

PEMANFAATAN ENSEMBLE LEARNING DAN PENGINDERAAN JAUH UNTUK PENGKLASIFIKASIAN JENIS LAHAN PADI

Arif Handoyo Marsuhandi¹, Agus Mohamad Soleh², Hari Wijayanto³,
Dede Dirgahayu Domiri⁴

¹Badan Pusat Statistik Provinsi Sulawesi Utara

²Institut Pertanian Bogor

³Institut Pertanian Bogor

⁴Lembaga Penerbangan dan Antariksa Nasional

arifhandoyo@bps.go.id

ABSTRAK

Pertanian adalah bidang yang sangat penting di Indonesia, sektor ini di tahun 2017 mencatat penyerapan tenaga kerja sebanyak 29.68% dari total seluruh pekerja (BPS, 2018), namun pentingnya sektor pertanian ini berbanding terbalik dengan data pertanian yang tersedia. Tahun 1998 Badan Pusat Statistik (BPS) bersama *Japan International Cooperation Agency* (JICA) telah mengisyaratkan overestimasi luas panen sekitar 17,07 persen. Ketidakuratan data pertanian ini kemudian diperbaiki pada tahun 2018 melalui kerjasama para *stakeholder* dengan menyusun suatu metodologi baru dalam menghitung luas lahan yang diberi nama kerangka sampel area. Selain metodologi yang sudah diperbarui, kemajuan teknologi dan teknik analisis di bidang ilmu pengetahuan juga mendukung perbaikan data pertanian. Citra satelit dan teknik klasifikasi menggunakan ensemble learning dapat dimanfaatkan dalam mengklasifikasikan jenis lahan padi. Pada penelitian ini digunakan citra satelit yang berasal dari *United States Geological Survey* (USGS) yaitu Landsat 8 dan teknik klasifikasi *ensemble learning*. Citra satelit dimanfaatkan untuk mengekstrak indeks vegetatif dari koordinat koordinat yang diteliti, sedangkan ensemble learning yang digunakan dalam penelitian ini yaitu *Random Forest* dan *Boosting*. Hasil pengolahan data menunjukkan *Random Forest* memiliki akurasi yang lebih tinggi daripada *Boosting* yaitu dengan nilai 76,52 untuk *Random Forest* dan 75,60 untuk *Boosting*. Keunggulan *Random Forest* terhadap *Boosting* tidak hanya dari sisi tingkat akurasi saja namun juga dari kestabilan model yang dibentuk.

Kata kunci: klasifikasi, ensemble, kerangka sampling area, penginderaan jauh.

ABSTRACT

Agriculture is a very important sector in Indonesia, this sector in 2017 recorded a labor absorption of 29.68% of the total workers (BPS, 2018), but the importance of the agricultural sector is inversely proportional to the available agricultural data. In 1998 the Central Statistics Agency (BPS) together with the Japan International Cooperation Agency (JICA) had indicated an estimated harvest area of around 17.07 percent. The inaccuracy of agricultural data was then corrected in 2018 through the collaboration of stakeholders by developing a new methodology in calculating the area of land that was named the area sample framework. In addition to the updated methodology, technological advances and analytical techniques in the field of science also support the improvement of agricultural data. Satellite imagery and classification techniques using ensemble learning can be utilized in classifying types of rice fields. In this study, satellite images from the United States Geological Survey (USGS) are used, namely Landsat 8 and ensemble learning classification techniques. Satellite imagery is used to extract the vegetative index from the coordinates studied, while the ensemble learning used in this study is Random Forest and Boosting. The results of data processing show that Random Forest has a higher accuracy than Boosting, with a value of 76.52 for Random Forest and 75.60 for Boosting. The advantages of Random Forest over Boosting are not only in terms of accuracy but also from the stability of the model.

Keywords: classification, ensemble, area sampling framework, remote sensing.

