Xplore (2022), Vol. 11, No. 2, 147-156
DOI: https://doi.org/10.29244/xplore.v11i2.919

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Perbandingan Pengklasifikasian Metode *Support Vector Machine* dan *Random Forest* (Kasus Perusahaan Kebun Kelapa Sawit)*

Nabila Destyana Achmad[1], Agus M Soleh[2‡], and Akbar Rizki[3]

[1]PT Indekstat, Indonesia
[2,3]Department of Statistics, IPB University, Indonesia
‡corresponding author: agusms@apps.ipb.ac.id

Abstract

Palm oil is one of the leading commodities that support the economy in Indonesia. One of the companies engaged in the oil palm plantation sector has 146 units of oil palm plantations. It is very important to optimize oil palm production, so it is necessary to classify the status of plantation units. Classification aims to predict new plantation units and find the most important variables in the modeling process. The variables used were the status of the garden as a response variable and nine explanatory variables, namely harvested area, rainfall, percentage of normal fruit, fresh fruit bunches production, oil palm loose fruits, production, harvest job performance, harvesting rotation, and farmers. The classification process is carried out using the Support Vector Machine and Random Forest methods to find which method is the best. The data is divided into 80% training data and 20% test data with ten iterations so that ten models are produced for each method. Comparing accuracy value, F1 score, and Area Under Curve (AUC) to evaluate the model. The modeling results show that the random forest method has better performance than the SVM method. The random forest has an average occuracy, F1 score, and AUC, respectively, 90%, 86%, and 89%. Variables of harvest job performance, oil palm loose fruits, harvested area, rainfall, and harvesting rotation are important variables that contribute more than 10% of the model. The results of the research are used for the evaluation and development process of oil palm companies by taking into account the result of important variables that affect productivity and predictive results of new plantation units.

Keywords: units of oil palm, classification, support vector machine, random forest

* Received: Jan 2022; Reviewed: May 2022; Published: May 2022

1. Pendahuluan

Perkebunan kelapa sawit merupakan salah satu sektor utama yang telah berkontribusi banyak dalam pembangunan ekonomi Indonesia. Sektor perkebunan kelapa sawit ini telah memberikan jutaan lapangan pekerjaan baik secara langsung dan tidak langsung (Kemenko 2021). Selain itu, sektor sawit ini juga telah membuat Indonesia menjadi produsen sawit terbesar di dunia yang dimana hal tersebut sangat berdampak positif bagi pemasukan negara.

Namun, meskipun Indonesia adalah produsen sawit terbesar dunia, nyatanya produktivitas kebun sawit di negara ini masih tergolong rendah. Bahkan, apabila dibandingkan dengan negara Malaysia dan Thailand, Indonesia masih kalah dalam hal produktivitas sawit. Kepala Badan Pengelola Dana Perkebunan Kelapa Sawit (BPDKS), Eddy Abdurrachman, mencatat nilai ekspor sawit tahun 2020 mencapai US$ 22,97 miliar atau setara Rp 321,5 triliun. Namun demikian, secara volume ekspor produksi kelapa sawit ini mengalami penurunan. Sepanjang tahun 2020, volume ekspor produk sawit mencapai sekitar 34 juta ton, dimana hasil ini lebih rendah dari tahun 2019 yang mencapai 37,39 ton (Pebrianto 2021). Penurunan produktivitas ini menyebabkan perlunya pengoptimalan unit kebun kelapa sawit agar tidak terus mengalami penurunan produksi, yaitu dengan mengetahui peubah-peubah penting yang memengaruhi produktivitas unit kebun kelapa sawit.

