BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Penelitian ini mengacu pada beberapa penelitian yang pernah dilakukan sebelumnya, yakni sebagai berikut:

Beny (2017), pada penelitian ini penulis melakukan proses *exploratory data analysis* pada pangkalan data sistem informasi Sekolah Tinggi Ilmu Kompter Dinamia Bangsa untuk mendapatkan gambaran pola-pola, sebaran data, dan korelasi antar variabel. Hasil dari penelitian ini terbentuklah data yang lebih rapih, tergambarkannya sebaran data nilai ipk berdasarkan beberapa variabel lainnya, dan teridentifikasinya beberapa variabel yang memiliki korelasi. Dari hasil yang didapatkan tersebut dapat disimpulkan bahwa variabel angkatan terhadap rata-rata ipk menunjukkan nilai korelasi *Pearson Product Moment* sebesar 0.91 dengan *Confidence Interval* 95% dan *p-value* 0.0003, ini menandakan korelasi yang secara *statistic* signifikan.

Rayhanali Heiko Amier dan Johan Setiawan (2019), tujuan dari penelitian ini adalah untuk memberikan gambaran yang komprehensif dengan membuat visualisasi nominasi Film Terbaik *Academy Awards* 1993 - 2017. Prosedur ini belum pernah dilakukan sebelumnya terutama dalam bentuk *dashboard* dengan menggunakan metode *Visual Data Mining* (VDM) pada *software tools* Tableau. Data yang

digunakan dalam penelitian ini diambil dari tiga sumber berbeda yang tersedia di Internet. Hasil dari penelitian ini adalah 5 *dashboard* yang berbeda dibuat untuk memvisualisasikan pencarian pola tren untuk parameter dan keberhasilan penerapan metode prediksi pemulusan eksponensial *Holt-Winters*. Penelitian ini divalidasi dengan menerapkan *User Acceptance Test* (UAT) kepada 5 responden.

Andrew T. Jebb, Scott Parrigon dan Sang Eun Woo (2017), peneliti yang membahas logika dan metode *exploratory data analysis* (EDA), mode analisis yang berkaitan dengan penemuan, eksplorasi, dan deteksi secara empiris fenomena dalam data. Peneliti memulai dengan mendeskripsikan latar belakang historis dan konseptual EDA. Kemudian membahas dua masalah yang berkaitan dengan EDA dan hubungannya dengan kredibilitas ilmiah. Pertama, peneliti berpendapat bahwa EDA mendorong ilmu berbasis replikasi dengan mensyaratkan validasi silang dan dengan menekankan *natural uncertainty of data patterns*. Kedua, peneliti mengklarifikasi bahwa EDA dapat dibedakan dari praktik eksplorasi lain yang dianggap dapat dipertanyakan secara ilmiah (misalnya, "peretasan p", "penangkapan data", dan "pengerukan data"). Pada bagian makalah berikut ini, kami menyajikan argumen terakhir untuk EDA: yang membantu memaksimalkan nilai data. Untuk mengilustrasikan poin ini, peneliti menyajikan beberapa metode grafis untuk mendeteksi pola.

Brillinger David R, Haiganoush K. Preisler, Alan A. Ager dan John G. Kie (2004), Dalam karya makalah ini, model berbasis persamaan diferensial stokastik

dikembangkan secara berurutan. Persamaan gerak ditetapkan dimotivasi oleh persamaan fisika yang sesuai. Parameter fungsional yang muncul dalam persamaan diestimasi secara nonparametrik dan plot bidang vektor pergerakan hewan disiapkan. Residu digunakan untuk mencari interaksi antar pergerakan hewan. Ada berbagai macam analisis eksplorasi. Kesimpulan statistik didasarkan pada transformasi Fourier dari data, yang jaraknya tidak sama.

Dey, SK, Rahman, MM, Siddiqi, UR, dan Howlader, A (2020), Dalam studi ini, peneliti menyajikan upaya untuk menghimpun dan menganalisis informasi epidemiologi wabah *COVID-19* berdasarkan beberapa *open dataset 2019-nCoV* yang disediakan oleh *Johns Hopkins University*, *World Health Organization, Chinese Center for Disease Control and Prevention, National Health Commission*, dan DXY. Analisis data eksplorasi dengan visualisasi telah dibuat untuk memahami jumlah kasus berbeda yang dilaporkan (dikonfirmasi, meninggal, dan pulih) di berbagai provinsi di China dan di luar China.

