MODEL KLASIFIKASI KEUNGGULAN KEILMUAN PUBLIKASI DOSEN PTN-BH MENGGUNAKAN ALGORITMA *DECISION TREE C5.0* DAN *RANDOM FOREST*

ERLENIH METASARI

PROGRAM MAGISTER ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul "Model Klasifikasi Keunggulan Keilmuan Publikasi Dosen PTN-BH Menggunakan Algoritma *Decision Tree C5.0* dan *Random Forest*" adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2026

Erlenih Metasari
M0503241011

RINGKASAN

ERLENIH METASARI. Model Klasifikasi Keunggulan Keilmuan Publikasi Dosen PTN-BH Menggunakan Algoritma *Decision Tree C5.0* dan *Random Forest*. Dibimbing oleh HARI AGUNG ADRIANTO, SONY HARTONO WIJAYA, SRI SUNING KUSUMAWARDANI.

Peningkatan daya saing ilmu pengetahuan dan teknologi (IPTEK) nasional sangat bergantung pada kualitas publikasi ilmiah. Perguruan Tinggi Negeri Badan Hukum (PTN-BH) memiliki mandat strategis dalam pengembangan *Center of Excellence* yang diakui secara nasional maupun internasional, sebagaimana diamanatkan oleh Panduan Pusat Unggulan IPTEK Perguruan Tinggi serta Undang-Undang Nomor 11 Tahun 2019 tentang Sistem Nasional IPTEK.

Namun, capaian publikasi antar PTN-BH belum seragam dari segi kuantitas, kualitas, dan arah keilmuan. Kondisi tersebut memerlukan pendekatan berbasis data untuk mengidentifikasi konsentrasi bidang keilmuan secara sistematis. Penelitian ini bertujuan mengevaluasi pola publikasi dosen PTN-BH, mengidentifikasi fitur yang berkontribusi terhadap klasifikasi, serta membandingkan kinerja *Decision Tree* C5.0 dan *Random Forest* dalam mengklasifikasikan keunggulan keilmuan publikasi dosen PTN-BH.

Penelitian menggunakan data sekunder profil dosen dan metadata publikasi periode 2021–2025 yang bersumber dari Sistem Informasi Sumberdaya Terintegrasi (SISTER), Pangkalan Data Pendidikan Tinggi (PDDikti), dan *Science and Technology Index* (SINTA). Data publikasi awal berjumlah 1.571.410 baris. Melalui seleksi domain, integrasi data, agregasi publikasi berdasarkan unit dosen PTN-BH, dan pembersihan data, diperoleh 27.532 observasi final. Judul publikasi setiap dosen digabungkan menjadi dokumen teks, dinormalisasi, kemudian direpresentasikan menggunakan *Term Frequency–Inverse Document Frequency* dengan kombinasi unigram dan bigram serta maksimum 1.000 fitur.

Pembentukan label bidang keilmuan dilakukan dengan membandingkan *Latent Dirichlet Allocation* (LDA), K-Means, dan *Agglomerative Hierarchical Clustering* pada tiga kelompok. Evaluasi menggunakan *Silhouette Score* dan *Davies–Bouldin Index* (DBI) menunjukkan bahwa LDA memberikan hasil terbaik dengan *Silhouette Score* 0,6983 dan DBI 0,3835. Setiap observasi kemudian diberi label berdasarkan topik LDA dengan probabilitas tertinggi sehingga terbentuk tiga kategori bidang keilmuan yang digunakan sebagai target klasifikasi.

Hasil pengujian menunjukkan bahwa *Random Forest* memiliki kinerja klasifikasi lebih baik dibandingkan *Decision Tree C5.0. Random Forest* menghasilkan akurasi 90% dan (ROC-AUC) 0,9824, sedangkan *Decision Tree C5.0* menghasilkan akurasi 82% dan ROC-AUC 0,9232. Distribusi ketiga bidang pada 23 PTN-BH menunjukkan perbedaan konsentrasi publikasi antar institusi. Keunggulan keilmuan didefinisikan sebagai bidang dengan konsentrasi publikasi relatif dominan berdasarkan pola semantik publikasi dosen. Hasil penelitian memberikan pemetaan empiris portofolio bidang keilmuan PTN-BH sebagai salah satu masukan berbasis data dalam evaluasi arah riset dan diferensiasi misi institusi.