PENDAHULUAN

LATAR BELAKANG

Pertanian tercatat menjadi sektor kehidupan ekonomi utama di Indonesia, pada tahun 2017 tercatat 26,68% Pekerja di Indonesia memiliki lapangan pekerjaan pada sektor pertanian (BPS, 2018). Besarnya *share* lapangan pekerjaan di bidang pertanian menyebabkan setiap keputusan atau

kebijakan yang diambil berdasarkan data yang ada harus akurat dan cepat, namun di lapangan data produksi padi di Indonesia sejak 1997 diduga terjadi *overestimate*. sebesar 17,07 persen(BPS Jawa Tengah 2018), pada tahun 2018 metode kerangka sampel area (KSA) diterapkan oleh Badan Pusat Statistik (BPS) dengan Badan Pengkajian dan Penerapan Teknologi (BPPT) dalam rangka memperbaiki metodologi pengumpulan data statistik pertanian yang konvensional menjadi lebih objektif dan modern. Metode KSA sendiri merupakan suatu survey yang dilaksanakan setiap bulan pada area yang terpilih menjadi sampel.

Pemanfaatan citra satelit yang merupakan salah satu big data yang sangat berguna untuk mengukur indeks vegetasi padi, mengelompokkan penggunaan lahan, dan lain-lain. Pemanfaatan citra satelit ini didukung oleh penemuan mengenai indeks-indeks spektral yang dapat digunakan sebagai peubah prediktor.

Teknik analisis yang semakin berkembang juga menjadikan pemanfaatan data dari citra satelit semakin mudah, beberapa penelitian sudah menerapkannya, hal ini dimungkinkan karena data data citra penginderaan jarak jauh sudah terbuka secara umum untuk dapat dimanfaatkan seluas-luasnya. Selain itu teknik analisis yang cenderung lebih mudah diterapkan dalam hal ini analisis klasifikasi, sangat mempermudah bagi peneliti yang ingin melakukan klasifikasi tanpa belajar secara mendalam.

TUJUAN PENELITIAN

Penelitian ini bertujuan untuk memilih model klasifikasi dari metode ensemble dengan akurasi tertinggi serta melihat peubah prediktor yang memiliki pengaruh yang besar terhadap model klasifikasi yang dibentuk.

TINJAUAN PUSTAKA

Ensemble Learning

Metode *ensemble learning* merupakan bagian dari *machine learning* yang memiliki tujuan utama untuk memprediksi nilai respon berdasarkan model klasifikasi dari set contoh yang telah memiliki peubah respon. *ensemble learning* dapat menghasilkan clasifier yang sangat tinggi akurasi dengan cara menggabungkan clasifier-clasifier yang memiliki akurasi yang rendah (Dietterich, 2000). Metode ensemble memiliki kelebihan apabila

Boosting

Metode *boosting* merupakan suatu algoritma yang secara teoritis dapat digunakan untuk mengurangi secara signifikan kesalahan pada berbagai macam "*weak*" *learning algorithm* yang secara konsisten menghasilkan *classifier* yang lebih baik sedikit pendugaannya daripada pendugaan secara acak(Freund dan Schapire 1996). Metode *boosting* menggunakan pembobotan sebagai ganti *bootstrap* dalam pengulangan untuk menghasilkan beberapa *weak classifier* yang selanjutnya dilakukan *voting* untuk mendapatkan *classifier boosting*. Boosting akan menghasilkan akurasi yang baik apabila dalam pengulangan *classifier* yang terbentuk mendekati nilai akurasi 50%. Berikut adalah tahapan dan algoritma dari *boosting*

Tahap inisiasi : untuk setiap observasi x , tetapkan nilai $D(x) = \frac{1}{N}$ dimana N adalah jumlah observasi.