Kementerian Pekerjaan Umum Dan Perumahan Rakyat (2017), mengacu pada Modul Sistem Informasi Banjir, subbab Siaga Banjir; Status siaga banjir merupakan hasil analisa dari informasi yang didapatkan dari stasiun-stasiun pengamatan Tinggi Muka Air (TMA) yang ada di sungai-sungai. Semakin tinggi TMA-nya, kian tinggi pula status siaganya. Penetapan status siaga sebagai berikut:

1. Siaga IV (Normal): Belum ada peningkatan debit air secara mencolok. komando di lapangan, termasuk membuka atau menutup pintu air serta akan

dikemanakan arah air cukup dilakukan oleh komandan pelaksana dinas atau wakil komandan operasional wilayah.

2. Siaga III: Hujan yang terjadi menyebabkan terjadinya genangan air di lokasi-lokasi tertentu tetapi kondisinya masih belum kritis dan membahayakan. Penanganannya diserahkan pada masing-masing suku dinas pembinaan mental dan kesejahteraan sosial (Bintal Kesos) di masing-masing wilayah.

3. Siaga II: Bila wilayah genangan air mulai meluas, maka akan ditetapkan Siaga II, penanggungjawab untuk siaga II ini adalah Ketua Harian Satkorlak Penanggulangan Bencana Provinsi (PBP) yaitu Sekretaris Daerah.

4. Siaga I: Bila dalam enam jam genangan air tersebut tidak surut dan kritis maka ditetapkan Siaga I. Penanggung jawab penanganan status siaga I langsung ditangan Gubernur.

Tabel 2.1 Tabel Penelitian Terkait

Penulis	Tahun	Data Penelitian	Metode	Hasil
Rayhanali Heiko Amier, Johan Setiawan	2019	Film award nominations	Virtual data mining and exploratory data analysis	5 different dashboards were created to visualize the trend pattern search for the parameters and the successful application of the

				Holt-Winters exponential smoothing prediction method.
Beny	2017	Database sistem informasi akademik STIKOM Dinamika Bangsa	Exploratory data analysis	Terbentuknya data yang lebih rapih, tergambarkannya sebaran data nilai ipk berdasarkan beberapa variabel lainnya, dan teridentifikasinya beberapa variabel yang memiliki korelasi.
Andrew T. Jebb, Scott Parrigon, Sang Eun Woo	2017	Exploratory data analysis	Exploratory data analysis	Final argument for EDA: that it helps maximize the value of data.
David R, Brillinger., Haiganoush K, Preisler, Alan A,	2004	Paths of moving animals	Exploratory data analysis	Statistical inferences are based on Fourier transforms of the data, which are

Ager, dan John G Kie				unequally spaced.
Dey, SK, Rahman, MM, Siddiqi, UR, dan Howlader, A	2020	Open datasets on 2019-nCoV provided by the Johns Hopkins University, World Health Organization, Chinese Center for Disease Control and Prevention, National Health Commission, and DXY	Exploratory data analysis	Visual exploratory data analysis approach.
Kementerian Pekerjaan Umum Dan Perumahan	2017	Modul Sistem Informasi Banjir		Mengacu pada Modul Sistem Informasi Banjir, subbab Siaga

Rakyat				Banjir; Status siaga banjir merupakan hasil analisa dari informasi yang didapatkan dari stasiun-stasiun pengamatan Tinggi Muka Air (TMA) yang ada di sungai-sungai.
Rahmat Hadi Suko Wijoyo	2021	Tinggi Muka Air di Jakarta	Exploratory data analysis	Tinggi air tertinggi tercatat pada nama_pintu_air PA. Manggarai dengan tinggi_air sebesar 9690, terletak pada lokasi Ciliwung berkoordinat pada latitude -6.207825 longitude 106.848458 pada

				tanggal 2020-01-02 00:40:00 dengan status_siaga tertinggi yaitu Status : Siaga 1.