Setiap unit kebun kelapa sawit di semua perusahaan pasti memiliki target produktivitasnya tersendiri yang harus dicapai. Sehingga perusahaan dapat melihat dan melabeli unit kebun kelapa sawit mana yang termasuk produktif dan kurang produktif. Biasanya, unit kebun kelapa sawit akan terus bertambah sesuai dengan kebijakan perusahaan dan kebutuhan ekspor negara. Penambahan unit kebun kelapa sawit ini pun telah menjadi suatu tantangan tersendiri bagi perusahaan untuk dapat memprediksi status unit kebun kelapa sawit agar dijadikan bahan evaluasi dan pengembangan perusahaan. Oleh karena itu, untuk mengoptimalkan evaluasi dan pengembangan perusahaan dapat dilakukan pengklasifkasian kebun untuk melakukan prediksi label unit kebun dan melihat peubah penting yang memengaruhi hasil produktivitas unit kebun tersebut. Metode yang dapat digunakan untuk menangani permasalahan ini adalah metode klasifikasi.

Penelitian menggunakan metode klasifikasi dilakukan oleh Qi-na dan Ting-hui (2020) dalam membandingkan kinerja metode klasifikasi *Naïve Bayes* dengan *Support Vector Machine* (SVM*).* Hasil penelitian tersebut menunjukkan bahwa metode SVM menghasilkan nilai akurasi terbaik. Metode klasifikasi lainnya yaitu *random forest*, metode yang menghasilkan performa yang sangat baik menurut hasil penelitian Santoso (2020) dalam mengklasifikasikan penyakit batang busuk pada sawit. Kedua penelitian ini mendasari peneliti untuk melihat model terbaik dari hasil pengklasifikasian metode *Support Vector Machine* (SVM) dan *random forest* yang akan digunakan untuk mengklasifikasikan unit kebun kelapa sawit dan mengidentifikasi peubah penting yang berpengaruh terhadap status unit kebun kelapa sawit.

2. Metodologi

2.1 Data

Data yang digunakan pada penelitian ini berasal dari data salah satu perusahaan kelapa sawit periode Januari – Juni 2021. Data terdiri dari sepuluh peubah

dengan jumlah amatan sebanyak 146 unit kebun kelapa sawit. Peubah yang digunakan adalah

status kebun sebagai peubah respon dan sebanyak sembilan peubah penjelas yang disajikan pada Tabel 1.

Status kebun merupakan peubah kategorik yang terdiri atas 1=Produktif dan 0= Kurang produktif. Status kebun merupakan peubah kategorik yang terdiri atas 1=Produktif dan 0= Kurang produktif. Penstatusan atau pelabelan ini berdasarkan tercapai atau tidaknya RKO (Rencana Kerja Operasional) yang telah ditetapkan oleh perusahaan pada setiap unit kebun kelapa sawit. Unit kebun yang kurang produktif adalah unit kebun yang tidak dapat mencapai RKO yang ditargetkan, sedangkan unit kebun produktif adalah unit kebun yang dapat mencapai RKO yang telah ditetapkan.

Peubah penjelas X1 yaitu peubah luas panen yang merupakan hasil tanaman yang dipungut hasilnya setelah cukup umur. Lalu peubah penjelas X2 adalah peubah curah hujan yang merupakan ketinggian air hujan yang terkumpul dalam penakar hujan. Peubah penjelas X3 yaitu peubah buah normal merupakan persenan buah normal yang dihasilkan selama panen. Selanjutnya, peubah X4 tandan buah segar merupakan sebutan dari buah kelapa sawit, sedangkan brondolan (X5) adalah buah sawit yang terjatuh sendiri karena sudah cukup matang. Peubah penjelas X7 yaitu prestasi panen adalah bobot total tandan yang didapatkan seorang pemanen. Peubah luas kelompok pusingan panen (X8) merupakan rotasi petani mengelilingi kebun.

Tabel 1 Peubah yang digunakan

Kode	Peubah	Jenis	Satuan
Y	Status Kebun	Kategorik (1=Produktif, 0=Kurang Produktif)	-
X1	Luas Panen	Numerik	Hektare
X2	Curah Hujan	Numerik	Milimiter
X3	Buah Normal	Numerik	Persen
X4	Tandan Buah Segar	Numerik	Ribu Tandan
X5	Brondolan	Numerik	Persen
X6	Produksi	Numerik	Ton
X7	Prestasi Panen	Numerik	Ton/Petani
X8	Luas Kelompok Pusingan Panen	Numerik	Hektar
X9	Tenaga Kerja	Numerik	Orang

.