Tahap pengulangan : untuk nilai $t = 1 \dots T$:

1. Cari *weak classifier* terbaik $H_t(x)$ menggunakan bobot $D_t(x)$
2. Hitung rasio galat ε_t

$$\varepsilon_t = \sum_{i=1}^N D_t(x) \cdot I[y_i \neq H_t(x_i)]$$

3. Tambahkah pembobot α_t ke dalam *classifier* $H_t(x)$ dimana

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right)$$

4. Untuk setiap nilai x_i , $D(x_i) = D(x_i) \cdot \exp(\alpha_t \cdot I[y_i \neq H_t(x_i)])$
5. Normalisasi $D(x_i)$ sehingga $\sum_{i=1}^N D_t(x_i) = 1$

$$H(x) = \text{SIGN} \left[\sum_{t=1}^T \alpha_t H_t(x) \right]$$

Random Forest

Random forests merupakan suatu kombinasi dari banyak pohon prediksi dimana tiap pohon bergantung pada nilai dari vektor contoh acak yang saling bebas dan berdistribusi sama pada semua pohon yang ada di dalam *Random Forests* (Breiman 2001). Random Forest memiliki tahapan yang mirip dengan bagging dalam *resampling* dan *voting*, namun yang berbeda adalah dalam pembentukan *classifier*.

Fase Pertumbuhan Padi

Menurut BPS dalam KSA 2018, Fase pertumbuhan padi meliputi :

1. Vegetatif awal: Fase tumbuh mulai dari awal tanam sampai anakan maksimum (1-35 hari).
2. Vegetatif akhir: Fase tumbuh mulai dari anakan maksimum sampai sebelum keluar malai (35-55).
3. Generatif: Fase tumbuh mulai dari keluar malai, pematangan, sampai sebelum panen (55-105 hari).
4. Panen: Fase pada saat padi sedang atau sudah dipanen.
5. Pesiapan lahan: Fase dimana lahan sawah mulai diolah untuk persiapan tanam padi.
6. Puso: Apabila terjadi serangan OPT atau bencana, sehingga produksi padi kurang 11 % dari normal.

Dilihat dari indeks vegetatifnya, tanaman padi memiliki pola tertentu mengikuti hari setelah tanam (HST). Dari hasil penelitian mengenai model pertumbuhan padi menggunakan data EVI MODIS temporal di pulau sulawesi (Dirgahayu, 2015). Tanaman padi apabila diukur menggunakan indeks vegetatif EVI polanya akan memiliki 1-3 puncak tiap periode tanam selama selang waktu 112-120 hari.

Kerangka Sample Area(KSA)

Survei Kerangka Sampel Area (KSA) adalah survei berbasis area yang dilakukan dengan pengamatan langsung terhadap sampel segmen dan bertujuan untuk mengestimasi luasan dengan ekstrapolasi dari sampel ke populasi dalam periode yang relatif pendek (rapid estimate). KSA memanfaatkan citra satelit dalam pembentukan grid segmen dan subsegmen serta teknologi Handphone Android untuk melakukan pemotretan dan pendataan pada koordinat area terpilih. Hasil dari KSA berupa koordinat dan label dari subsegmen. Label subsegmen ini berupa kode yaitu :

1. Vegetatif awal
2. Vegetatif akhir
3. Generatif
4. Panen
5. Pesiapan Lahan
6. Puso
7. Sawah bukan padi
8. Bukan sawah
9. Panen diantara Survey
10. Bera

Citra Satelit Landsat 8

Satelit Landsat merupakan produk dari *U.S. Geological Survey* yang telah digunakan sejak tahun 1972. Satelit Landsat 8 saat ini memiliki resolusi maksimum 15-meter per pixel untuk beberapa *band* yang disediakan, sedangkan untuk waktu pemanfaatan arsip citra Landsat 8 dapat diakses dari April 2013 dan tersedia 16 hari sekali untuk satu citra. Landsat 8 memiliki beberapa band yang dijelaskan pada tabel 1.

Tabel 1. Band Landsat 8 beserta kegunaannya.