2.2. Dasar Teori

2.2.1. Python

Python adalah bahasa pemrograman interpretatif multiguna. Tidak seperti bahasa lain yang susah untuk dibaca dan dipahami, *Python* lebih menekankan pada keterbacaan kode agar lebih mudah untuk memahami sintaks. Hal ini membuat *Python* sangat mudah dipelajari baik untuk pemula maupun untuk yang sudah menguasai bahasa pemrograman lain.

Bahasa pemrograman *Python* muncul pertama kali pada tahun 1991, yang dirancang oleh seseorang bernama Guido van Rossum. Sampai saat ini *Python* masih dikembangkan oleh *Python Software Foundation*. Bahasa *Python* mendukung hampir semua sistem operasi, bahkan untuk sistem operasi *Linux*, dan hampir semua distronya sudah menyertakan *Python* di dalamnya.

2.2.2. Jupyter Notebook

Jupyter Notebook merupakan *tools* yang populer untuk pengolahan data dengan bahasa pemrograman *Python*. *Jupyter Notebook* memungkinkan untuk mengintegrasikan antara kode dengan *output* di dalam satu dokumen secara interaktif. *Jupyter* (https://jupyter.org/) sendiri adalah organisasi non-profit untuk mengembangkan *software* interaktif dalam berbagai bahasa pemrograman. Sedangkan *Notebook* adalah satu *software* buatan *Jupyter*, berupa aplikasi *web open-source* yang memungkinkan untuk membuat dan berbagi dokumen interaktif yang berisi *live code*, persamaan, visualisasi, dan teks naratif yang kaya.

2.2.3. Exploratory Data Analysis

Exploratory Data Analysis mengacu pada proses kritis dalam melakukan investigasi awal pada data untuk menemukan pola, anomali, menguji hipotesis, dan memeriksa asumsi dengan bantuan statistik ringkasan dan representasi grafis. Dengan melakukan EDA, maka kondisi *dataset* yang dimiliki dapat lebih dipahami. Memahami kondisi *dataset* dapat merujuk pada sejumlah hal termasuk namun tidak terbatas pada poin-poin berikut:

1. Mengekstrak variabel penting dan meninggalkan variabel yang tidak berguna (*data preprocessing*).
2. Mengidentifikasi nilai yang hilang (*missing values*) dan kesalahan manusia (*human error*).
3. Memahami hubungan, atau kekurangan, antar variabel.

4. Memaksimalkan wawasan yang dimiliki atas kondisi data dan meminimalkan potensi kesalahan di kemudian hari.

EDA menjadi penting karena tanpa melakukan *Exploratory Data Analysis*, maka banyak informasi penting yang terdapat di dalam *dataset* bisa hilang begitu saja. Selain itu, meskipun memakan waktu yang relatif cukup lama, sebenarnya EDA akan menghemat waktu pengerjaan proyek *data science*. Karena apabila dilakukan *data modelling* tanpa menerapkan EDA, besar kemungkinan model yang akan dikerjakan memiliki performa yang kurang baik karena model yang dibuat tanpa benar-benar memahami kondisi data yang dimiliki. Lalu besar kemungkinan akan menghabiskan banyak waktu untuk mencari kesalahan apa yang harus diperbaiki, dan mengulang proses *data modelling* kembali. Tentu proses ini sangat memakan waktu. Dengan melakukan EDA, maka tidak perlu melakukan pengulangan seperti ini atau setidaknya mengurangi kemungkinannya.

2.2.4. Pandas

Pandas merupakan sebuah *library* pada *Python* yang berlisensi *BSD* dan *open source* yang menyediakan struktur data dan analisis data yang mudah untuk digunakan. *Pandas* melakukan tugas penting seperti menyelaraskan data untuk perbandingan dan penggabungan *dataset*, penanganan data yang hilang, dan lain sebagainya.

Struktur data *Pandas* dinamakan *dataframe*, yaitu sebuah koleksi kolom yang berurutan dengan nama dan jenis. Dengan adanya fitur *dataframe* memudahkan untuk

membaca sebuah file dan menjadikannya sebuah *table*. *Dataframe* juga dapat mengolah suatu data dengan menggunakan operasi seperti *join*, *distinct*, *group by*, agregasi, dan fitur lainnya yang terdapat pada *SQL*. Format file yang didukung *Pandas* meliputi file dengan ekstensi .txt, .csv, .tsv dan lainnya (Mutmainnah, 2019).

