BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam keseharian manusia pasti ada peristiwa yang terjadi, baik yang disadari maupun yang tidak disadari. Setiap peristiwa tersebut yang sudah, sedang, dan akan terjadi masing-masing memiliki sebuah nilai yang disebut dengan peluang terjadinya peristiwa tersebut. Nilai-nilai tersebut akan menyebar dan membentuk suatu pola sebaran tertentu, yang disebut dengan distribusi probabilitas. Distribusi probabilitas adalah suatu fungsi yang mendefinisikan nilai dari suatu percobaan acak yang disebut variabel random dimana domain dari variabel random bergantung kepada ukuran sampel dari setiap distribusi probabilitas. Secara umum distribusi pobabilitas terbagi atas dua jenis, yaitu distribusi kontinu dan distribusi diskrit. Distribusi peluang kontinu digunakan ketika variabel random terdapat secara acak dalam suatu interval, dan distribusi peluang diskrit digunakan ketika variabel random merupakan suatu nilai spesifik (Thomopoulos, 2017).

Setiap distribusi memiliki kemiripan dengan distribusi yang lain. Namun dalam melihat kemiripan antar distribusi, terdapat dua sudut pandang yaitu berdasarkan sifat dan berdasarkan komputasi. Pada penelitian yang dilakukan oleh Gupta dan Kundu (2001), dijelaskan bahwa berdasarkan teori, fungsi hazard distribusi Exponensial memiliki sifat yang mirip dengan distribusi Gamma dimana jika nilai parameter bentuk distribusi gamma > 1, maka fungsi hazard akan bergerak dari 0 menuju ke suatu nilai berhingga kemudian akan bergerak secara konstan di nilai berhingga tersebut. Sedangkan jika nilai parameter bentuk distribusi Weibull > 1, maka fungsi hazard akan bergerak dari 0 menuju tak hingga. Namun berdasarkan komputasi, jika distribusi Gamma, Weibull, dan Eksponensial digunakan untuk memodelkan suatu sebaran data, misalnya pada penelitian Gupta dan Kundu (2001) mengenai data uji ketahanan bantalan bola alur dalam, hasil dari distribusi Exponensial dalam hal ini nilai ekspektasi, justru lebih mirip dengan distribusi Weibull.

Perbedaan ini mengakibatkan banyak peneliti yang mengalami kekeliruan dalam memilih dua distribusi atau lebih untuk digunakan secara bersamaan atau

sebaliknya membandingkan distribusi-distribusi tersebut. Pada penelitian yang dilakukan oleh Manurung, Ariswoyo dan Sembiring (2013) dibandingkan distribusi Binomial dan Poisson karena dalam kondisi tertentu nilai binomial akan sulit untuk ditentukan. Sehingga dilakukan pendekatan dengan menggunakan distribusi Poisson untuk kondisi tertentu. Selain itu, Stevanovic (2020) melakukan penelitian terkait penggunaan distribusi Weibull dan Poisson untuk menilai panjang umur sebuah saklar dengan menggunakan data tahun 2017-2020. Hasil penelitian menunjukkan terjadi peningkatan angka koefisien korelasi, yang berarti data pada tahun 2017-2020 semakin mendekati model sebaran distribusi Weibull jika dibandingkan dengan penelitian sebelumnya. Kemudian Stevanovic menguji data yang sama dengan menggunakan distribusi Poisson. Hasil penelitian menunjukkan bahwa hampir 90% dari data tersebut mengikuti sebaran poisson. Sehingga Stevanovic menyimpulkan bahwa sisa waktu pakai untuk sebuah saklar dapat dihitung dengan menggunakan distribusi Weibull dan Poisson.

Untuk mengatasi masalah di atas, dalam skripsi ini penulis menggunakan nilai statistik uji berdasarkan perhitungan komputasi. Selanjutnya nilai statistik uji dikelompokkan menggunakan salah satu metode pengelompokan di bidang Data Mining yaitu klasterisasi. Klasterisasi atau pengelompokan data pertama kali ditemukan oleh Driver dan Kroeber (1932). Kemudian seiring berjalannya waktu, klasterisasi data semakin populer. Seperti yang dilakukan oleh Moraru (2019) dalam penelitiannya, dilakukan klasterisasi data hasil pemindaian otak. Dimana Moraru dkk. menggunakan Gausian Mixture Model (GMM) untuk mendeteksi kesimetrisan antara otak kiri dengan otak kanan, kemudian hasil yang peroleh akan diklaster berdasarkan jenis kerusakan jaringan pada otak dengan menggunakan metode k-means. Hasil penelitian menunjukkan 18 sub-kelas dari data pemindaian otak diklaster ke dalam 6 tingkat pembobotan difusi yang berbeda dan 3 jaringan otak utama. Lima (2021) melakukan klasterisasi wilayah di Rio de Janeiro berdasarkan intensitas curah hujan ekstrim dengan menggunakan metode Ward. Hasil penelitian menunjukkan berdasarkan intensitas curah hujan ekstrim, wilayah di Rio de Janeiro terbagi atas 5 klaster. Safdar (2022) melakukan klasterisasi pasien berdasarkan jenis pasien menggunakan metode k-means. Hasil penelitian

menunjukkan terbentuk 4 klaster pasien. Namun, sampai saat ini belum ada peneliti yang melakukan penelitian mengenai klasterisasi distribusi probabilitas.