Band	Panjang Gelombang	Kegunaan
Band 1 – <i>coastal aerosol</i>	0.43 – 0.45	Studi pesisir dan <i>aerosol</i>
Band 2 – Biru	0.45 – 0.51	Pemetaan batimetri, membedakan tanah dari vegetasi, dan deciduous dari vegetasi konifera .
Band 3 – Hijau	0.53 – 0.59	Penelitian vegetasi, yang berguna untuk menilai kekuatan tanaman.
Band 4 – Merah	0.64 – 0.67	Pembedaan lereng vegetasi.
Band 5 – <i>Near Infrared (NIR)</i>	0.85 – 0.88	Menekankan konten biomassa dan garis pantai.
Band 6 – <i>Short-wave Infrared (SWIR 1)</i>	1.57 – 1.65	Membedakan kelembapan tanah dan tanaman, menembus awan tipis
Band 7 – <i>Short-wave Infrared (SWIR 1)</i>	2.11 – 2.29	Meningkatkan kelembapan tanah dan tanaman, menembus awan tipis.
Band 8 – <i>Panchromatic</i>	0.50 – 0.68	Resolusi 15 meter, menajamkan definisi gambar.
Band 9 – <i>Cirrus</i>	1.36 – 1.38	Peningkatan deteksi kontaminasi awan cirrus
Band 10 – TIRS 1	10.60 – 11.19	Resolusi 100 meter, pemetaan termal dan perkiraan kelembapan tanah.
Band 11 – TIRS 2	11.5 – 12.51	Resolusi 100 meter, meningkatkan kualitas pemetaan termal dan perkiraan kelembapan tanah.

Sumber: *U.S. Geological Survey*

METODE

Peubah Penelitian

Peubah penelitian berasal dari KSA sebagai peubah respon dan citra satelit Landsat 8 sebagai peubah prediktornya. Data KSA yang diambil adalah koordinat sampel terpilih dan label KSA, sedangkan untuk citra satelit Landsat 8 diambil band 1 sampai dengan band 11 yang kemudian akan disusun menjadi indeks spektral yaitu *Enchanted Vegetation Index (EVI)*, *Normalize Difference Vegetation Index (NDVI)*, *Modified Soil Adjusted Vegetation Index (MSAVI)*(Qi et al. 1995), *Modified Normalize Difference Water Index (MNDWI)* (Xu , 2005) dan *Normalize Difference Building Index (NDBI)*

Sumber Data

Data pada penelitian ini diambil dari citra satelit Landsat 8 bulan Januari – Desember 2018 yang diolah melalui Google Earth Engine sehingga menghasilkan data mentah dalam format *comma separated value* berisi tanggal pemotretan citra, koordinat, dan indeks vegetatif. Untuk data respon diperoleh dari survey KSA Badan Pusat Statistik selama tahun 2018. Data label ini berisi peubah respon dari pengamatan titik koordinat yang menjadi sampel KSA 2018. Data selanjutnya berasal dari Badan Informasi Geospasial (BIG) berupa shapefile batas wilayah administratif kabupaten Brebes yang nantinya akan digunakan untuk mengekstrak koordinat-koordinat didalam kabupaten Brebes.

Persiapan Data

Tahapan pertama pada pembentukan penelitian ini adalah ekstrasi data citra satelit Landsat 8. Data-data spektral dari Landsat 8 berupa band-band akan digunakan sebagai peubah bentuan

untuk membuat indeks baru yaitu indeks spektral yang digunakan pada penelitian ini, yaitu : EVI, MSAVI, MNDWI dan NDBI. Data-data ini diperoleh dari google earth engine, dimana data yang diambil adalah data yang memiliki label KSA pada koordinat tertentu di kabupaten Brebes dan data pada waktu 3 periode sebelumnya.