2.2.5. Numpy

Numpy merupakan salah satu *library* dalam *Python*. *Library* ini dapat digunakan untuk banyak *case* dalam *data science*. Dengan adanya *library* ini, maka tidak diperlukan lagi baris kode yang panjang untuk menjalankan program *machine learning*. *Numpy* sendiri merupakan singkatan dari *Numerical Python*. Pada umumnya penggunaan *library* ini untuk menghitung operasi matematika pada *array*.

2.2.6. Matplotlib

Matplotlib adalah *library Python* 2 Dimensi yang dapat menghasilkan plot dengan kualitas tinggi dalam berbagai format dan dapat digunakan di banyak *platform*. *Matplotlib* dapat digunakan sebagai pembuat grafik dalam berbagai *platform*, seperti *Python* dan *Jupyter Notebook*. Grafik yang dapat dibuat oleh matplotlib cukup beragam, seperti grafik garis, batang, lingkaran, histogram dan berbagai macam lainnya.

2.2.7. Seaborn

Seaborn adalah *library* yang bertujuan untuk membuat grafik dan statistik dengan menggunakan *Python*. *Seaborn* dibangun berdasarkan *library matplotlib* serta terintegrasi dengan struktur data pada *Pandas*.

2.2.8. Plotly

Plotly adalah *library* untuk pembuatan plot yang tersedia dalam bahasa pemrograman *Python* dan *R*. dari segi kompatibilitas pada diagram, *library* ini tidak jauh berbeda dengan *matplotlib*. Plot garis, diagram batang, hingga *heatmaps* merupakan keunggulan dari *Plotly*. Pada *Plotly* secara *default* sudah tersedia beberapa *tools* yang mendukung interaksi pada plot, sebagai contoh pembesaran diagram dan tombol *screenshot* tersedia secara otomatis. Berbeda dengan *matplotlib*, penulisan kode secara manual diperlukan untuk menyimpan hasil diagram dari *Plotly*.

2.2.9. Pandas Profiling

Pandas Profiling adalah *open source library* yang menghasilkan laporan interaktif untuk kumpulan data apa pun, hanya dengan menggunakan satu baris kode. *Pandas Profiling* menghasilkan laporan profil dari *Pandas dataframe*. Fungsi *Pandas* df.describe() sedikit mendasar untuk analisis data eksplorasi, maka *Pandas Profiling* memperluas fungsi *dataframe* dengan menggunakan ProfileReport(df) untuk proses analisis data yang lebih cepat.

Untuk setiap kolom, statistik berikut disajikan dalam laporan HTML interaktif:

- *Overview* yang terdiri dari *Overview, Warnings* dan *Reproduction*
- *Variables* yang terdiri dari *Statistics, Histogram, Common values* dan *Extreme values*
- *Quantile statistics* seperti nilai *minimum, Q1, median, Q3, maximum, range* dan *interquartile range*
- *Descriptive statistics* seperti standard deviation, *coefficient of variation, kurtosis, mean, median absolute deviation, skewness, sum, variance* dan *monotonicity*
- *Interactions* menyoroti interaksi antar variabel data
- *Correlations* menyoroti variabel yang saling berkorelasi (matriks Pearson, Spearman, Kendall, Phik dan Cramer)
- *Missing values* menampilkan adanya data yang hilang atau tidak. Laporan disajikan dalam *Count* dan *Matrix*
- *Sample* yang terdiri dari *First rows* dan *Last rows* data
- *Duplicate rows* menampilkan baris data yang saling memiliki kesamaan antar satu dengan yang lainnya

2.3.0. Proses Pengambilan Data

Pengambilan data dilakukan untuk memperoleh informasi yang dibutuhkan dalam rangka mencapai tujuan penelitian. Sebelum melakukan penelitian, seorang peneliti biasanya telah memiliki dugaan berdasarkan teori yang ia gunakan, dugaan tersebut disebut dengan hipotesis. Untuk membuktikan hipotesis secara empiris,

seorang peneliti membutuhkan pengambilan data untuk diteliti secara lebih mendalam.