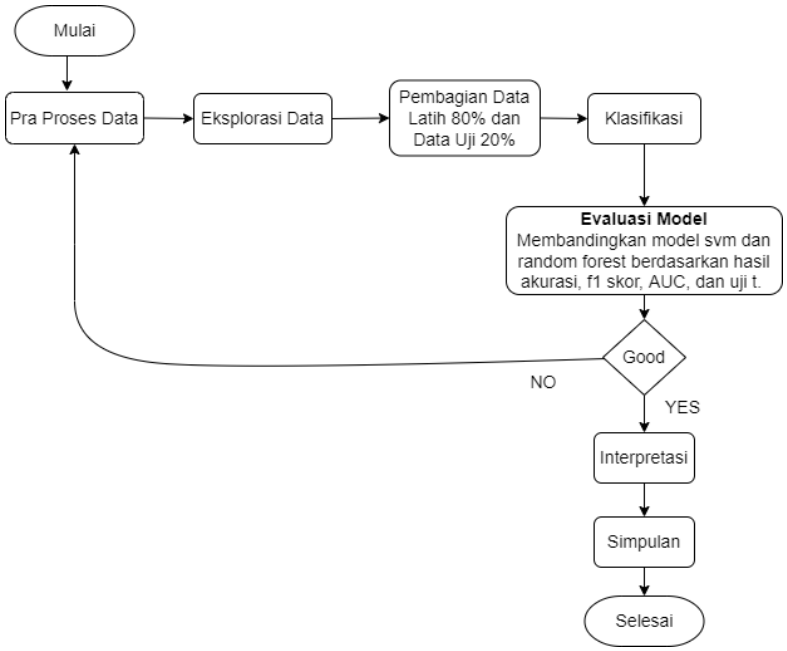
2.2 Prosedur Analisis Data

Tahapan analisis pada penelitian ini adalah sebagai berikut **:**

1. Melakukan pra proses data yaitu mengatasi nilai hilang atau *missing value* dengan total 14 data hilang dari 146 data. *Missing value* terdapat di peubah curah hujan dan luas kelompok pusingan panen Proses mengatasi *missing value* dengan menggunakan imputasi nilai median masing-masing peubah yang terdapat *missing value.*

2. Mengeksplorasi data dengan melihat sebaran data menggunakan diagram batang, dan *pie chart* untuk melihat proporsi label.
3. Melakukan standardisasi data sehingga data memiliki nilai rataan nol dan nilai standard deviasi sama dengan satu (Thara *et al.* 2019) dengan rumus $z = \frac{x-\mu}{\sigma}$.
4. Melakukan pembagian data menjadi 80% data latih dan 20% data uji dengan melakukan ulangan 10 kali untuk proses evaluasi model.
5. Melakukan pemodelan menggunakan metode *Support Vector Machine*
   a. Mencari nilai *hyperplane* yang memiliki nilai margin maksimum dengan meminimalkan nilai $^{min}_{w}\tau(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||^2$ dengan batasan $y_i(\boldsymbol{x_i}\,.\,\mathbf{w} + b) - 1 \geq 0, \forall_i$.
   b. Menjumlahkan persamaan yang didapatkan dari poin a untuk mendapatkan nilai **w** dan b untuk membuat *hyperplane* sebagai fungsi klasifikasi.
   c. Melakukan evaluasi terhadap data uji dengan melihat nilai akurasi, skor F1, dan nilai AUC.
6. Melakukan pemodelan klasifikasi dengan *random forest* yang menerapkan metode bootstrap aggregating (*bagging*) dan *random feature selection* (Pamungkas *et al.* 2020). banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), yang dilakukan dengan cara berikut (Breiman 2001):
   a. Melakukan tahapan *resampling bootstrap with replacement* pada data latih sebanyak 100 kali dikarenakan untuk metode *random foret* ini akan menggunakan 100 pohon.
   b. Memilih m peubah penjelas secara acak pada setiap simpul, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan *feature randomness.* Dalam penelitian ini menggunakan m = $\sqrt{p}$.
   c. Melakukan evaluasi terhadap data uji dengan melihat nilai akurasi, skor F1, dan nilai AUC.
7. Membandingkan hasil model SVM dan *random forest* berdasarkan hasil rataan nilai akurasi, skor F1 dan nilai AUC.
8. Menginterpretasi hasil.