Kata kunci: *decision tree c5.0,* keunggulan keilmuan, *machine learning*, publikasi dosen PTN-BH, *random forest.*

SUMMARY

ERLENIH METASARI. Classification Model of Lecturers' Publication-Based Scientific Excellence in Academic State Universities with Legal Entity State University (PTN-BH) Using Decision Tree C5.0 and Random Forest. Supervised by HARI AGUNG ADRIANTO of 1st SUPERVISOR 1st, SONY HARTONO WIJAYA of 2nd SUPERVISOR, and SRI SUNING KUSUMAWARDANI of 3rd SUPERVISOR.

The Enhancing national competitiveness in science and technology (IPTEK) strongly depends on the quality of scientific publications. Legal Entity State Universities (PTN-BH) have a strategic mandate to develop *Centers of Excellence* recognized nationally and internationally, as stipulated in the Guidelines for University-Based Centers of Excellence in Science and Technology and Law Number 11 of 2019 concerning the National System of Science and Technology.

However, publication achievements across PTN-BH institutions vary in quantity, quality, and scientific orientation. This condition requires a data-driven approach to identify concentrations of scientific fields. This study aims to evaluate publication patterns among PTN-BH lecturers, identify features that contribute to classification, and compare the performance of Decision Tree C5.0 and Random Forest in classifying scientific excellence based on PTN-BH lecturers' publications.

This study used secondary data consisting of lecturer profiles and publication metadata for the 2021–2025 period obtained from the Integrated Academic Resource Information System (SISTER), the Higher Education Database (PDDikti), and the Science and Technology Index (SINTA). The publication dataset consisted of 1,571,410 records. Through domain selection, data integration, aggregation at the lecturer PTN-BH level, and data cleaning, 27,532 observations were obtained. Publication titles for each lecturer were concatenated into text documents, normalized, and represented using Term Frequency–Inverse Document Frequency with unigram and bigram combinations and a maximum of 1,000 features.

Scientific-field labels were generated by comparing Latent Dirichlet Allocation (LDA), K-Means, and Agglomerative Hierarchical Clustering using a three-group configuration. Evaluation using the Silhouette Score and Davies–Bouldin Index (DBI) showed that LDA achieved the best results, with a Silhouette Score of 0.6983 and a DBI of 0.3835. Each observation was assigned a label based on the LDA topic with the highest probability, resulting in three scientific-field categories used as classification targets.

The results showed that Random Forest outperformed Decision Tree C5.0, achieving 90% accuracy and a ROC-AUC of 0.9824, compared with 82% and 0.9232, respectively. The three scientific fields showed different publication concentrations across 23 PTN-BH institutions. Scientific excellence refers to the dominant field based on publication semantic patterns. The findings provide a data-driven mapping of PTN-BH scientific-field portfolios to support research direction evaluation and institutional mission differentiation.

Keywords: decision tree c5.0, machine learning, PTN-BH lecturer publications, random forest, scientific excellence.

MODEL KLASIFIKASI KEUNGGULAN KEILMUAN PUBLIKASI DOSEN PTN-BH MENGGUNAKAN ALGORITMA *DECISION TREE C5.0* DAN *RANDOM FOREST*

ERLENIH METASARI

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Program Magister pada
Program Studi Ilmu Komputer

PROGRAM MAGISTER ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

Tim Penguji pada Ujian Tesis:

Dr. Toto Haryanto, S.Kom.,M.Si

Judul Tesis : Model Klasifikasi Keunggulan Keilmuan Publikasi Dosen PTN-BH Menggunakan Algoritma *Decision Tree* C5.0 dan *Random Forest*
Nama : Erlenih Metasari
NIM : M0503241011

Disetujui oleh

Pembimbing 1:
Hari Agung Adrianto, S.Kom., M.Si., Ph.D.

Pembimbing 2:
Dr. Sony Hartono Wijaya, S.Kom., M.Kom.

Pembimbing 3:
Prof. Dr. Ir. Sri Suning Kusumawardani, S.T., M.T.

