjadi dua bagian yaitu data latih (*training set*) yang digunakan untuk pembentukan model dan *testing set* yang digunakan untuk menilai tingkat akurasi dari klasifikasi yang dilakukan. Salah satu metode yang seing digunakan untuk pembagian data tersebut adalah *k-fold cross validation*. Metode ini membagi data secara acak menjadi k bagian untuk kemudian dilatih dengan menggunakan beberapa bagian data dan diuji dengan bagian lainnya. Ide dasar dari metode ini adalah ketika akurasi hanya dari data sampel, maka akan sangat mungkin untuk bias. Oleh karena itu, metode *cross validation* digunakan untuk menghindari *overlapping* pada data *testing* sehingga dapat menghindari bias. Adapun nilai k yan paling umum digunakan adalah 10 (*10-fold cross validation*). Hal ini dikarenakan hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis menunjukkan bahwa *10-fold cross validation* adalah pilihan terbaik karena dapat memberikan hasil yang optimum terkait akurasi dan *running time*. Dalam hal ini, *10-fold cross validation* akan mengulang pengujian sebanyak 10 kali dimana hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian tersebut (Han dan Kamber, 2012).

2.3. *Preprocessing Data*

Preprocessing data merupakan hal vital yang perlu dilakukan dalam proses *data mining*. Hal ini dikarenakan data mentah yang tersedia tidak terlepas dari unsur ketidaklengkapan, *noisy*, dan inkonsistensi data. Data yang tidak lengkap menunjukkan adanya nilai atribur yang kurang, tidak disertakan, atau hanya memuat data agregat. Sementara itu, data yang bersifat *noisy* adalah data yang masih memuat *error* dan atau data pencilan. Adapun data yang tidak konsisten merupakan data yang memuat perbedaan seperti dalam pemberian kode atau nama. Hal-hal di atas dapat menyebabkan data menjadi 'kotor' dan menurunkan kualitas data. Data yang berkualitas rendah dapat menghasilkan kualitas *mining* yang rendah pula. Oleh karena itu, *Preprocessing data* dilakukan untuk mengakomodasi hal tersebut. Dalam hal ini, terdapat empat peran utama *preprocessing data* yaitu meliputi *data cleaning* (mengisi *missing data*, mengenali atau menghilangkan data pencilan, melakukan *smoothing* data, serta memperbaiki inkonsistensi pada data), *data integration* (mengintegrasikan data yang tersebar di beberapa *database*/*file*), *data reduction* (mengurangi ukuran data baik dari sisi ukuran, dimensi, ataupun keragaman), serta *data transformation* dan *data discretization* (melakukan tranformasi dan reduksi data dengan tujuan tertentu) (Han dan Kamber, 2012).

2.4. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi dan termasuk dalam *non-ensemble method*. Metode SVM memakai ruang hipotesis dalam bentuk fungsi-fungsi *linear* dalam sebuah fitur berdimensi tinggi dan dilatih dengan

algoritma *learning* yang didasarkan pada teori optimasi. Dalam hal ini, SVM bekerja dengan menemukan *hyperplane* terbaik untuk memisahkan dua kelas yang berbeda dengan memaksimumkan *margin* yang diperoleh dari *support vector*. Selain diterapkan pada data *linear*, SVM juga dapat diterapkan pada data *nonlinear* dengan menggunakan *kernel trick*. untuk menemukan *hyperplane*, *kernel trick* akan mentransformasi *dataset* ke dalam ruang vektor dengan dimensi tinggi dimana proses pengklasifikasian dilakukan di ruang vektor tersebut (Maulana dan Irhamah, 2018). Lebih lanjut, penentuan fungsi *kernel* serta parameter yang digunakan dalam pengklasifikasikan akan sangat berpengaruh terhadap tingkat akurasi yang dihasilkan. Adapun fungsi *kernel* yang paling banyak digunakan dalam SVM meliputi fungsi *kernel linear, polynomial*, dan *Radial Basis Function* (RBF) (Feta dan Ginanjar, 2019). Secara teoretis, SVM merupakan metode *machine learning* yang superior dengan hasil performa yang baik dibandingkan metode *non-ensemble* lainnya (Tzotsos dan Argialas, 2008). Berbagai penelitian membuktikan bahwa SVM memiliki performa yang lebih unggul dibandingkan *non-ensemble method* lainnya seperti metode *K-Nearest Neighbors*, *Naive Bayes*, *Quadratic Bayes Normal*, *Nearest Mean* (Amami dkk., 2012), *decision tree* dan *rule-learners* (Osisanwo dkk., 2017), serta regresi logistik (Salazar dkk., 2012).