Pra Pengolahan

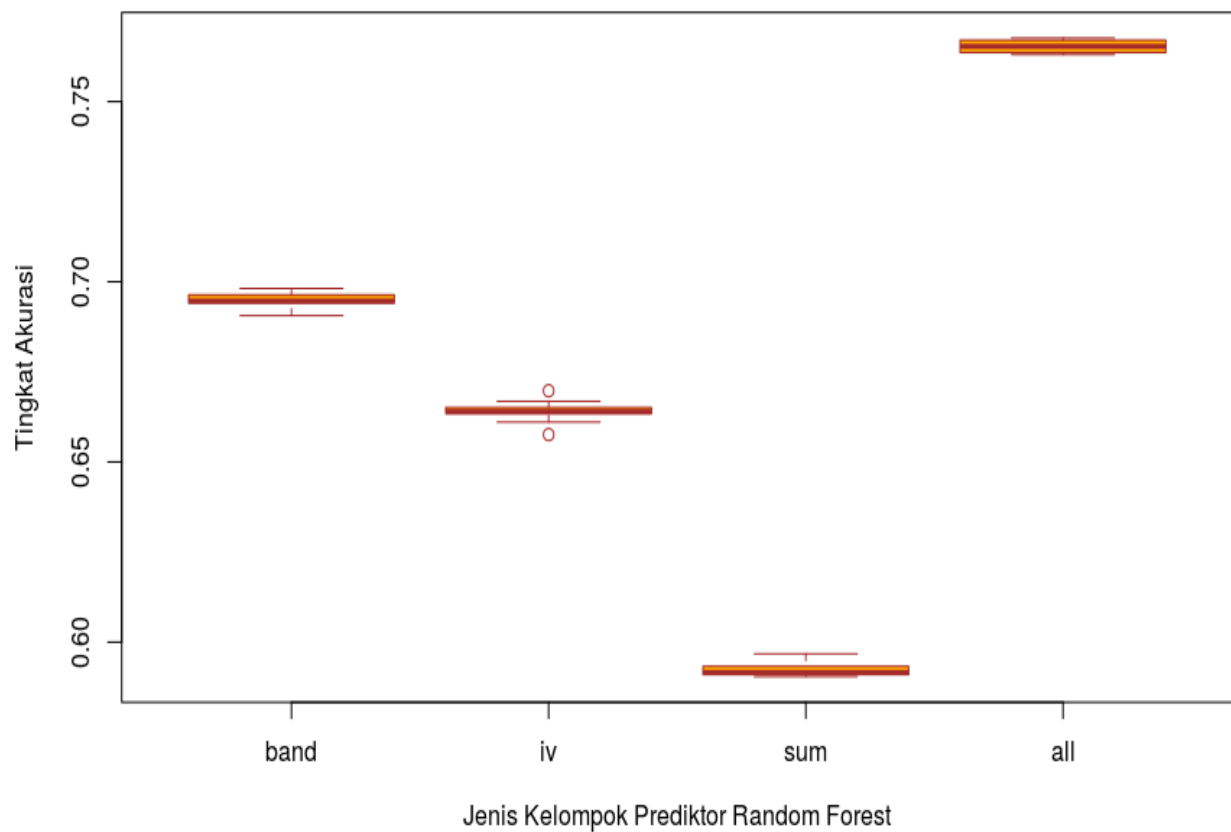
Pada tahap ini akan dilakukan proses pembentukan format data dan imputasi data. Imputasi dilakukan untuk mengantisipasi data-data yang bersesuaian dengan data KSA namun memiliki indeks spektral *Not Available/NA*, hal ini disebabkan oleh tutupan awan yang terjadi pada area penelitian. Selanjutnya data yang memiliki nilai NA akan diimputasi menjadi suatu nilai yang sangat besar. Proses imputasi dilakukan agar data dapat diolah dengan metode ensemble namun masih memiliki history/flag tentang ketidaktersediaan data tersebut.