Berdasarkan uraian diatas penulis tertarik untuk melakukan pengelompokan distribusi probabilitas dengan distribusi yang digunakan adalah distribusi probabilitas kontinu yaitu distribusi Erlang, Fatigue Life, Frechet, Gamma, Log Logistik, Pareto, Pearsson Tipe 5 dan Weibull. Delapan distribusi probabilitas tersebut kemudian akan diklaster ke dalam beberapa klaster dengan menggunakan metode Single Linkage Clustering berdasarkan nilai statistik uji yang diperoleh melalui Uji Kolmogorv-Smirnov dan Uji Anderson-Darling. Uji Kolmogorv-Smirnov dan Uji Anderson-Darling merupakan teknik uji hipotesis yang memeriksa apakah sebuah sebaran sampel mengikuti sebaran distribusi tertentu.

Pada penelitian ini, digunakan *Index* Davies-Bouldin dalam menentukan jumlah klaster optimal. *Index* Davies-Bouldin merupakan salah satu metode evaluasi klaster dalam pengelompokan data dengan prinsip memaksimalkan jarak antar klaster dan meminimalkan jarak antar data pada sebuah klaster. Semakin kecil nilai Index Davies-Bouldin yang diperoleh menunjukkan bahwa klaster yang terbentuk semakin optimal.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah :

1. Bagaimana hasil estimasi parameter setiap distribusi dengan metode estimasi kemungkinan maksimum?
2. Distribusi apa saja yang mirip berdasarkan metode *Single Linkage Clustering* ?
3. Berapa jumlah klaster optimal menggunakan *Index* Davies-Bouldin ?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah :

1. Data yang digunakan adalah data yang digenerate secara acak dengan menggunakan *software* Matlab,
2. Distribusi yang digunakan adalah distribusi probabilitas kontinu dengan 2 parameter yaitu parameter bentuk dan parameter skala,

3. Jumlah klaster terbaik ditentukan dengan menggunakan metode *Index* Davies-Bouldin.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini adalah :

1. Untuk mengetahui hasil estimasi parameter setiap distribusi dengan metode estimasi kemungkinan maksimum,
2. Untuk mengetahui distribusi yang mirip menggunakan metode *Single Linkage Clustering*,
3. Untuk mengetahui jumlah klaster optimal menggunakan *Index* Davies-Bouldin.

1.5 Manfaat Penelitian

Adapun manfaat dari tulisan ini,

Bagi penulis :

1. Untuk memahami kolaborasi antara bidang statistik infrensi dan data mining,
2. Untuk memahami cara mengimplementasikan *software* Matlab dan RStudio.

Bagi pembaca :

1. Sebagai acuan bagi mahasiswa jurusan matematika dalam mempelajari mata kuliah Statistik Matematika,
2. Sebagai acuan bagi para peneliti bidang Bioekologi Laut dalam menentukan Model sebaran data spesies,
3. Untuk memberikan alternatif bagi para peneliti di bidang statistika dalam memilih model sebaran kontinu.

1.6 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah metode kuantitatif. Dimana dalam penelitian ini terdapat proses pengumpulan data, kemudian data akan dianalisis dengan menggunakan teknik statistik, dan akan dilakukan penarikan kesimpulan dari hasil yang diperoleh.

1.7 Sistematika Penulisan

Adapun sistematika penulisan yang digunakan dalam penelitian ini adalah sebagai berikut :

BAB I : Bab ini merupakan bagian pendahuluan dari tulisan ini yang berisi tentang penjelasan dari latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian, serta sistematika penelitian.

BAB II : Bab ini membahas tentang beberapa teori yang nantinya digunakan dalam pembahasan masalah yang akan dipaparkan dalam BAB IV.

BAB III : Bab ini membahas tentang jenis dan sumber data, waktu dan tempat penelitian, metode penelitian, serta tahapan penelitian.

BAB IV : Bab ini membahas tentang uraian pembahasan masalah, yakni tentang analisis klastering multi-distribusi kontinu.

BAB V : Bab ini merupakan bagian penutup dari tulisan ini yang terdiri dari kesimpulan dan saran.