

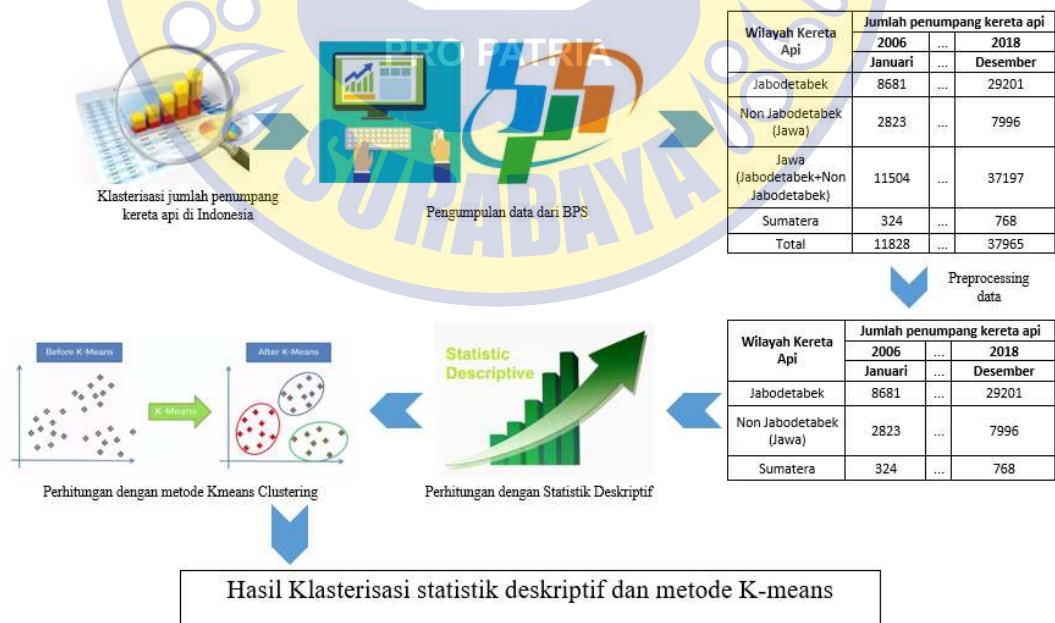
BAB III

METODELOGI PENELITIAN

Pada bab III ini menjelaskan tentang metodologi penelitian dan metode penelitian yang dilakukan pada penelitian ini.

3.1 Metodologi Penelitian

Pada penelitian ini terdapat 5 alur proses penelitian untuk melakukan klasterisasi jumlah penumpang kereta api di Indonesia, yakni dimulai dengan pengumpulan data, preprocessing data, perhitungan menggunakan statistik deskriptif, perhitungan menggunakan K-means, dan hasil klasterisasi. Seperti ditunjukkan pada gambar 3.1 Rancangan Sistem dibawah ini :



Gambar 3. 1 Rancangan Sistem

3.2 Pengumpulan Data

Data penelitian yang digunakan berasal dari Badan Pusat Statistik (BPS) yakni data jumlah penumpang kereta api di Indonesia. Tahapan awal pengumpulan data ini yakni memasuki website BPS kemudian masuk ke halaman publikasi lalu unduh data jumlah penumpang kereta api di Indonesia. Dalam penelitian ini penulis mengambil data jumlah penumpang kereta api di Indonesia dalam range waktu 2006 sampai 2018 per bulan. Data jumlah Dalam penelitian ini data yang digunakan adalah data kumlah penumpang kereta api di Indonesia dari wilayah Jabodetabek, luar Jabodetabek, dan Sulawesi. Jumlah data yang digunakan yakni sebanyak 159 record dengan 3 variabel pada setiap wilayah. Data yang didapat berupa format excel .xls (*Extensibe Stylesheet Language*). Data yang di gunakan atau yang akan di olah dalam penelitian ini memiliki satu type data yaitu integer. Pada penelitian ini penulis menggunakan googlecolab untuk melakukan perhitungan dengan program python yang telah disusun menggunakan algoritma K-means untuk melakukan klasterisasi terhadap jumlah penumpang kereta api di Indonesia.