Pembentukan model & Proses estimasi

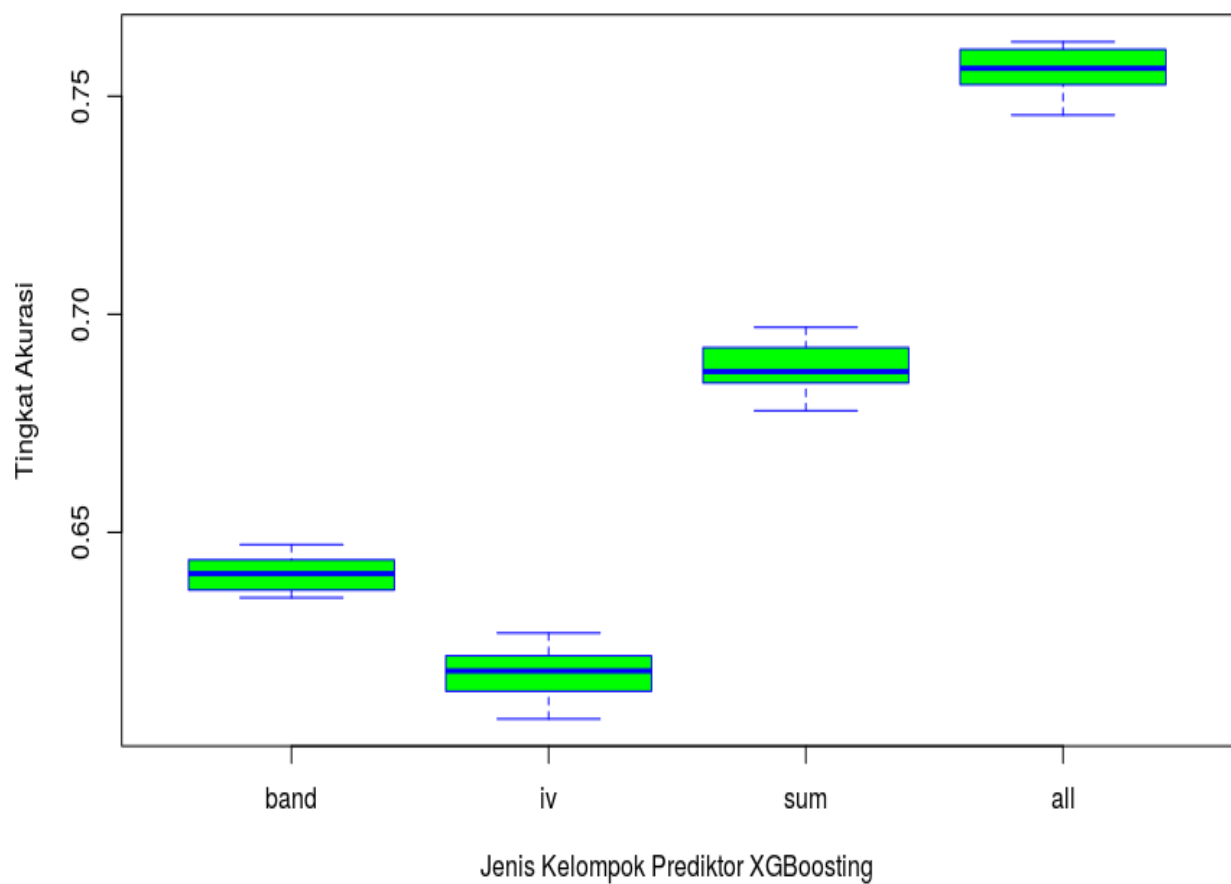
Pembentukan model klasifikasi dengan menggunakan Algoritma *Ensemble* yaitu boosting dan random forest. Selama pembentukan model, dilakukan pengelompokan *feature* yang digunakan yaitu band saja, indeks vegetatif saja, statistik dari band dan indeks vegetatif dan gabungan dari band, indeks vegetatif, dan statistiknya. Model yang terbentuk akan dilihat ukuran keakuratannya untuk dipilih model terbaik yang akan digunakan dalam melakukan klasifikasi lahan pertanian. Keakuratan menjadi penting namun kestabilan model juga tidak kalah penting, oleh karena itu agar dapat menjaga kestabilan model, dilakukan pengulangan selama 10 kali pengulangan agar dapat dicek kestabilannya melalui statistik rata-rata dan simpangan bakunya.