2.5. *Random Forest*

Random forest merupakan algoritma *machine learning* pada klasifikasi yang berupa sekumpulan metode pembelajaran (*ensemble method*) menggunakan *bagging* (Breiman, 2001). *Ensemble method* merupakan metode untuk meningkatkan performa klasifikasi (Bühlmann, 2012). Sehubungan dengan itu, penggunaan *bagging* bekerja dengan mengurangi varians (*noise*) pada metode-metode dasar (Breiman, 1996). Metode *random forest* memanfaatkan pohon keputusan yang digunakan untuk *base classifier* dengan cara dibangun dan dikombinasikan. Adapun aspek penting dalam metode ini adalah melakukan *bootstrap sampling* untuk membentuk pohon prediksi, kemudian setiap pohon keputusan akan memprediksi dengan prediktor acak, dan *random forest* akan memprediksi dengan melakukan kombinasi dari hasil setiap pohon keputusan dengan menggunakan *voting* terbanyak (*majority vote*) untuk keperluan klasifikasi, serta dengan menggunakan rata-rata untuk keperluan regresi (Sadewo dkk., 2017),

2.6. Kriteria Evaluasi dan Validasi Model

Secara umum, dalam memilih model yang bagus, diperlukan kriteria evaluasi dan validasi terhadap model yang terbentuk. Penelitian ini menggunakan *confusion matrix* untuk memberikan rincian terkait hasil klasifikasi. *Confusion matrix* merupakan suatu alat yang dapat digunakan untuk menganalisis seberapa baik suatu metode klasifikasi mengenali *tuple* dari kelas yang berbeda. *True positive* (TP) dan *true negative* (TN) memberikan informasi ketika *classifier* benar, sedangkan *false positive* (FP) dan *false negative* (FN) memberitahu ketika *classifier* salah. Ukuran yang dihasilkan dari *confusion matrix* diantaranya meliputi akurasi, *sensitivity*/*recall*, *specificity*, presisi/*Positive Predictive Value* (PPV), dan *Negative Predictive Value* (NPV). Akurasi adalah ukuran dari seberapa baik model dalam mengkorelasikan antara hasil dengan atribut dalam data yang sudah ada dan merupakan metode yang paling umum digunakan untuk mengevaluasi apakah sebuah model bagus atau tidak.

Sensitivity/*recall* mengukur proporsi data *true positive* yang teridentifikasi dengan tepat. *Specificity* mengukur proporsi data *true negative* yang teridentifikasi dengan tepat. presisi/PPV mengukur proporsi data dengan hasil positif yang didiagnosis dengan benar sedangkan NPV mengukur proporsi data dengan hasil negatif yang didiagnosis dengan benar (Han dan Kamber, 2012).

3. Perancangan dan Implementasi

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian merupakan data sekunder yang diperoleh dari *raw data* SAKERNAS pada Bulan Agustus tahun 2018. Cakupan wilayah dalam penelitian ini adalah seluruh wilayah di Indonesia dengan unit analisis penduduk berumur 15-24 tahun. Dari keseluruhan data sampel, sebanyak 20.915 penduduk termasuk ke dalam kategori umur tersebut.

