BAB I. PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang terus berkembang pesat, *email* telah menjadi salah satu sarana komunikasi utama yang digunakan, baik dalam kehidupan pribadi maupun dalam lingkungan profesional (Rahma *et al*., 2021). Seiring dengan meningkatnya penggunaan *email*, ancaman *spam email* juga semakin meningkat. *Spam email* adalah pesan yang tidak diinginkan dan sering kali bersifat mengganggu yang dikirim ke sejumlah besar penerima untuk tujuan komersial, penipuan, atau distribusi malware (Bachtiar *et al*., 2023). Masalah ini menimbulkan tantangan besar bagi penyedia layanan *email* dan pengguna, karena *spam* tidak hanya mengurangi produktivitas tetapi juga menimbulkan risiko keamanan.

Email merupakan sarana komunikasi yang memanfaatkan perangkat elektronik dan transmisi data melalui jaringan komputer untuk mengirimkan pesan kepada satu atau lebih penerima melalui internet (Ikbar Athallah Taufik, 2023). Perkembangan teknologi informasi yang sangat pesat menyebabkan penggunaan *email* sebagai sarana komunikasi menjadi sangat umum di berbagai lapisan masyarakat. Namun, bersamaan dengan popularitas email, media ini juga menjadi sasaran utama bagi serangan *spam* yang dapat mengganggu efisiensi dan keamanan komunikasi elektronik. Serangan *spam* tidak hanya mengakibatkan kerugian waktu, tetapi juga berpotensi membahayakan keamanan data pribadi maupun bisnis karena sering kali menyertakan tautan berbahaya atau lampiran berbahaya yang dapat merusak sistem.

Klasifikasi *spam email* merupakan proses mengidentifikasi dan memisahkan *email spam* dari *email* yang sah (*non-spam*) (Bachtiar *et al*., 2023). Sistem klasifikasi *spam* yang efektif dapat membantu mengurangi jumlah *spam* yang mencapai kotak masuk pengguna, sehingga meningkatkan efisiensi dan keamanan komunikasi melalui *email*. Banyak metode dan algoritma telah dikembangkan untuk mengatasi masalah ini, termasuk algoritma pembelajaran mesin yang memanfaatkan berbagai teknik klasifikasi.

Salah satu algoritma klasifikasi yang digunakan dalam pembelajaran mesin adalah *Naive Bayes***.** *Naive Bayes* merupakan metode klasifikasi probabilistik

sederhana yang bekerja dengan menghitung sejumlah probabilitas berdasarkan frekuensi serta kombinasi nilai dari data yang tersedia. Algoritma ini menggunakan teorema bayes dengan asumsi bahwa setiap fitur bersifat independen satu sama lain. Metode ini dikenal memiliki proses komputasi yang sederhana dan efisien, serta mampu menghasilkan kinerja klasifikasi yang baik pada berbagai jenis data, termasuk dalam klasifikasi teks seperti *Spam* email. Oleh karena itu, algoritma *Naive Bayes* banyak dimanfaatkan dalam penelitian yang berkaitan dengan klasifikasi data.

Berbagai studi sebelumnya turut memperkuat bahwa algoritma *Naive Bayes* memiliki performa yang baik dalam tugas klasifikasi. Penelitian oleh (Fitriyah *et al*., 2020) mengimplementasikan algoritma *Naive Bayes* untuk mengelompokkan email ke dalam kategori *spam* dan non-*spam*. Berdasarkan evaluasi menggunakan metode *k-fold cross validation*, diperoleh rata-rata akurasi sebesar 84,8%, dengan nilai *precision* 86% dan *recall* 85%. Hasil ini mengindikasikan bahwa *Naive Bayes* mampu memberikan kinerja yang cukup optimal dalam proses klasifikasi email.

Penelitian lain yang dilakukan oleh (Firmansyah *et al*., 2025) mengombinasikan algoritma *Naive Bayes* dengan teknik seleksi fitur Chi-Square dalam kerangka Knowledge Discovery in Databases (KDD). Dari hasil pengujian, model yang dihasilkan mencapai tingkat akurasi sebesar 81,00%, *precision* 100%, *recall* 65%, serta *F1-score* 79%. Temuan tersebut menunjukkan bahwa penerapan *Naive Bayes* yang dipadukan dengan seleksi fitur dapat meningkatkan efektivitas dalam pengolahan dan klasifikasi data.

Selain itu, studi oleh (Firdaus & Walid, 2022) mengembangkan *Naive Bayes* menjadi *Gaussian Naive Bayes* dan memperoleh hasil performa yang sangat tinggi. Model tersebut mencatat akurasi sebesar 0,9688, *precision* 0,97, *recall* 1,00, serta *F1-score* 0,98. Hasil ini menegaskan bahwa *Gaussian Naive Bayes* sangat baik digunakan untuk menangani data dengan karakteristik distribusi kontinu, sehingga mampu menghasilkan performa klasifikasi yang lebih unggul.

Meskipun berbagai penelitian terdahulu telah menunjukkan bahwa algoritma *Naive Bayes* mampu memberikan performa yang baik dalam klasifikasi *spam* email, sebagian besar penelitian masih berfokus pada penggunaan *Naive Bayes* standar atau mengombinasikannya dengan metode seleksi fitur tertentu.