HASIL DAN PEMBAHASAN

Pengklasifikasian dengan metode *Random Forest* dan *Boosting* dilakukan dengan 4 tahapan yaitu dengan menggunakan *Band* saja sebagai peubah prediktor, Indeks Vegetatif sebagai prediktor, Statistik dari Band dan Indeks Vegetatif, serta Band, Indeks Vegetatif dan Statistiknya atau dengan kata lain keseluruhan peubah digunakan. Hal ini dapat menggambarkan tingkat kepentingan peubah respon serta dan keunggulan dari metode yang akan digunakan. selain membagi kelompok prediktor, dilakukan pengulangan sebanyak 10 kali kemudian hasilnya dieksplorasi menggunakan boxplot.



Gambar 1. Boxplot jenis kelompok prediktor dan tingkat akurasi dari Metode Random Forest.



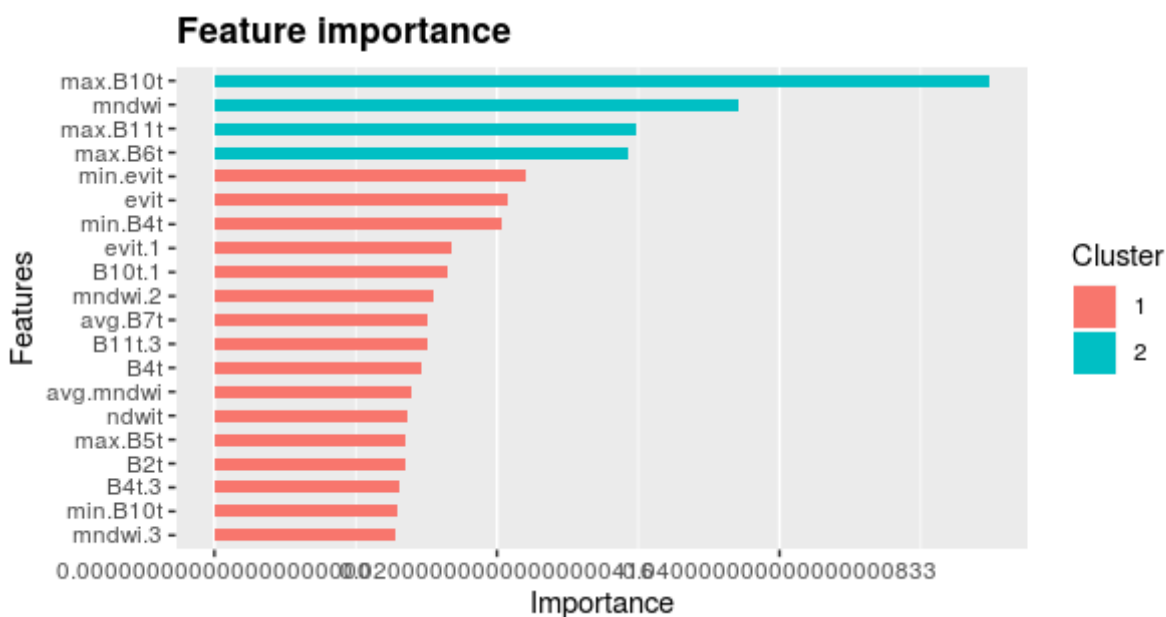
Gambar 2. Boxplot jenis kelompok prediktor dan tingkat akurasi dari XGBoost.

Tabel 2. Rataan dan Simpangan Baku Akurasi dari *Random Forest* dan *Boosting* menurut kelompok *feature*.

Kelompok Peubah Prediktor	Rataan		Simpangan Baku	
	<i>Random Forest</i>	<i>Boosting</i>	<i>Random Forest</i>	<i>Boosting</i>
Band	67,47	64,02	0.00217	0.00389
Indeks Vegetatif	66,40	61,77	0.00324	0.00602
Statistik Band dan Indeks Vegetatif	59,23	68,77	0.00197	0.00563
Semua Prediktor	76,52	75,60	0.00167	0.00523

dari hasil eksplorasi boxplot dan tabel diatas terlihat Random forest memiliki akurasi yang lebih stabil daripada Xgboost. Hal ini ditandai dengan kecilnya angka simpangan baku dari akurasi tiap kelompok peubah prediktor dan juga ditandai dengan boxplot yang melebar. Pada Xgboost boxplot lebih memanjang dan simpangan baku akurasi juga lebih besar, namun dari kedua metode tersebut tidak ada yang simpangan baku akurasinya menyentuh angka satu persen, hal ini lebih meyakinkan bahwa model ensemble stabil digunakan dalam mengklasifikasikan jenis lahan padi.

