BAB I
PENDAHULUAN

1.1. Latar Belakang

Pemanfaatan data pada era digital saat ini sangatlah penting, dikarenakan dari hasil pendataan tersebut dapat menentukan, maupun memberikan gambaran serta langkah apa yang harus dilakukan oleh pemegang data tersebut. Hampir 80% data bersifat tekstual dan tidak terstruktur, akan tetapi dianggap sebagai sumber informasi yang sangat berharga (El Rifai et al., 2022). Sebagaimana halnya informasi gender, yang tidak lagi menjadi input yang diperlukan saat mendaftar dalam pembuatan akun, keikut sertaan pelatihan, maupun keikut sertaan dalam kegiatan lainnya (Hu et al., 2021). Sejatinya data gender sangat bermanfaat dalam beberapa hal, seperti dalam kegiatan industri, yang mana data gender berperan untuk memperoleh informasi tambahan mengenai pelanggan, dan dapat digunakan dalam pemasaran yang ditargetkan ataupun pengiklanan (Jia & Zhao, 2019). Serta dalam keikutsertaan program pelatihan, data gender berperan untuk memperoleh informasi jumlah atau seberapa banyak peserta yang mengikuti program pelatihan berdasarkan jenis kelaminnya, yang nantinya data tersebut dapat membantu dalam membuat program pelatihan baru yang terkhususkan.

Pemegang data memiliki dokumen terkait data personal tetapi tidak dengan variabel gender (Jia & Zhao, 2019). Dalam hal ini, memungkinkan pemilik data mengklasifikasikan variabel gender berdasarkan nama yang ada secara manual, akan tetapi mengklasifikasikan dokumen secara manual oleh para ahli dinilai tidak seefisien seperti dulu, dikarenakan jumlah data yang ada semakin meningkat (El Rifai et al., 2022).

Dari permasalahan tersebut disinilah algoritma *supervised learning* pada *machine learning* berperan, sebagai cara alternatif dalam pengklasifikasian dokumen (El Rifai et al., 2022). Pengklasifikasian yang dimaksud dilakukan dengan pendekatan *machine learning* berbasis teks untuk menyimpulkan jenis kelamin berdasarkan namanya dengan membandingkan terhadap daftar nama yang sudah terkait dengan jenis kelamin (Rego et al., 2021; Septiandri, 2017).

Dalam hal ini dilakukan penelitian terkait identifikasi jenis kelamin berdasarkan nama menggunakan metode fitur ekstraksi dan penerapan algoritma klasifikasi teks pada *machine learning*, dikarenakan peneliti sudah melakukan eksperimen pada metode *deep learning* dan mengalami *hang (*perangkat tidak beroperasi*)*, dikarenakan perangkat yang digunaka pada penelitian ini tidak cukup memumpuni. Dalam penelitian ini berfokus pada nama-nama orang indonesia dan penggunaan *bag-of-word* serta pemanfaatan *n-gram* sebagai fitur ekstarksi. Algoritma klasifikasi yang digunakan yaitu *logistic regression* (LR) dan *multinomial naive bayes* (MNB) yang keduanya akan dibandingkan tingkat akurasi, presisis, recall, dan f1-score menggunakan *confusion matrix*, yang semua metode di atas diakses dalam pustaka *python* yaitu *sklearn.* Model yang terbaik akan diimplementasikan ke dalam aplikasi identifikasi jenis kelamin berbasis web dengan memanfaatkan pustaka *python* yaitu *streamlit*.

1.2. Perumusan Masalah

Berdasarkan latar belakang yang sudah dijabarkan di atas, pada penelitian ini terdapat beberapa rumusan masalah sebagai berikut:

1. Bagaimana kemampuan model *Logistic Regression* dan *Multinomial Naive Bayes* dalam memprediksi jenis kelamin berdasarkan nama?
2. Bagaimana mendapatkan model yang optimal untuk memprediksi jenis kelamin berdasarkan data keikutsertaan kepelatihan?
3. Belum adanya pengimplementasikan dalam mengidentifikasi jenis kelamin berbasis web?

1.3. Tujuan Penelitian

Adapun tujuan dari dilakukannya penelitian ini yaitu:

1. Mengukur kemampuan model supervised learning untuk memprediksi jenis kelamin berdasarkan nama
2. Mendapatkan model yang optimal untuk memprediksi jenis kelamin berdasarkan data keikutsertaan kepelatihan

3. Mempermudah melakukan prediksi menggunakan algoritma terbaik dengan aplikasi berbasis web

1.4. Manfaat Penelitian

Dalam penelitian ini diharapkan dapat bermanfaat bagi seluruh pihak yang terkait, diantaranya:

1. Secara Aplikatif
   a. Diharapkan dari aplikasi identifikasi jenis kelamin berdasarkan nama ini dapat bermanfaat dalam memperkirakan atau kemungkinan jumlah laki-laki atau perempuan dalam suatu dokumen.
   b. Memberikan output variabel (kolom) jenis kelamin pada dokumen dan dapat diunduh dengan format file xlsx.
2. Secara Akademis
   a. Memberikan wawasan terkait penggunaan *machine learning* dalam mengidentifkasi jenis kelamin berdasrakan nama, yang ada pada dokumen.
   b. Memberikan cara atau tahapan pengimplementasian metode data mining dan algoritma *supervised learning* dalam mengidentifikasi jenis kelamin berdasarkan nama yang ada pada dokumen.

1.5. Batasan Masalah

Terdapat beberapa batasan masalah pada penelitian ini, agar dapat dilaksanakan secara spesifik. Batasan masalah tersebut meliputi sebagai berikut:

1. Dataset yang digunakan hanya nama-nama orang indonesia.
2. Penelitian berfokus pada kategori biner atau 2 jenis kelamin.
3. Evaluasi model berdasarkan confusion matrix.
4. Berfokus dalam pembandingan model, dan akan diimplementasikan sebagai sistem web app machine learning identifikasi jenis kelamin.
5. Tidak berfokus dalam pengembangan web.
6. Aplikasi berbasis web hanya menerima inputan berformat csv.