Selain itu, hasil performa yang diperoleh pada setiap penelitian menunjukkan variasi yang cukup signifikan, baik dari sisi *accuracy, precision, recall*, maupun *F1-score.* Perbedaan hasil tersebut menunjukkan bahwa performa algoritma sangat dipengaruhi oleh karakteristik data, teknik prapemrosesan, serta metode yang digunakan. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengimplementasikan serta menganalisis kinerja algoritma *Gaussian Naive Bayes* dalam proses klasifikasi *spam* email.

Gaussian Naive Bayes merupakan salah satu varian dari algoritma *Naive Bayes* yang dimanfaatkan dalam proses klasifikasi. Algoritma ini mengasumsikan bahwa fitur-fitur yang digunakan untuk klasifikasi mengikuti distribusi *Gaussian* (normal) (Mujahidin *et al*., 2022). *Gaussian Naive Bayes* sangat berguna ketika fitur kontinu dapat diasumsikan mengikuti distribusi normal. Keuntungan utama dari *Gaussian Naive Bayes* adalah kesederhanaannya dan efisiensi komputasinya, terutama ketika fitur-fitur kontinu mengikuti distribusi normal. Algoritma ini tetap dapat beroperasi secara optimal meskipun asumsi independensi antar fitur tidak sepenuhnya terpenuhi (Mujahidin *et al*., 2022).

Gaussian Naive Bayes telah dimanfaatkan pada beragam bidang, di antaranya pengenalan teks, identifikasi wajah, hingga klasifikasi data medis (Hilda Rachmi, 2023). Oleh karena itu, penelitian ini difokuskan pada penerapan algoritma *Gaussian Naive Bayes* untuk klasifikasi email *spam*. Selain itu, penelitian ini juga bertujuan untuk mengimplementasikan sekaligus menilai performa algoritma tersebut dalam proses klasifikasi *spam* email. Proses penelitian mencakup beberapa tahapan utama, yaitu *pre-processing* data, pembangunan dan pelatihan model, tahap pengujian, serta evaluasi kinerja menggunakan metrik seperti *accuracy, precision, recall,* dan *F1-score*. Dengan pendekatan tersebut, diharapkan hasil penelitian ini mampu memberikan kontribusi dalam pengembangan sistem deteksi *spam* email yang lebih optimal, baik dari segi efektivitas maupun efisiensi.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah diuraikan pada latar belakang, dapat dirumuskan permasalahan yang akan dikaji pada penelitian ini sebagai berikut:

1. Bagaimana implementasi algoritma *Gaussian Naive Bayes* dalam melakukan

klasifikasi *spam email*?

2. Berapa nilai akurasi, presisi, *recall*, dan *F1-score* yang diperoleh dari penerapan algoritma *Gaussian Naive Bayes* pada proses klasifikasi *spam email?*

1.3 Tujuan Penelitian

Adapun tujuan penelitian ini berdasarkan rumusan masalah yang telah ditetapkan adalah sebagai berikut:

1. Mengimplementasikan Algoritma *Gaussian Naive Bayes* untuk Klasifikasi *Spam Email* dengan mengembangkan dan menerapkan model *Naive Bayes* untuk mengklasifikasikan *email* sebagai *spam* atau *non-spam* berdasarkan fitur-fitur yang relevan.
2. Mencari nilai akurasi, presisi, *recall*, dan *F1-score* untuk memastikan efektivitas model dalam mengklasifikasikan *email*.

1.4 Manfaat Penelitian

Berdasarkan latar belakang yang diuraikan di atas, manfaat dari penelitian ini adalah:

1. Sebagai referensi yang baik untuk penelitian

   Penelitian ini diharapkan dapat menjadi referensi yang berharga bagi para peneliti, terutama bagi mereka yang tertarik dalam mengimplementasikan algoritma *Gaussian Naive Bayes* untuk klasifikasi *spam email*. Dengan perbaikan yang tepat, penelitian ini akan menjadi sumber informasi yang lebih lengkap dan sempurna.
2. Sebagai landasan untuk pengembangan skripsi yang berkualitas

   Penelitian ini diharapkan dapat menjadi fondasi yang kuat bagi peneliti skripsi, memberikan landasan yang solid untuk mengembangkan skripsi yang lebih berkualitas. Dengan memanfaatkan temuan dan metodologi yang terdokumentasi dengan baik, peneliti dapat mengeksplorasi topik lebih dalam untuk mencapai kesuksesan dalam menyelesaikan studi mereka.

1.5 Batasan Masalah

Agar pembahasan dalam penelitian ini tetap terarah dan tidak menyimpang dari topik yang sudah ditetapkan, maka penulis membuat batasan dalam penelitian ini antara lain:

1. Data yang digunakan dalam penelitian ini terdiri dari sampel *email* yang telah dikumpulkan sebanyak 5.136 data dan diklasifikasikan sebelumnya sebagai *spam* atau *non-spam.*
2. Pada penelitian ini penulis hanya berfokus untuk melakukan analisis data *spam email.*
3. Data *email* yang digunakan berbahasa inggris
4. *Tool* yang digunakan adalah *Google Colab*