Selanjutnya apabila dilihat dari tingkat akurasi yang digambarkan oleh rata-rata akurasi dari Random Forest dan XGBoost sama sama menunjukkan hal yang sama yaitu penggunaan seluruh peubah semakin meningkatkan akurasi dari model.



Gambar 3. 20 Feature Importance paling tinggi dari XGBoost.

Peubah prediktor yang memiliki peranan terpenting dalam model yang dibangun oleh Xgboost adalah statistik maksimum dari band 10 periode yang sama, bulan citra satelit diambil, NDVI, statistik maksimum dari band 6 periode yang sama, dan statistik maksimum dari band 11 periode yang sama. Peubah bulan pengambilan citra sateit menjadi *feature* yang penting selain band satelit karena dalam pertanian ada periode sirkular yang memiliki pola sehingga model dapat dengan baik menangkap pola tersebut.

KESIMPULAN

Metode ensemble dapat dengan baik mengklasifikasikan jenis lahan padi, baik dari sisi akurasi maupun dari kestabilan model. Akurasi dari Random Forest adalah yang tertinggi yaitu 76,52 dan juga paling stabil dengan simpangan baku yang kecil. Sedangkan baik band maupun indeks vegetatif juga memiliki peran yang besar ditandai dengan band dan indeks vegetatif yang memiliki *importance feature* yang tinggi. Kesalahan akurasi lebih banyak ditemukan karena pra proses yang masih belum

banyak dilakukan, seperti masalah *imbalance class*, teknik imputasi yang tepat tanpa menimbulkan masalah baru yaitu *overfitting model*. Kedepannya dapat juga dilakukan eksplorasi lebih jauh mengenai citra satelit yang lain seperti modis dan satelit cuaca untuk menangkap *feature* yang penting lainnya.

DAFTAR PUSTAKA

- Badan Pusat Statistik. (2018). Statistik Indonesia 2018. Badan Pusat Statistik, Jakarta.
- Badan Pusat Statistik. (2018). Upaya Perbaikan Data Padi Dengan Metode Kerangka Sampel Area (KSA) 2018. Badan Pusat Statistik, Jakarta.
- Badan Pusat Statistik Provinsi Jawa Tengah. (2018). Berita Resmi Statistik Luas Panen dan Produksi Padi di Jawa Tengah 2018 (Perbaikan Metodologi Perhitungan Data Produksi Beras dengan Metode Kerangka Sampel Area). Badan Pusat Statistik Provinsi Jawa Tengah, Semarang.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.
- Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, Berlin, Heidelberg. P 1-15
- Dirgahayu, D., Noviar, H., Anwar, S. (2015). Pengembangan Model Pertumbuhan Tanaman Padi Menggunakan Data Evi Modis Multitemporal (Studi Kasus Di Pulau Sulawesi). *Seminar Nasional Penginderaan Jauh 2015*.
- Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *icml* (Vol. 96, pp. 148-156).
- LAPAN. (2015). Pedoman Pemantauan Fase Pertumbuhan Tanaman Padi Menggunakan Data Satelit Penginderaan Jauh. LAPAN, Jakarta.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote sensing of environment*, 48(2), 119-126.
- Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, 185, 46-56.
- Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14), 3025-3033.
- Zha, Y., Ni, S. X., & Yang, S. (2003). An effective approach to automatically extract urban land-use from TM imagery. *JOURNAL OF REMOTE SENSING-BEIJING-*, 7(1), 37-40.