Adapun diagram alir untuk analisis penelitian ini disajikan pada Gambar 1 di bawah ini.



Gambar 1 Diagram alir prosedur analisis data

3. Hasil dan Pembahasan

3.1 *Pre-processing* Data

Unit kebun kelapa sawit di perusahaan ini memiliki jumlah keseluruhan unit sebanyak 146 unit kebun. Data tersebut merupakan hasil dari pemantauan perusahaan terhadap masing-masing unit kebun selama kurun waktu 6 bulan (Januari 2021-Juni 2021). beberapa kebun sudah memberikan kinerja yang baik dan beberapa lainnya masih belum mencapai target yang diharapkan oleh perusahaan dalam kurun waktu tersebut. Setelah ditelusuri lebih lanjut, ternyata terdapat sekitar 9,5% atau sebanyak 14 data yang hilang. Semua amatan yang hilang tersebut berasal dari peubah luasan kelompok pusingan panen dan curah hujan.

Hasil uji Kolmogorov-smirnov menunjukkan bahwa kedua peubah tidak berdistribusi normal, dikarenakan p-value yang dihasilkan hampir mendekati nol. Dikarenakan data tidak berdistribusi normal, maka data hilang ini ditangani dengan menggunakan imputasi nilai median. Metode imputasi median digunakan untuk mengisi nilai pada data yang tidak berdistribusi normal (Jadhav *et al.* 2019). Tabel 2 menunjukkan sejumlah data sebelum dan sesudah imputasi.

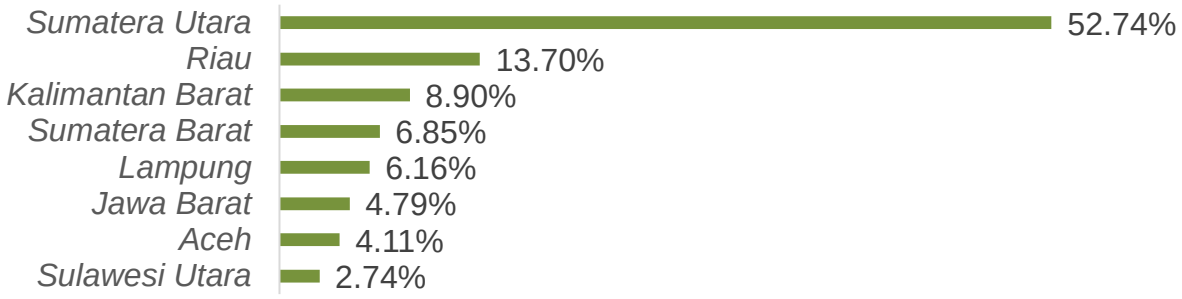
Tabel 2 Data sebelum dan sesudah imputasi

Data	Sebelum Imputasi Median		Setelah Imputasi Median	
	Curah Hujan	Luas Kelompok Pusingan	Curah Hujan	Luas Kelompok Pusingan
1	NA	NA	926	470
2	681	NA	681	470
3	783	NA	783	470
4	1493	NA	1493	470
5	1169	NA	1169	470
6	464	NA	464	470
7	1269	NA	1269	470
8	785	NA	785	470
9	1801	NA	1801	470
10	1784	NA	1784	470
11	705	NA	705	470
12	1468	NA	1468	470
13	725	NA	725	470

