ANALISIS *MACHINE LEARNING* DENGAN METODE *REGRESI LINIER* SEDERHANA DAN *NAÏVE BAYES* UNTUK MEMPREDIKSI POTENSI BANJIR

TUGAS AKHIR

Digunakan Sebagai Syarat Pengerjaan
Tugas Akhir Mata Kuliah *Machine Learning*

Oleh Kelompok 7 :

DAWAM ILHAMI ASSIDIQI	**NIM.2041720108**
IFTITAH HIDAYATI	**NIM.2041720006**
SAFIRA ISTIFARINI	**NIM.2041720229**

TI-3H

POLITEKNIK NEGERI MALANG
JURUSAN TEKNOLOGI INFORMASI
PROGRAM STUDI D4 TEKNIK INFORMATIKA

2022

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Banjir merupakan salah satu bencana alam yang paling sering terjadi di Indonesia. Dari total 31.659 desa/kelurahan di Indonesia yang mengalami bencana alam, 19.675 di antaranya adalah bencana banjir (Badan Pusat Statistik, 2018). Salah satu peristiwa yang dapat dapat mengakibatkan terjadinya banjir adalah penumpukan sampah.

Permasalahan sampah di Indonesia merupakan masalah yang menjadi mimpi buruk bangsa ini selama bertahun-tahun. Infrastruktur pengelolaan sampah di Indonesia dinilai masih belum memadai. Sementara itu, program untuk menanggulangi masalah pengelolaan sampah ini dinilai masih belum maksimal (Sembiring, 2018). Hal tersebut selaras dengan penelitian yang dilakukan di Desa Kerobokan Kelod, Kuta Utara dan Desa Pengastulan, Seririt yang menyatakan bahwa terdapat beberapa permasalahan dalam pengelolaan sampah di antaranya, pengumpulan sampah yang belum maksimal karena hanya dilakukan di kawasan tepi jalan raya yang dilengkapi dengan Tempat Pembuangan Sementara (TPS), kontainer pengangkut sampah sulit menjangkau kawasan terpencil (gang yang jauh dari lintasan truk sampah), memiliki jumlah terbatas, dan jarang beroperasi utamanya di masa pandemi, serta minimnya jumlah tempat sampah sehingga masyarakat kebingungan untuk membuah sampah dan berujung dengan menjadikan sungai sebagai TPA sehingga terjadi penumpukan sampah yang mengakibatkan terjadinya banjir (Putra dan Mandala, 2020; Wirawan dan Nandari, 2020).

Sampah yang menumpuk terlalu banyak menjadi penghambat aliran air di sungai sehingga disaat musim penghujan tiba volume air aliran sungai tersebut menjadi meluap dan menggenangi jalan raya, pasar dan pemukiman yang berada di dekat sungai (Putra dan Mandala, 2020; Wirawan dan Nandari, 2020). Berdasarkan penelitian yang dilakukan oleh Setiawan dkk., (2020) di Kota Samarinda, terdapat dua faktor utama penyebab banjir yaitu faktor alam (*natural*) dan faktor manusia (*man made*). Faktor alam berkaitan dengan curah hujan, topografi wilayah, dan pasang surut air sungai. Sedangkan faktor manusia berkaitan dengan alih guna lahan yang mengakibatkan berkurangnya daerah resapan air dan peningkatan produksi sampah yang tidak dibarengi dengan pengelolaan yang baik sehingga mengakibatkan terjadinya penyumbatan saluran drainase dan aliran sungai.

Bencana banjir mendatangkan berbagai kerugian terhadap masyarakat di daerah terdampak. Berdasarkan penelitian yang dilakukan di Kelurahan Rawa Makmur, Kota Bengkulu, terdapat dampak sosial ekonomi yang timbul akibat terjadinya banjir pada bulan April 2019 berupa terhambatnya aktivitas sehari-hari masyarakat seperti bekerja dan sekolah, timbulnya penyakit, dan kerugian ekonomi. Estimasi kerugian langsung masyarakat sebesar Rp. 1.742.957.130,66,- (Santri, Apriyanto dan Utama, 2020).Tidak hanya itu, bencana banjir juga memberikan efek

secara psikologis kepada korban terdampak seperti timbulnya perasaan pesimis terhadap masa depan, sering sakit kepala bila ada pikiran, mudah marah, sedih, dan gelisah (Amalia, Suzanna dan Dewi, 2021). Bahkan banjir juga dapat menimbulkan korban jiwa karena minimnya pencegahan terhadap akibat dari bencana banjir (Muzakky dkk., 2018).