Diketahui oleh

Ketua Program Studi:
Dr. Toto Haryanto, S.Kom., M.Si.
NIP 198211172014041001

Dekan Sekolah Sains Data, Matematika, dan Informatika:
Prof. Dr. Ir. Agus Buono, M.Si., M.Kom.
NIP 19660702 199302 1 001

Tanggal Ujian:
31 Juli 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur tak terhingga penulis panjatkan kepada Allah Subḥānahu wa Taʿālā atas segala karunia, kasih sayang, pertolongan, dan hidayah-Nya sehingga karya ilmiah ini berhasil diselesaikan. Sungguh, semua karena Allah, dan hanya Allah tempat terbaik dalam meminta dan memohon pertolongan. Tema penelitian yang dilaksanakan sejak bulan Oktober 2024 sampai bulan Oktober 2025 ini berkaitan dengan bioinformatika, dengan judul "Model Klasifikasi Keunggulan Keilmuan Publikasi Dosen PTN-BH Menggunakan Algoritma *Decision Tree C5.0* dan *Random Forest*".

Ucapan terima kasih secara khusus penulis haturkan kepada:

1. Bapak Hari Agung Adrianto, S.Kom., M.Si., Ph.D., Dr. Sony Hartono Wijaya, S.Kom., M.Kom. dan Ibu Prof. Dr. Ir. Sri Suning Kusumawardani, S.T., M.T. selaku komisi pembimbing yang telah membimbing, mengarahkan, memberi nasehat serta memberikan tunjuk ajar kepada penulis.
2. Beasiswa Unggulan (BU) Pegawai dari Pusat Pembiayaan Pendidikan di bawah Kementerian Pendidikan Dasar dan Menengah Republik Indonesia yang telah memberikan pembiayaan pendidikan sehingga penulis dapat melanjutkan studi dengan lancar.
3. Kepada Ayah Resim (Alm), Ibu Sarkem (Almh), Keluarga Serumah (Teteh Endang, Mas Agung, Ara, Jarier, Kiya) dan Saudara Kakak Adik (Teteh Rasitem, Teteh Endang, Teteh Ecih dan Azun) serta seluruh keluarga besar yang senantiasa selalu mendoakan, memberikan dukungan serta kasih sayang kepada Penulis.
4. Sahabat seperjuangan Penulis, teman-teman Ilmu Komputer angkatan 2024 dan 2023, teman-teman sobat Sarana dan Prasarana terutama (Mba Windi, Mas Bagus, dan Ibu Rini) serta Mas Tito Tim SISTER, Direktorat Sumber Daya Ditjen Diktiristek, Kemendiktisaintek,
5. Sahabat penulis sebagai *support system* khususnya Yulia dan Saripudin serta pihak dan rekan lainnya yang tidak dapat penulis tuliskan satu per satu yang telah memberikan semangat, do'a, dan dukungan selama menempuh Pendidikan Magister di Bogor. Semoga kelak kita semua akan berjumpa kembali.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Agustus 2026

Erlenih Metasari

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISTILAH

Istilah	Definisi
Accuracy	Tingkat ketepatan model dalam mengklasifikasikan data dengan benar terhadap seluruh data pengujian.
Agglomerative Hierarchical Clustering	Metode pengelompokan hierarkis yang membentuk cluster secara bertahap dengan menggabungkan objek yang memiliki kemiripan tertinggi hingga jumlah cluster yang ditentukan tercapai.
Bagging (Bootstrap Aggregating)	Teknik ensemble yang membangun banyak model dari sampel bootstrap untuk meningkatkan stabilitas dan akurasi prediksi.
Bibliometrik	Pendekatan kuantitatif untuk menganalisis karakteristik publikasi ilmiah, seperti produktivitas, sitasi, kolaborasi, dan perkembangan bidang ilmu.
Cluster	Sekelompok objek yang memiliki karakteristik atau tingkat kemiripan yang tinggi dibandingkan objek pada cluster lain.
Clustering	Teknik unsupervised learning yang bertujuan mengelompokkan objek berdasarkan tingkat kemiripan karakteristiknya.
Confusion Matrix	Tabel evaluasi yang menunjukkan jumlah prediksi benar dan salah untuk setiap kelas hasil klasifikasi.
Corpus	Kumpulan dokumen teks yang menjadi sumber data dalam proses text mining.
Data Mining	Proses menemukan pola, hubungan, atau pengetahuan baru dari kumpulan data menggunakan teknik statistik dan machine learning.
Decision Tree	Algoritma klasifikasi berbentuk struktur pohon yang membagi data berdasarkan atribut terbaik hingga menghasilkan keputusan akhir.
Domain Restriction	Proses membatasi ruang lingkup analisis hanya pada dokumen atau data yang relevan dengan tujuan penelitian.
Feature	Karakteristik atau atribut yang digunakan sebagai masukan (input) dalam proses machine learning.
Feature Engineering	Tahapan membangun, memilih, atau mentransformasi fitur agar lebih representatif terhadap proses pembelajaran model.
Feature Extraction	Proses mengubah data mentah menjadi representasi numerik yang dapat diproses oleh algoritma machine learning.
F1-Score	Nilai harmonik antara precision dan recall yang digunakan untuk mengevaluasi keseimbangan performa model klasifikasi.
Hyperplane	Batas pemisah antar kelas dalam ruang multidimensi (istilah umum dalam machine learning).