3.2 Eksplorasi Data

Unit kebun kelapa sawit di perusahaan memiliki jumlah keseluruhan unit sebanyak 146 unit kebun yang tersebar di sepuluh provinsi di Indonesia. Proporsi sebaran unit kebun kelapa sawit ini dapat dilihat pada Gambar 2. Unit kebun kelapa sawit mayoritas berada di provinsi Sumatera Utara yaitu sebesar 52,74% atau sebanyak 77 unit kebun berada di provinsi tersebut. Hampir setengah dari kebun yang ada di provinsi Sumatera Utara ini berada di kota Medan, yaitu sebanyak 24,66% kebun (36 unit), sisanya menyebar di beberapa kabupaten provinsi Sumatera Utara. Provinsi Riau memiliki sebanyak 13,70% unit kebun sawit yang menempati urutan

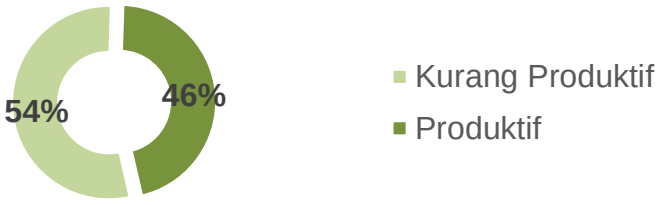
kedua pada sebaran tersebut. Selanjutnya, wilayah yang memiliki sedikit unit kebun kelapa sawit yaitu wilayah Sulawesi Utara yang hanya sekitar 2,74% unit kebun atau sekitar 4 unit kebun saja dari total keseluruhan 146 unit kebun kelapa sawit.



Gambar 2 Sebaran unit kebun kelapa sawit berdasarkan provinsi

Unit data kebun kelapa sawit dibagi menjadi dua kategori, yaitu unit kebun produktif (1) dan unit kebun kurang produktif (0). Pengkategorian ini dilihat dari hasil pemantauan selama 6 bulan periode Januari 2021 – Juni 2021. Unit kebun produktif adalah kebun yang mencapai atau melebihi target perusahaan. Sedangkan unit kebun kurang produktif adalah unit kebun yang masih belum mencapai target yang telah diterapkan oleh perusahaan.

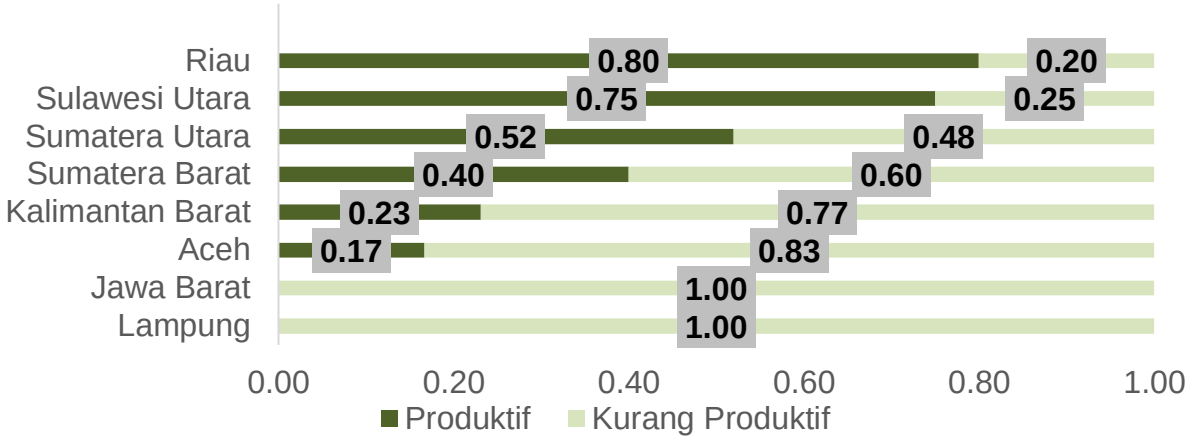
Proporsi masing-masing label dapat dilihat pada Gambar 3. Berdasarkan Gambar 3, menunjukkan bahwa proporsi antar labelnya bisa dikatakan seimbang dengan persentase unit kebun produktif sebesar 46% dan unit kebun kurang produktif sebesar 54%. Sedangkan data dengan kelas biner dikatakan tidak seimbang apabila rasio ukuran sampel kelas minor dengan kelas mayor mencapai 1:100, 1:1000, atau bahkan lebih besar dari itu (Sun et al. 2009). Hal ini membuktikan bahwa data unit kebun kelapa sawit tersebut tidak memiliki masalah data yang tidak seimbang. Sehingga tidak diperlukan penerapan metode untuk mengatasi masalah data tidak seimbang.