3.3 Preprocessing data

Data yang didapat dari Badan Pusat Statistik (BPS) yakni memiliki 5 variabel yang akan di *preprocessing* dahulu agar dapat dilakukan perhitungan yang baik dan akurat. Berikut tabel data dari BPS sebelum di *preprocessing* untuk tampilan type data dan sampel data ditunjukkan pada tabel 3.1 tipe data BPS dan tabel 3.2 Sampel data jumlah penumpang dibawah ini:

Tabel 3. 1 Tipe data BPS

Variabel	Keterangan	Jenis data
Jabodetabek	Jumlah Penumpang	Integer
Luar Jabodetabek	Jumlah Penumpang	Integer
Jawa (Jabodetabek + Luar Jabodetabek)	Jumlah Penumpang	Integer
Sumatera	Jumlah penumpang	Integer
Total	Jumlah Penumpang	Integer

Tabel 3. 2 Sampel data jumlah penumpang

Tahun dan Bulan	Wilayah Kereta Api					
	Jabodetabek	Non Jabodetabek	Jawa (Jabodetabek + Luar Jabodetabek)	Sumatera	Total	
2006	Januari	8681000	2823000	11504000	324000	11828000
	Februari	8144000	3561000	11705000	226000	11931000
	Maret	8920000	4146000	13066000	248000	13314000
	April	8462000	4195000	12657000	252000	12909000
	Mei	8899000	4413000	13312000	263000	13575000
	Juni	8606000	4323000	12929000	274000	13203000
	Juli	8787000	5300000	14087000	346000	14433000
	Agustus	8661000	4330000	12991000	264000	13255000
	September	8829000	4354000	13183000	253000	13436000
	Oktober	8767000	5182000	13949000	341000	14290000
	November	8895000	4474000	13369000	262000	13631000
	Desember	8774000	4570000	13344000	270000	13614000
2007	Januari	8723000	4924000	13647000	313000	13960000
	Februari	7566000	3193000	10759000	210000	10969000
	Maret	9009000	4158000	13167000	242000	13409000
	April	10206000	3983000	14189000	226000	14415000
	Mei	10608000	4331000	14939000	293000	15232000

Lanjutan Tabel 3.2 Sampel Data Jumlah Penumpang

Tahun dan Bulan	Wilayah Kereta Api					
	Jabodetabek	Non Jabodetabek	Jawa (Jabodetabek + Luar Jabodetabek)	Sumatera	Total	
2007	Juni	10310000	4510000	14820000	284000	15104000
	Juli	10761000	5331000	16092000	362000	16454000
	Agustus	10653000	4491000	15144000	275000	15419000
	September	10446000	4346000	14792000	241000	15033000
	Oktober	9887000	5578000	15465000	401000	15866000
	November	10031000	4110000	14141000	250000	14391000
	Desember	9895000	4871000	14766000	318000	15084000
2008	Januari	9868000	4829000	14697000	330000	15027000
	Februari	9673000	4443000	14116000	262000	14378000
	Maret	10469000	5287000	15756000	315000	16071000
	April	10562000	4873000	15435000	276000	15711000
	Mei	10582000	5472000	16054000	309000	16363000
	Juni	10824000	5812000	16636000	374000	17010000
	Juli	11206000	6257000	17463000	425000	17887000
	Agustus	11129000	5641000	16769000	339000	17108000
	September	10481000	5123000	15604000	275000	15879000
	Oktober	10379000	6522000	16901000	436000	17337000
	November	10583000	5113000	15696000	277000	15973000
	Desember	9695000	5316000	15011000	321000	15332000

Dari tabel 3.1 tipe data BPS merupakan tipe data yang ada pada data jumlah penumpang kereta api di Indonesia, yaitu integer. Tabel 3.2 Sampel data jumlah penumpang merupakan jumlah penumpang kereta api di Indonesia yang terdiri dari 5 variabel yang tertera yaitu Jabodetabek, luar Jabodetabek, Jawa (Jabodetabek + luar Jabodetabek) dan Sumatera. Kemudian dari hasil pengambilan data dilakukan preprocessing data berikut tabel data dari hasil *preprocessing* untuk tampilan type data dan sampel data ditunjukkan dalam tabel