Dampak bencana banjir juga menyebabkan kerugian dari berbagai sektor infrastruktur, karena dapat menghambat perkembangan dan kemajuan kota. Ketersediaan dan pelayanan infrastruktur yang baik dalam mengatur dan mengelola kegiatan publik dalam kehidupan sehari-hari sangat berpengaruh terhadap pengembangan sebuah kota dalam menentukan perekonomian di daerah tersebut (Saidi dkk., 2018; Yilema dan Gianoli, 2018). Kerusakan dan kerugian infrastruktur publik akan membebani pemerintah dalam tahap pemulihan pasca bencana banjir. Sampai saat ini, pembiayaan pemulihan akibat bencana banjir sebagian besar masih menjadi tanggung jawab pemerintah (Putra, Hermawan dan Hatmoko, 2020). Sebagai contoh, pada tahun 2020, pemerintah Kota Samarinda mengajukan anggaran penanggulangan banjir sebesar Rp. 315.000.000.000 dari APBD Provinsi Kalimantan Timur dan APBD murni Kota Samarinda sebesar Rp. 131.000.000.000 (Hutauruk, Kusuma dan Ningsih, 2020). Mengingat kejadian bencana banjir tidak bisa diprediksi kapan terjadinya dan keterbatasan anggaran untuk pemulihan pasca bencana, maka harus merencanakan strategi untuk penanggulangan risiko akibat bencana banjir sehingga dapat mengurangi beban anggaran pemerintah (Putra, Hermawan dan Hatmoko, 2020).

Penelitian yang dapat digunakan untuk mengantisipasi permasalahan banjir tersebut adalah inovasi pendeteksi terjadinya banjir berdasarkan jumlah produksi sampah seluruh daerah Jawa Timur menggunakan dua metode dalam machine learning yaitu regresi linier sederhana dan naive bayes. Berdasarkan jumlah produksi sampah pada sungai nantinya dilakukan suatu perkiraan atau prediksi terjadinya banjir dari jumlah sampah tersebut. Penelitian ini bertujuan untuk memprediksi adanya potensi banjir dari data produksi sampah.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini dapat disusun sebagai berikut:

1. Bagaimana tahapan pelaksanaan penelitian prediksi banjir berdasarkan jumlah sampah dengan menggunakan metode regresi linier sederhana dan naive bayes?
2. Bagaimana analisa dan perancangan kode program untuk memprediksi banjir berdasarkan jumlah sampah dengan menggunakan metode regresi linier sederhana?
3. Bagaimana analisa dan perancangan kode program untuk memprediksi banjir berdasarkan jumlah sampah dengan menggunakan metode naive bayes?

1.3 Batasan Masalah

Batasan masalah yang terapkan dari penelitian adanya potensi banjir tersebut adalah :

1. Lingkup daerah pengambilan data terletak di Provinsi Jawa Timur.
2. Studi kasus penyebab banjir meliputi jumlah sampah.
3. Penelitian terdiri dari dua metode yaitu Decision Tree dan K-Mean.

1.4 Tujuan

Tujuan yang diharapkan dari penelitian adanya potensi banjir tersebut yaitu :

1. Mengetahui tahapan pelaksanaan penelitian prediksi banjir berdasarkan jumlah sampah dengan menggunakan metode regresi linier sederhana
2. Membuat sebuah prediksi perkiraan banjir berdasarkan jumlah sampah.
3. Memberikan perbandingan antara metode Regresi linier sederhana dan naive bayes untuk dijadikan pengambilan keputusan terbaik.
4. Membantu memenuhi nilai UAS dari tim ini.

1.5 Manfaat

Manfaat yang diharapkan dari penelitian adanya potensi banjir tersebut adalah :

1. Penelitian ini dapat digunakan sebagai pendeteksian jumlah sampah sebagai petunjuk untuk mendeteksi banjir.
2. Sebagai pemenuhan tugas akhir dari mata kuliah machine learning.

BAB II
METODE PELAKSANAAN

2.1 Landasan Teori

2.1.1. Banjir dan Penyebabnya

Banjir merupakan fenomena alam yang biasa terjadi di suatu kawasan yang banyak dialiri oleh aliran air. Secara sederhana banjir dapat didefinisikan sebagainya hadirnya air di suatu kawasan luas sehingga menutupi permukaan bumi kawasan tersebut. Banjir dapat disebabkan oleh faktor alam maupun faktor manusia. Kebanyakan banjir disebabkan oleh faktor manusia, yaitu salah satunya adalah kegiatan membuang sampah sembarangan yang terjadi di beberapa tempat seperti, sungai, danau, bahkan selokan.

Sampah - sampah yang dibuang lalu menyumbat saluran air dan menyebabkan air meluap lebih cepat ketika musim hujan tiba. Lalu air yang meluap tersebut dapat masuk ke pemukiman warga dan dapat menimbulkan korban jiwa maupun harta yang tidak sedikit. Karena itu, diperlukan upaya pencegahan untuk mengatasi banjir, yaitu salah satunya dengan membersihkan sampah yang menumpuk di sekitar tempat aliran air.