K-Means	Algoritma clustering berbasis centroid yang mengelompokkan data berdasarkan jarak ke pusat cluster.
Keunggulan Keilmuan	Kondisi yang menunjukkan tingkat kekuatan atau dominasi suatu bidang ilmu berdasarkan karakteristik publikasi ilmiah dosen.
Label Keunggulan	Kategori hasil pengelompokan dosen ke dalam kelompok kategori keunggulan (klister 0, 1 dan 2) dan digunakan sebagai variabel target pada tahap klasifikasi.
Machine Learning	Cabang kecerdasan buatan yang memungkinkan sistem mempelajari pola dari data untuk melakukan prediksi atau klasifikasi secara otomatis.
Model Klasifikasi	Model pembelajaran terawasi yang digunakan untuk memetakan suatu observasi ke dalam salah satu kelas kategorikal berdasarkan pola yang dipelajari dari data pelatihan.
N-Gram	Kombinasi beberapa kata berurutan yang digunakan sebagai fitur dalam analisis teks.
Normalisasi Teks	Tahapan prapemrosesan yang bertujuan menyeragamkan format teks agar mudah dianalisis.
Precision	Proporsi prediksi positif yang benar dibandingkan seluruh prediksi positif yang dihasilkan model.
Preprocessing	Tahapan awal pengolahan data yang meliputi pembersihan, normalisasi, tokenisasi, dan transformasi data sebelum proses analisis.
PTN-BH	Perguruan Tinggi Negeri Badan Hukum yang memiliki otonomi dalam pengelolaan akademik maupun non-akademik sesuai ketentuan peraturan perundang-undangan.
Random Forest	Algoritma ensemble yang membangun banyak decision tree dan menggabungkan hasil prediksinya melalui mekanisme voting mayoritas.
Recall	Kemampuan model menemukan seluruh data yang benar-benar termasuk dalam suatu kelas tertentu.
Silhouette Index	Ukuran kualitas clustering yang menunjukkan tingkat kedekatan objek terhadap cluster sendiri dibandingkan cluster lain. Nilai berkisar antara -1 hingga 1.
Davies–Bouldin Index (DBI)	Indeks evaluasi clustering yang mengukur rasio kedekatan antar cluster. Semakin kecil nilai DBI, semakin baik kualitas cluster.
Stopword	Kata umum yang memiliki sedikit makna informatif sehingga dihilangkan dalam proses text mining.
Stemming	Proses mengubah kata menjadi bentuk dasar dengan menghilangkan imbuhan.
Text Mining	Proses ekstraksi informasi dan pengetahuan dari data teks menggunakan teknik komputasi dan analisis bahasa alami.
Title Chars	Jumlah karakter dari seluruh judul publikasi yang digabungkan untuk setiap dosen.

TF-IDF	Metode pembobotan kata yang mengukur tingkat kepentingan suatu kata dalam dokumen berdasarkan frekuensi kemunculan pada dokumen dan keseluruhan korpus.
Tokenisasi	Proses memecah teks menjadi unit-unit kecil berupa kata atau token sebagai dasar analisis teks.
Unsupervised Learning	Metode machine learning yang mempelajari pola data tanpa menggunakan label kelas sebagai target pembelajaran.
Vector Space Model	Representasi dokumen dalam bentuk vektor numerik sehingga dapat diproses menggunakan algoritma machine learning.
Word Cloud	Visualisasi kata berdasarkan frekuensi kemunculannya dalam kumpulan dokumen.