3.3 tipe data hasil *preprocessing* dan tabel 3.4 sampel data hasil *preprocessing* dibawah ini;

Tabel 3. 3 Tipe data hasil preprocessing

Variabel	Keterangan	Jenis data
Jabodetabek	Jumlah Penumpang	Integer
Luar Jabodetabek	Jumlah Penumpang	Integer
Sumatera	Jumlah penumpang	Integer

Tabel 3. 4 Sampel data hasil preprocessing

Tahun dan Bulan	Wilayah Kereta Api			
	Jabodetabek	Non Jabodetabek	Sumatera	
2006	Januari	8681000	2823000	324000
	Februari	8144000	3561000	226000
	Maret	8920000	4146000	248000
	April	8462000	4195000	252000
	Mei	8899000	4413000	263000
	Juni	8606000	4323000	274000
	Juli	8787000	5300000	346000
	Agustus	8661000	4330000	264000
	September	8829000	4354000	253000
	Oktober	8767000	5182000	341000
	November	8895000	4474000	262000
	Desember	8774000	4570000	270000
2007	Januari	8723000	4924000	313000
	Februari	7566000	3193000	210000
	Maret	9009000	4158000	242000

	April	10206000	3983000	226000
	Mei	10608000	4331000	293000

Lanjutan Tabel 3.4 Sampel data hasil *preprocessing*

Tahun dan Bulan		Wilayah Kereta Api		
		Jabodetabek	Non Jabodetabek	Sumatera
2007	Juni	10310000	4510000	284000
	Juli	10761000	5331000	362000
	Agustus	10653000	4491000	275000
	September	10446000	4346000	241000
	Oktober	9887000	5578000	401000
	November	10031000	4110000	250000
	Desember	9895000	4871000	318000
2008	Januari	9868000	4829000	330000
	Februari	9673000	4443000	262000
	Maret	10469000	5287000	315000
	April	10562000	4873000	276000
	Mei	10582000	5472000	309000
	Juni	10824000	5812000	374000
	Juli	11206000	6257000	425000
	Agustus	11129000	5641000	339000
	September	10481000	5123000	275000
	Oktober	10379000	6522000	436000
	November	10583000	5113000	277000
	Desember	9695000	5316000	321000

Preprocessing data dengan menghapus 2 variabel yang tidak diperlukan yakni variabel Jawa (Jabodetabek + luar Jabodetabek) dan total. Menghapus variabel Jawa (jabodetabek + luar Jabodetabek) dikarenakan hal tersebut adalah hasil penjumlahan dari Jabodetabek dan luar Jabodetabek. Menghapus variabel total dikarenakan total adalah penjumlahan dari jumlah penumpang kereta api di

Indonesia dari setiap wilayah. Jadi ada 3 variabel yang digunakan dalam penelitian ini yaitu Jabodetabek, luar Jabodetabek, dan Sumatera.

3.4 Statistik Deskriptif

Berdasarkan data yang telah di preprocessing, semua data yang ada akan dilakukan perhitungan dengan statistik deskriptif bertujuan untuk mendapatkan hasil pengelompokan. Proses perhitungan dimulai dengan cara menghitung Mean dari seluruh data dengan $\frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$, dimana x adalah data ke N dan N adalah banyaknya data. Setelah mendapatkan hasil Mean, dilakukan perhitungan Standart Deviasi dengan $\sqrt{\frac{\sum (x_i - \text{Mean})^2}{N}}$, dimana x adalah data ke N, N adalah banyaknya data, dan Mean ialah nilai rata-rata data. Menetapkan data ke N tersebut masuk kedalam kluster atau kelompok yang telah ditetapkan. Ada 3 kluster yang ditetapkan dalam penelitian ini yaitu rendah, sedang, dan tinggi.