Gambar 3 Persentase label unit kebun kelapa sawit

Pada Gambar 4, terlihat bahwa provinsi Riau dan Sulawesi Utara memiliki proporsi kebun berstatus produktif lebih dari 50%. Provinsi Sumatera Utara memiliki status unit kebun yang seimbang antara status produktif dan kurang produktif. Terdapat dua provinsi yang bahkan tidak memiliki kebun berstatus produktif karena

seluruh kebun yang berada di wilayah tersebut berstatus kurang produktif, yaitu provinsi Lampung dan Jawa Barat.



Gambar 4 Sebaran status kebun per provinsi

3.3 Perbandingan Hasil Klasifikasi SVM dan *Random Forest*

Performa kedua metode yaitu hasil pengklasifikasian SVM dan *random forest* diperoleh setelah melakukan semua prosedur analisis, hal tersebut yang disajikan pada Tabel 3. Performa yang diperlihatkan terdiri dari nilai akurasi, skor F1 dan nilai AUC. Hasil pemodelan menunjukkan bahwa rataan ketiga evaluasi model menunjukkan model *random forest* memiliki performa yang lebih baik dibandingkan dengan model SVM. Model *random forest* menghasilkan rataan akurasi 90%, skor F1 86% dan skor AUC 89%. Namun model SVM hanya menghasilkan rataan akurasi sekitar 81%, skor F1 80% dan skor AUC 81%.

Tabel 3 Hasil pemodelan klasifikasi SVM dan *random forest*

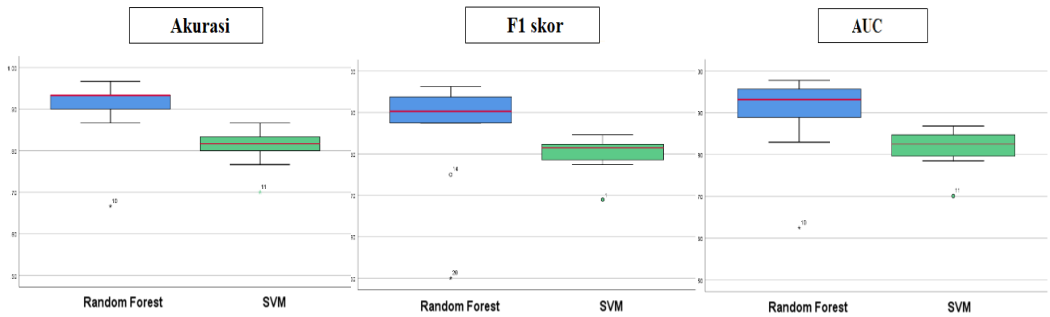
Model	**Accuracy**		**Skor F1**		**AUC**	
	SVM	RF	SVM	RF	SVM	RF
1	0,70	0,93	0,69	0,88	0,70	0,89
2	0,80	0,93	0,79	0,88	0,80	0,96
3	0,77	0,93	0,77	0,93	0,79	0,94
4	0,83	0,87	0,81	0,75	0,83	0,83
5	0,80	0,97	0,82	0,94	0,80	0,98
6	0,83	0,97	0,81	0,96	0,87	0,97
7	0,83	0,90	0,81	0,89	0,85	0,90
8	0,80	0,93	0,79	0,94	0,82	0,93
9	0,83	0,93	0,83	0,92	0,83	0,93
10	0,87	0,67	0,85	0,50	0,86	0,63
Rataan	**0,81**	**0,90**	**0,80**	**0,86**	**0,81**	**0,89**