Proses pengambilan data ditentukan oleh beragam variabel yang terdapat dalam hipotesis. Pengambilan data dilakukan terhadap sampel yang telah ditentukan sebelumnya. Data sendiri merupakan sesuatu yang belum memiliki arti bagi penggunanya dan masih membutuhkan adanya suatu pengolahan. Data bisa memiliki berbagai wujud, mulai dari gambar, suara, huruf, angka, bahasa, simbol, bahkan keadaan.

2.3.1. Proses Import Data

Import data adalah suatu metode untuk pengambilan data atau *file* dari luar baik itu dari aplikasi itu sendiri maupun dari aplikasi lain. *Import data* bekerja dengan mengunggah *file* ke dalam suatu aplikasi yang akan digunakan. Data yang di *import* dapat digunakan untuk analisis maupun proses pengolahan data lainnya.

2.3.2. Proses Preprocessing Data

Preprocessing data merupakan teknik awal *data mining* untuk mengubah data mentah yang dikumpulkan dari berbagai sumber menjadi informasi yang dapat digunakan untuk pengolahan data selanjutnya. 3 masalah umum yang diselesaikan dalam tahap *preprocessing data* adalah menangani *missing value*, *data noise*, dan data yang tidak konsisten.

Missing value merupakan data yang tidak akurat dikarenakan informasi yang hilang menyebabkan informasi yang ada di dalamnya tidak relevan. *Missing value* sering terjadi ketika terjadi masalah dalam proses pengumpulan data, seperti kesalahan dalam *entry data* atau masalah dalam penggunaan biometrik. *Data noise* berisi data yang salah yang dapat ditemukan di kumpulan data. Beberapa penyebab adanya *data noise* adalah karena kesalahan manusia, berupa kesalahan pemberian label dan masalah lain selama pengumpulan data. Inkonsistensi data terjadi ketika seseorang menyimpan file yang berisi data yang sama dengan format yang berbeda-beda. Beberapa inkonsisten data adalah duplikasi dalam format yang berbeda, kesalahan pada kode nama, dan lain sebagainya.

2.3.3. Acuan Status Siaga

Parameter atau nilai acuan status siaga untuk tiap pintu air seperti yang terlampir pada Tabel 2.2 di dapatkan dari Pusat Data dan Informasi Sumber Daya Air Dinas Sumber Daya Air Provinsi DKI Jakarta (http://poskobanjirdsda.jakarta.go.id/Pages/grafikDataTinggiMukaAir.aspx).

Untuk tiap pintu air memiliki batas ketinggian status siaga yang berbeda-beda. Sebagai contoh nama pintu air Bendung Cibalok – Gadog memiliki batas Status Siaga : Normal dengan ketinggian <150, Status Siaga : Siaga 3 dengan ketinggian 150, Status Siaga : Siaga 2 dengan ketinggian 250, dan Status Siaga : Siaga 1 dengan ketinggian >300.

Tabel 2.2 Parameter Status Siaga Tiap Pintu Air

Nama Pintu Air	Status Siaga : Normal	Status Siaga : Siaga 3	Status Siaga : Siaga 2	Status Siaga : Siaga 1
Bendung. Cibalok - Gadog	<150	150	250	>300
Bendung. Katulampa (Hulu)	<80	80	150	>200
PS. Depok	<200	200	270	>350
PA. Manggarai	<750	750	850	>950
PS. Krukut Hulu	<150	150	250	>300
Pompa Cideng	<150	150	250	>300
P.A. Karet	<450	450	550	>600
P.A. Marina Ancol	<170	170	200	>250
Pompa Pasar Ikan	<170	170	200	>250
Pompa. Pluit	<-50	-50	0	>45
PS. Pesanggrahan	<150	150	250	>350
PS. Angke Hulu	<150	150	250	>300
PS. Sunter Hulu	<140	140	200	>250
PA. Pulo Gadung	<550	550	700	>770
Pompa Yos Sudarso 1	<140	140	200	>250

PS. Cipinang Hulu	<150	150	200	>250
Pompa Kali Duri (Kalijodo)	<220	220	270	>320
P.A. Istiqlal	<250	250	300	>350
P.A. Jembatan Merah	<150	150	200	>250
P.A. Flusing Ancol	<180	180	190	>220
P.A. Hek	<200	200	250	>300