Statistik deskriptif memiliki aturan menentukan kluster rendah, sedang dan tinggi. Kluster rendah ialah $X < \text{Mean} - \text{Standart Deviasi}$ dimana X adalah data ke N yang akan masuk ke kluster rendah apabila nilai X tidak lebih besar dari hasil rata-rata data dikurangi Standart Deviasi. Kluster sedang ialah $\text{Mean} - \text{Standart Deviasi} < X < \text{Mean} + \text{Standart Deviasi}$ dimana X adalah data ke N yang akan dimasukkan ke kluster sedang apabila nilai X lebih besar dari pada hasil rata-rata data dikurangi Standart Deviasi dan tidak lebih besar atau kurang dari hasil Mean ditambah Standart Deviasi.

Kluster tinggi ialah $X > \text{Mean} + \text{Standart Deviasi}$ dimana X adalah data ke N yang masuk ke kluster tinggi apabila nilai X lebih besar dari hasil rata-rata data

ditambah Standar Deviasi. Sehingga setiap data dapat mendapatkan klaster yang telah di tetapkan.

3.5 K-means

Berdasarkan variabel yang dimiliki, semua data yang ada akan dilakukan pengelompokkan dengan metode K-means. Proses clustering dimulai dengan mengidentifikasi data yang akan di cluster , $X_{ij}(i=1,\dots,n; j=1,\dots,m)$ dengan n adalah jumlah data yang akan di cluster dan m adalah jumlah variabel. Pada awal iterasi, cluster yang ditentukan ditetapkan secara acak atau *random*, $C_{ij}(k=1,\dots,k; j=1,\dots,l)$. Kemudian dihitung jarak antara setiap data dengan cluster yang telah ditentukan secara acak diawal. Berikut rumus Euclidiean Distance, seperti pada persamaan (1), yaitu :

$$(X,Y), (A,B) = \sqrt{(X - A)^2 + (Y - B)^2} \quad (1)$$

Suatu data akan menjadi anggota dari cluster ke-k apabila data tersebut ke pusat cluster k-k bernilai paling kecil jika dibandingkan dengan jarak ke cluster lainnya. Kelompokan data sesuai dengan clusternya. Nilai cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data yang menjadi anggota pada cluster tersebut, dengan menggunakan rumus pada persamaan (2) :

$$C_{ij} = \frac{\sum_i^p X_{ij}}{p} \quad (2)$$

Dimana X_{ij} cluster ke -k

P = banyaknya anggota cluster ke k

Proses Clustering menggunakan metode algoritma K-means memiliki langkah-langkah sebagai berikut :

- a. Inisialisasi : tentukan berapa K sebagai jumlah cluster yang diinginkan.
- b. Pilih K buah titik centroid secara acak atau *random*.
- c. Lakukan perhitungan data yang tersedia dengan K yang telah dipilih, sehingga menghasilkan nilai keterdekatan masing-masing data terhadap K.
- d. Kelompokkan data sehingga terbentuk K buah cluster dengan titik centroid dari setiap cluster merupakan titik centroid yang telah dipilih sebelumnya.
- e. Perbaharui nilai titik centroid
- f. Kemudian hitung kembali centroid berdasarkan data yang mengikuti cluster masing-masing
- g. Ulangi langkah c, d, e hingga kondisi konvergen tercapai yaitu perubahan fungsi objektif sudah dibawah ambang batas yang diinginkan, atau tidak ada yang berpindah cluster atau posisi centroid sudah di ambang batas yang ditetapkan.

3.6 Rencana Kerja

Penelitian ini menggunakan dataset jumlah penumpang kereta api di Indonesia dari BPS. Data jumlah penumpang kereta api yang didapatkan penulis akan disesuaikan dengan kebutuhan penelitian. Penulis hanya akan mengambil data jumlah penumpang kereta api di Indonesia. Adapun rencana kerja pada penelitian yakni sebagai berikut pada tabel 3.3 waktu pelaksanaan :