Gambar 5 menunjukkan visualisasi performa akurasi kedua metode dalam *boxplot*. Kedua metode menghasilkan performa yang sangat berbeda, dapat dilihat bahwa metode *random forest* memperoleh nilai rataan yang lebih besar dibandingkan dengan SVM dalam ketiga performa evaluasi. Nilai maksimum performa metode *random forest* pada akurasi, skor F1 dan AUC berturut-turut yaitu 97%, 96%, dan 98%. Lalu nilai

maksimum performa metode SVM yaitu sebesar 87% untuk akurasi, 85% skor F1, dan 87% nilai maksimum AUC.

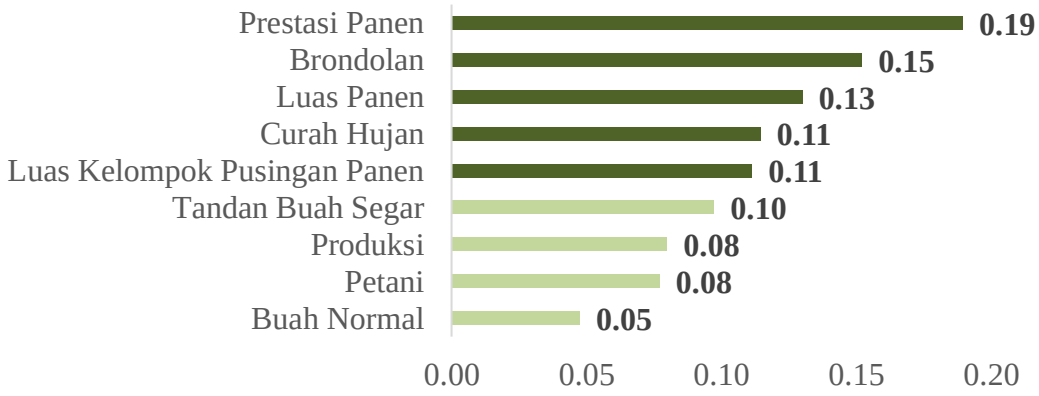
Perbandingan hasil model juga akan dilihat dari nilai *skewness* dan kurtosis kedua metode. Nilai *skewness* merupakan nilai kemenjuluran sebaran data yang bernilai nol akan mengikuti sebaran data normal (simetris), bernilai negatif bahwa sebaran data memiliki ekor di sebelah kiri yang lebih panjang dari ekor kanan, dan kondisi sebaliknya bagi *skewness* yang bernilai positif. Selanjutnya nilai kurtosis adalah ukuran keruncingan data. Semakin tinggi nilai kurtosis, maka kurva akan semakin runcing.

Hasil distribusi nilai akurasi model *random forest* memiliki nilai *skewness* sebesar -2,5 dan nilai kurtosis sebesar 7,1. Model SVM memiliki nilai *skewness* sebesar -1,3 dan nilai kurtosis sebesar 2,5. Berdasarkan hasil *skewness* dan kurtosis model yang dihasilkan oleh metode *random forest* dan SVM menunjukkan bahwa kedua metode menjulur negatif dikarenakan keduanya menghasilkan nilai *skewness* negatif. Namun dikarenakan hasil *skewness* dan kurtosis *random forest* lebih besar dibandingkan SVM, maka model *random forest* memiliki kurva yang lebih menjulur ke kiri dan lebih runcing. Hal ini menunjukkan bahwa nilai akurasi model *random forest* lebih memusat di satu titik yaitu di sebelah kanan sebaran, dengan kata lain nilai akurasi model *random forest* secara umum memusat pada nilai yang tinggi. Nilai *skewness* dan kurtosis untuk skor F1 dan AUC pada kedua model menghasilkan kesimpulan yang sama. Oleh karena itu, penetapan performa model terbaik pada penelitian ini dihasilkan oleh metode *random forest.*