3.2. Pengolahan Data

Tahap pengolahan data diawali dengan pembentukan klasifikasi pada variabel dependen. Variabel dependen dalam penelitan ini adalah status NEET. Untuk mendapatkan klasifikasi tersebut, dilakukan pengolahan terhadap variabel b5_r5a1 (bekerja minimal 1 jam tanpa terputus), b5_r6 (sementara tidak bekerja minimal 1 jam tanpa terputus), b4_k8 (partisipasi sekolah), dan b5_r1f (sedang mengikuti pelatihan/kursus). Individu tergolong ke dalam NEET apabila tidak sedang melakukan aktivitas dari indikator-indikator di atas (b5_r5a1 = 2, b5_r6 = 2, b4_k8 = 1 atau 3, dan b5_r1f = 2).

Tahapan selanjutnya, dilakukan *preprocessing data* yang meliputi *data reduction* dan *data transformation*. *Data reduction* dilakukan dengan melakukan reduksi terhadap dimensi data. Dengan menggunakan nilai koefisien korelasi, kemudian dilakukan pemilihan terhadap atribut yang paling relevan untuk menunjang pengklasifikasian status NEET, yaitu atribut-atribut dengan nilai korelasi tertinggi. Sehubungan dengan hal tersebut, variabel-variabel yang digunakan sebagai variabel independen meliputi klasifikasi, b2_r1 (jumlah anggota rumah tangga), b4_k3 (hubungan dengan kepala rumah tangga), b4_k6 (jenis kelamin), b4_k8 (umur), b4_k10 (status perkawinan), b5_r1a (ijazah tertinggi yang dimiliki), b5_r1d (pernah mendapat pelatihan/kursus dan memperoleh sertifikat), b5_r4a (kesulitan/gangguan melihat), b5_r4b (kesulitan/gangguan mendengar), b5_r4c (kesulitan/gangguan berjalan), b5_r4d (kesulitan/gangguan memegang), b5_r4e (kesulitan/gangguan berbicara), dan b5_r4f (kesulitan/gangguan lain).

Tabel 1. Definisi Operasional Variabel

Variabel	Tipe Data	Kategori
Status NEET (kelas)	Kategorik	1: Ya 0: Tidak
Klasifikasi	Kategorik	1: Perkotaan 0: Perdesaan
Jumlah anggota rumah tangga	Numerik	
Status kepala rumah tangga (KRT)	Kategorik	1: KRT 0: Bukan KRT

Jenis kelamin	Kategorik	1: Laki-laki 0: Perempuan
Umur	Numerik	
Status perkawinan	Kategorik	1: Kawin 0: Tidak kawin
Ijazah tertinggi yang dimiliki	Kategorik	1: SD/sederajat 2: SMP/sederajat 3: SMA/sederajat 4: Di atas SMA 0: Tidak memiliki ijazah
Pernah memperoleh pelatihan/kursus	Kategorik	1: Ya 0: Tidak
Penyandang disabilitas	Kategorik	1: Ya 0: Tidak

Tahapan terakhir dilakukan transformasi data yang meliputi agregasi dan normalisasi. Agregasi dilakukan untuk variabel b5_r4a hingga b5_r4f dengan membentuk variabel disabilitas dengan dua kategori (penderita disabilitas dan bukan penderita disabilitas). Kemudian, dilakukan penggabungan beberapa ketagori pada variabel b4_k3 (KRT dan bukan KRT), b4_k10 (kawin dan tidak kawin), b5_r1a (tidak memiliki ijazah, ijazah SD/sederajat, ijazah SMP/sederajat, ijazah SMA/sederajat, dan ijazah di atas SMA). Selanjutnya, dilakukan normalisasi terhadap variabel numerik yaitu pada variabel b2_r1 (jumlah anggota rumah tangga) dan b4_k8 (umur). Hasil dari pengolahan data ini adalah dataset yang siap digunakan pada algoritma SVM dan *random forest* untuk pengklasifikasian seperti yang disajikan pada Tabel 1.