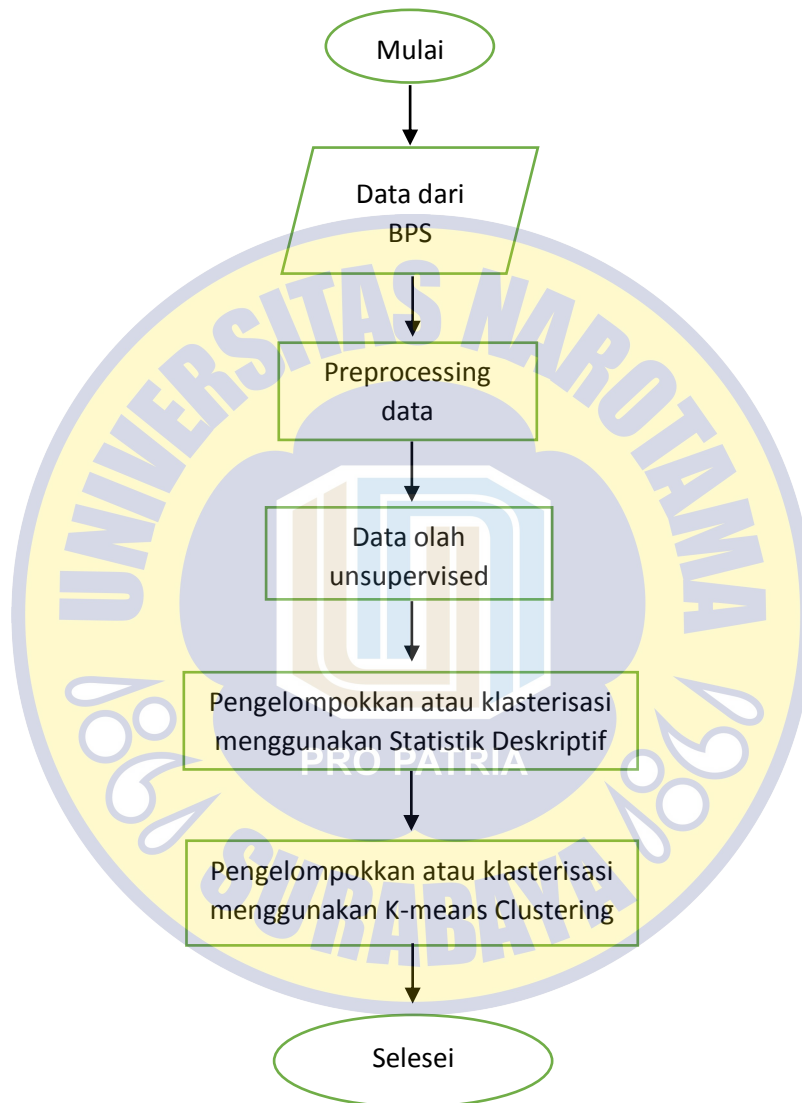
Tabel 3. 5 Waktu Pelaksanaan

Proses	Maret 2019				April 2019				Mei 2019			
	M 1	M 2	M 3	M 4	M 1	M 2	M 3	M 4	M 1	M 2	M 3	M 4
Perancangan proses kerja												
Pengumpulan data												
Data preprocessing												
Klustersasi dengan statistik deskriptif												
Klasterisasi dengan K-means												
Hasil Penelitian												

Pada minggu pertama dan kedua bulan Maret 2019 dilakukan perancangan proses kerja yang meliputi agenda langkah-langkah melakukan penelitian klasterisasi jumlah penumpang kereta api. Minggu kedua dan ketiga bulan Maret 2019 sepakat mengumpulkan data dari BPS.

Minggu terakhir Maret 2019 *preprocessing* data agar dapat dilakukan perhitungan yang baik. Pada minggu awal dan minggu kedua April 2019 dilakukan perhitungan menggunakan statistik deskriptif. Sedangkan minggu ketiga, keempat April 2019, minggu pertama dan minggu kedua Mei 2019 dilakukan perhitungan klasterisasi menggunakan metode K-means yang dilakukan perhitungannya menggunakan python di googlecolab. Pada minggu ketiga dan keempat Mei 2019 mendapatkan hasil penelitian klasterisasi menggunakan statistik deskriptif dan metode K-means.

Tahapan penelitian yang dilakukan dan dijelaskan dengan flowchart berikut ini :



Gambar 3. 2 Flowchart metodologi penelitian