Gambar 5 *Boxplot* evaluasi kedua metode

Peubah penjelas yang paling penting (*feature importance*) dalam pemodelan klasifikasi *random forest* ditunjukkan pada Gambar 6. Berdasarkan hasil persenan peubah penting, akan diambil peubah penting dengan nilai kepentingan lebih dari 10%. Hal ini dikarenakan hasil dari peubah penting ini tidak terlalu jauh berbeda. Oleh karena itu, terdapat lima peubah penting yang memiliki kepentingan lebih dari 10%, yaitu peubah prestasi panen, brondolan, luas panen, curah hujan, dan luas kelompok pusingan panen.



Gambar 6 Peubah penting dalam pemodelan *random forest*

4. Simpulan

Hasil dari perbandingan pengklasifikasian metode SVM dan *random forest* menghasilkan bahwa metode *random forest* memiliki performa yang lebih baik dibandingkan SVM. Kesimpulan didapatkan dari hasil membandingkan nilai rataan akurasi, skor F1, dan AUC hasil pemodelan kedua metode. Selain itu juga dilihat dari hasil sebaran nilai pemodelan dengan membandingkan nilai *skewness* dan kurtosis.

Tingkat kepentingan peubah dipilih berdasarkan peubah yang memiliki kontribusi kepentingan lebih dari 10% dalam model, yaitu peubah prestasi panen, brondolan, luas panen, curah hujan, dan luas kelompok pusingan panen. Oleh karena itu, diharapkan perusahaan kelapa sawit dapat lebih memperhatikan kelima peubah tersebut pada setiap unit kebun sawit.

Daftar Pustaka

Breiman L. 2001. *Machine Learning: Random Forests*. California : Kluwer Academic. DOI: 10.1023/A:1010933404324

Jadhav A, Pramod D, Ramanathan K. 2019. *Comparison of performance of data imputation methods for numeric dataset.* AAI. DOI : 10.1080/08839514.2019.1637138.

[Kemenko] Kementrian Koordinator Bidang Perekonomian. 2021. PERS : Industri Kelapa Sawit Indonesia Menjaga Keseimbangan Aspek Sosial, Ekonomi, dan Lingkungan. Siaran pers NO HM.4.6/82/SET.M.EKON.3/04/2021.

Qi-na, Ting-hui. 2020. *Research on the application of naïve bayes ad support vector machine algorithm on excersises classification*. J. Phys.: Conf. Ser. 14377 012071.

Pamungkas FJ, Prasetyo BD, Kharisudin I. 2020. Perbandingan metode klasifikasi supervised learning pada data bank customers menggunakan python. PRISMA. 3:689-694

Pebrianto F. 2021 Feb 10. Ekspor produk sawit 2020 capai rp 321 triliun, tumbuh 136 persen. tempo.co. Rubrik Bisnis. [diakses 2021 Nov 25]. https://bisnis.tempo.co/read/1431588/ekspor-produk-sawit-2020-capai-rp-321-t-tumbuh-136-persen/full&view=okn

Santoso H. 2020. Performance of random forest group basal stem rot disease classification caused by Ganoderma boninense in oil palm plantation. *Jurnal Penelitian Kelapa Sawit.* 28(3): 133-146. https://doi.org/10.22302/iopri.jur.jpks.v28i3.116.

Sun Y, Wong AKC, Kamel MS. 2009. Classification of imbalanced data: a review. International Journal of Pattern Recognition. 23(4):687-719.

Thara DK, PremaSudha BG, dan Fan Xiong. 2019. *Auto-detection of epileptic seizure events using a deep neural network with different feature scaling techniques. Pattern Recognition Letters.* 544-550.