3.3. Pengimplementasian Metode SVM dan *Random Forest*

Pada tahapan ini, dilakukan pengimplementasian metode SVM dan *random forest*. Metode SVM dipilih untuk mewakili *non-ensemble method* yang menurut penelitian terdahulu memiliki performa lebih unggul dibandingkan metode *non-ensemble* lainnya. Sementara itu, *random forest* dipilih untuk mewakili *ensemble method*. Pada metode SVM, digunakan fungsi *kernel linear*, *radial*, dan *polynomial* pada fungsi pemisah (*hyperplane*) SVM. Dalam hal ini, penentuan parameter terbaik pada masing-masing fungsi *kernel* dilakukan dengan metode *10-fold cross validation*. Setelah dilakukan penyeleksian terhadap parameter, diperoleh parameter terbaik yaitu dengan fungsi *kernel linear* dengan tipe *C-classification* dan nilai *cost* = 1. Sementara itu, dalam metode *random forest*, dilakukan penentuan jumlah variabel yang digunakan sebagai *split* pada setiap pohon yang terbentuk (mtry) yaitu sebanyak 2, 7, dan 12 yang dilakukan dengan metode *10-fold cross validation*. Setelah dilakukan penyeleksian terhadap parameter, diperoleh bahwa model terbaik adalah model dengan nilai mtry = 2.

4. Hasil

Tabel 2. Hasil *Confusion Matrix* dengan Metode SVM

		Nilai Sebenarnya	
		Bukan NEET	NEET
Nilai Perkiraan	Bukan NEET	15363	2709
	NEET	1384	1459

Tabel 2 menunjukkan hasil validasi dari klasifikasi dengan metode SVM yang disajikan dalam *confusion matrix* untuk mengetahui seberapa tepat prediksi model dengan nilai yang sebenarnya. Berdasarkan Tabel 2, diketahui bahwa dari 18.072 penduduk muda yang diprediksi bukan termasuk NEET, metode SVM dapat dengan benar memprediksi 15.363 penduduk yang sebenaranya bukan termasuk NEET, sedangkan 2.709 penduduk yang diprediksi bukan termasuk NEET ternyata NEET. Sementara itu, dari 2.843 penduduk muda yang diprediksi termasuk NEET, metode ini dapat dengan benar mengklasifikasikan 1.459 penduduk yang sebenarnya NEET, sedangan 1.384 lainnya merupakan penduduk yang sebenarnya bukan tergolong NEET.

Tabel 3. Hasil *Confusion Matrix* dengan Metode *Random Forest*

		Nilai Sebenarnya	
		Bukan NEET	NEET
Nilai Perkiraan	Bukan NEET	16143	2965
	NEET	604	1203

Sementara itu, Tabel 3 menunjukkan hasil *confusion matrix* dengan metode *random forest*. Berdasarkan tabel 3, diketahui bahwa dari 19.108 penduduk muda yang diprediksi bukan termasuk NEET, metode *random forest* dapat dengan benar memprediksi 16.143 penduduk yang sebenaranya bukan termasuk NEET, sedangkan 2.965 penduduk yang diprediksi bukan termasuk NEET ternyata NEET. Sementara itu, dari 1.807 penduduk muda yang diprediksi termasuk NEET, metode ini dapat dengan benar mengklasifikasikan 1.203 penduduk yang sebenarnya NEET, sedangan 604 lainnya merupakan penduduk yang sebenarnya bukan tergolong NEET.

Tabel 4. Ukuran Perbandingan Evaluasi Model SVM dan *Random Forest*

Kriteria Evaluasi	SVM	*Random Forest*
Akurasi	0,8043	0,8294
95% CI	0,7989-0,8097	0,8242-0,8344
Sensitivity/*Recall*	0,9174	0,9693
Specificity	0,35	0,2886
Presisi/PPV	0,8501	0,8448
NPV	0,5132	0,6657

