

## ABSTRAK

Perguruan Tinggi Negeri memiliki perpustakaan sebagai fasilitas penunjang pendidikan dan ilmu pengetahuan para mahasiswa. Di UPN “Veteran” Yogyakarta khususnya Fakultas Teknik Industri (FTI) memiliki perpustakaan yang didalamnya menyimpan berbagai buku, jurnal maupun tugas akhir. Perpustakaan FTI sendiri menyimpan jurnal, tugas akhir/skripsi dari berbagai jurusan yang ada di Fakultas Teknik Industri salah satunya dari jurusan Teknik Informatika. Perpustakaan FTI sendiri menyimpan jurnal, tugas akhir/skripsi dari berbagai jurusan yang ada di Fakultas Teknik Industri salah satunya dari jurusan Teknik Informatika. Pada pembaca/mahasiswa sendiri untuk mencari referensi tugas akhir mahasiswa datang ke perpustakaan, membaca daftar skripsi yang ada, memilih judul yang diinginkan kemudian akan dicarikan laporan tugas akhir berdasarkan judul tersebut. Di perpustakaan FTI sendiri tugas akhir jurusan Teknik Informatika belum dikelompokkan sesuai bidang kajian, dimana didalam jurusan Teknik Informatika memiliki 4 bidang kajian sesuai KBK (Kurikulum Berbasis Kompetensi) antara lain Geoinformatika, Komputasi dan Jaringan, Sistem Cerdas serta Multimedia.

Pada penelitian ini menggunakan metode pengklasifikasian *K-Nearest Neighbor* (K-NN) untuk mengkategorikan judul skripsi jurusan Teknik Informatika berdasarkan bidang kajian. Kemudian menggunakan fitur seleksi *information gain* untuk menangani data *imbalance* dimana data yang digunakan komposisi datanya disetiap bidang kajian berbeda atau tidak seimbang. Proses pengklasifikasian melalui beberapa tahap antara lain *preprocessing*, pembobotan dengan TF-IDF, fitur seleksi *information gain*, perhitungan jarak terdekat dengan *cosine similarity* lalu klasifikasi dengan K-NN.

Berdasarkan hasil pengujian yang telah dilakukan hasil akurasi tertinggi apabila menggunakan fitur seleksi *information gain* dengan data latih sebanyak 276 data ada pada persentase data latih (80%) dan data uji (20%) sebesar 87,5% dan dengan parameter nilai  $K=5$ . Hasil akurasi tertinggi apabila tidak menggunakan fitur seleksi *information gain* dengan data latih sebanyak 276 data ada pada persentase data latih (90%) dan data uji (10%) sebesar 92,9% dan dengan parameter nilai  $K=5, 7$  dan  $9$ . Sehingga untuk penggunaan fitur seleksi *information gain* kurang maksimal untuk data berupa judul tugas akhir dikarenakan fitur-fitur yang ada pada judul tugas akhir termasuk fitur penting semua.

**Kata kunci** : klasifikasi dokumen, judul skripsi, bidang kajian, fitur seleksi, *information gain*, *K-Nearest Neighbor* (K-NN)