Dari *confusion matrix* yang diperoleh dari Tabel 2 dan Tabel 3, dapat ditentukan ukuran-ukuran evaluasi lain yang dapat digunakan untuk menentukan seberapa baik kinerja dari model yang terbentuk. Berdasarkan Tabel 4, diperoleh nilai akurasi dari model yang terbentuk oleh metode SVM dan *random forest* secara berturut-turut adalah sebesar 0,8043 dan 0,8294. Artinya, dari keseluruhan penduduk muda yang menjadi sampel penelitian, terdapat sebanyak 80,43 persen penduduk yang tergolong NEET dan penduduk bukan NEET yang dapat diklasifikasikan dengan benar oleh metode SVM. Sementara itu, penduduk yang tergolong NEET dan penduduk bukan NEET yang dapat diklasifikasikan dengan benar oleh metode *random forest* adalah sebesar 82,94 persen. Dengan demikian, *error* yang terbentuk dari metode SVM dan *random forest* secara berturut-turut adalah sebesar 19,57 persen dan 17,06 persen.

Nilai *sensitivity* dan *recall* adalah sebesar 0,9174 untuk metode SVM dan 0,9693 untuk metode *random forest* yang berarti bahwa metode SVM mampu memprediksi 91,74 persen dan metode *random forest* mampu memprediksi 96,93 persen dengan tepat penduduk muda bukan termasuk NEET yang sebenarnya bukan termasuk NEET. Nilai *specificity* yang dihasilkan oleh metode SVM dan *random forest* secara berturut-turut adalah sebesar 0,35 dan 0,2886 yang menunjukkan bahwa metode SVM mampu memprediksi dengan tepat 35 persen penduduk muda NEET yang sebenarnya termasuk NEET sedangkan metode *random forest* mampu memprediksi dengan tepat 28,86 persen penduduk muda NEET yang sebenarnya termasuk NEET.

Nilai presisi atau PPV adalah sebesar 0,8501 untuk metode SVM dan 0,8448 untuk metode *random forest* yang berarti bahwa dari seluruh penduduk muda yang diprediksi bukan merupakan NEET, metode SVM mampu mengklasifikasikan dengan benar prediksi tersebut sebesar 85,01 persen untuk metode SVM dan 84,48 persen untuk metode *random forest*. Sementara itu, NPV yang dihasilkan dari metode SVM dan *random forest* secara berturut-turut adalah sebesar 0,5132 dan 0,6657 yang menunjukkan bahwa dari seluruh penduduk muda yang diprediksi merupakan NEET, metode SVM mampu mengklasifikasikan 51,32 persen prediksi tersebut dengan benar sedangkan metode *random forest* mampu mengklasifikasikan 66,57 persen dengan benar.

Berdasarkan perbandingan ukuran evaluasi, dapat diketahui bahwa secara rata-rata, metode *random forest* memiliki performa yang lebih baik dibandingkan dengan metode SVM. Menurut Twala (2010), *Random forest* merupakan algoritma *machine learning* dengan *ensemble method* menggunakan *bagging* yang bertujuan agar memiliki performa yang lebih stabil pada perubahan kecil di dalam *dataset* sehingga dapat meningkatkan tingkat akurasi pada pengklasifikasian. Dengan demikian, metode *random forest* secara teoretis lebih unggul dibandingkan metode SVM dan metode *non-ensemble* lainnya karena melibatkan penggunaan beberapa model sekaligus (simultan) yang kemudian dilakukan *majority vote*. Sehingga, metode ini mampu menghasilkan performa yang lebih baik dibandingkan metode SVM yang hanya melibatkan penggunaan satu model.

5. Kesimpulan

Berdasarkan hasil penelitian mengenai pengklasifikasian status NEET terhadap penduduk muda di Indonesia dengan menggunakan metode SVM dan *random forest*, diperoleh kesimpulan bahwa metode *random forest* memberikan hasil akurasi yang lebih tinggi. Dengan demikian, metode *random forest* memiliki kemampuan yang lebih baik dalam mengklasifikasikan penduduk muda menurut status NEET sehingga dapat digunakan untuk memprediksi status NEET dalam rangka menunjang usaha pengurangan persentase NEET di Indonesia.

Kemudian, untuk pengembangan penelitian di masa depan, disarankan melakukan penambahan skema untuk mengatasi *imbalanced data* (*undersampling*, *oversampling*, dan *combined method*) mengingat kelas *dataset* memiliki rasio yang tidak seimbang. Selain itu, pengembangan juga dapat dilakukan dengan memodifikasi nilai parameter serta melakukan pengembangan terhadap metode yang digunakan untuk meningkatkan performa klasifikasi.

Daftar Pustaka

Amami, R., Ayed, D. B., & Ellouze, N. (2015). An Empirical comparison of SVM and some supervised learning algorithms for vowel recognition. *International Journal of Intelligent Information Processing (IJIIP)*, 3(1).

Badan Perencanaan Pembangunan Nasional. (2017). *Bonus Demografi 2030-2040: Strategi Indonesia Terkait Ketenagakerjaan dan Pendidikan*. Jakarta: Bappenas.

Badan Pusat Statistik. (2019). *Tingkat Pengangguran Terbuka Berdasarkan Kelompok Umur, 2015-2019*. Jakarta: BPS.

Bay, A. H., dan Blekeseanu, M. (2002). Youth, Unemployment, and Political Marginalisation. *International Journal of Social Welfare*, 11(2), 132-139.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics* (pp. 985-1022). Springer, Berlin, Heidelberg.

Feta, N. R. dan Ginanjar, A. R. (2019). "Komparasi Fungsi *Kernel* Metode *Support Vector Machine* Untuk Pemodelan Klasifikasi Terhadap Penyakit Tanaman Kedelai"., *BRITech, Jurnal Ilmiah Ilmu Komputer, Sains dan Teknologi Terapan*, vol.1(1), pp. 33-39.

Han, J., Kamber, M & Jian Pei. (2012). *Data Mining Concepts and Techniques: Third Edition*. USA. Morgan Kaufmann Publications.

International Labour Organization. (2017). *Indonesia Jobs Outlook 2017: Harnessing Technology for Growth and Job Creation*. Jakarta: ILO.

Maulana, J. P. dan Irhamah. (2018). "Klasifikasi Kabupaten di Provinsi Jawa Timur Berdasarkan Indikator Daerah Tertinggal dengan metode *Support Vector Machine* (SVM) dan *Entropy Based Fuzzy Support Vector Machine* (EFSVM)"., *Inferensi*, vol. 1(1), pp. 9-15.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), 128-138.

Revar, A., Andhariya, M., Sutariya, D., & Bhavsar, M. (2010). Load balancing in grid environment using machine learning-innovative approach. *International Journal of Computer Applications*, 8(10), 31-34.

Sadewo, M. G., Windarto, A. P., & D. Hartama D. (2017). Penerapan *Data Mining* Pada Populasi Daging Ayam RAS Pedaging di Indonesia Berdasarkan Provinsi Menggunakan K-Means Clustering. *InfoTekJar (Jurnal Nasional Informatika dan Teknik Jaringan)*. 2017; 2(1): 60-67.

Salazar, D. A., Vélez, J. I., & Salazar, J. C. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate?. *Revista Colombiana de Estadística*, 35(SPE2), 223-237.

Scarpetta, S., A. Sonnet and T. Manfredi. (2010). "Rising Youth Unemployment During the Crisis: How to Prevent Negative Long-term Consequences on a Generation?", *OECD Social, Employment and Migration Working Papers*, No. 106, OECD Publishing.

Schell, George P. dan Raymond McLeod. (2007). *Management Information Systems*. New Jersey. Pearson/Prentice Hall.

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 37(4), 3326-3336.

Tzotsos, A., & Argialas, D. (2008). Support vector machine classification for object-based image analysis. In *Object-Based Image Analysis* (pp. 663-677). Springer, Berlin, Heidelberg.

Wickremeratne, N., & Dunusinghe, P. (2018). Youth Not in Education, Employment and Training (NEET) in Sri Lanka. *Advances in Economics and Business*, 6(5), 339-352.