2.1.2 Metode Regresi Linier Sederhana

a) Pengertian

Regresi Linear Sederhana adalah Metode Statistik yang berfungsi untuk menguji sejauh mana hubungan sebab akibat antara Variabel Faktor Penyebab (X) terhadap Variabel Akibatnya. Faktor Penyebab pada umumnya dilambangkan dengan X atau disebut juga dengan Predictor sedangkan Variabel Akibat dilambangkan dengan Y atau disebut juga dengan Response.

Regresi Linear Sederhana atau sering disingkat dengan SLR (Simple Linear Regression) juga merupakan salah satu Metode Statistik yang dipergunakan dalam produksi untuk melakukan peramalan ataupun prediksi tentang karakteristik kualitas maupun Kuantitas.

b) Model Persamaan

Model Persamaan Regresi Linear Sederhana adalah seperti berikut ini :

$Y = a + bX$

Dimana :

- Y = Variabel Response atau Variabel Akibat (Dependent)
- X = Variabel Predictor atau Variabel Faktor Penyebab (Independent)
- a = konstanta
- b = koefisien regresi (kemiringan); besaran Response yang ditimbulkan oleh Predictor.

Nilai-nilai a dan b dapat dihitung dengan menggunakan Rumus dibawah ini :

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

c) Urutan langkah - langkah regresi linier

Berikut ini adalah Langkah-langkah dalam melakukan Analisis Regresi Linear Sederhana :

1) Tentukan Tujuan dari melakukan Analisis Regresi Linear Sederhana
2) Identifikasikan Variabel Faktor Penyebab (Predictor) dan Variabel Akibat (Response)
3) Lakukan Pengumpulan Data
4) Hitung X^2, Y^2, XY dan total dari masing-masingnya
5) Hitung a dan b berdasarkan rumus diatas.
6) Buatkan Model Persamaan Regresi Linear Sederhana.
7) Lakukan Prediksi atau Peramalan terhadap Variabel Faktor Penyebab atau Variabel Akibat.

2.1.3 Metode Naive Bayes

a) Pengertian

Naive Bayes adalah metode yang cocok untuk klasifikasi biner dan multiclass. Metode yang juga dikenal sebagai Naive Bayes Classifier ini menerapkan teknik supervised klasifikasi objek di masa depan dengan menetapkan label kelas ke instance/catatan menggunakan probabilitas bersyarat. Probabilitas bersyarat adalah ukuran peluang suatu peristiwa yang terjadi berdasarkan peristiwa lain yang telah (dengan asumsi, praduga, pernyataan, atau terbukti) terjadi.

b) Tipe Naive Bayes

Metode Naive Bayes digolongkan menjadi beberapa tipe berdasarkan fungsinya. Berikut ini penjelasannya.

1. Multinomial Naive Bayes

   Salah satu tipe metode Naive Bayes adalah Multinomial yang sebagian besar digunakan untuk mengklasifikasi kategori dokumen. Sebuah dokumen dapat dikategorikan bertema olahraga, politik, teknologi, atau lain-lain berdasarkan frekuensi kata-kata yang muncul dalam dokumen.

2. Bernoulli Naive Bayes

   Tipe ini mirip dengan tipe Multinomial, namun klasifikasinya lebih berfokus pada hasil ya/tidak. Prediktor yang di-input adalah variabel boolean. Misalnya, prediksi atas sebuah kata muncul dalam teks atau tidak.

3. Gaussian Naive Bayes

   Distribusi Gaussian adalah asumsi pendistribusian nilai kontinu yang terkait dengan setiap fitur berisi nilai numerik. Ketika diplot, akan muncul kurva berbentuk lonceng yang simetris tentang rata-rata nilai fitur.

c) Model Persamaan

Model Persamaan

Teorema Bayes :

$P(A \mid B) = P(B \mid A)P(A)P(B)$

Keterangan:

- $P(A \mid B)$: Probabilitas A terjadi dengan bukti bahwa B telah terjadi (probabilitas superior)
- $P(B \mid A)$: Probabilitas B terjadi dengan bukti bahwa A telah terjadi
- P(A) : Peluang terjadinya A
- P(B) : Peluang terjadinya B

d) Langkah - langkah

Klasifikasi Naive Bayes menghitung probabilitas suatu peristiwa dalam langkah-langkah berikut:

Langkah 1: Hitung probabilitas sebelumnya untuk label kelas yang diberikan.

Langkah 2: Temukan probabilitas Peluang dengan setiap atribut untuk setiap kelas.

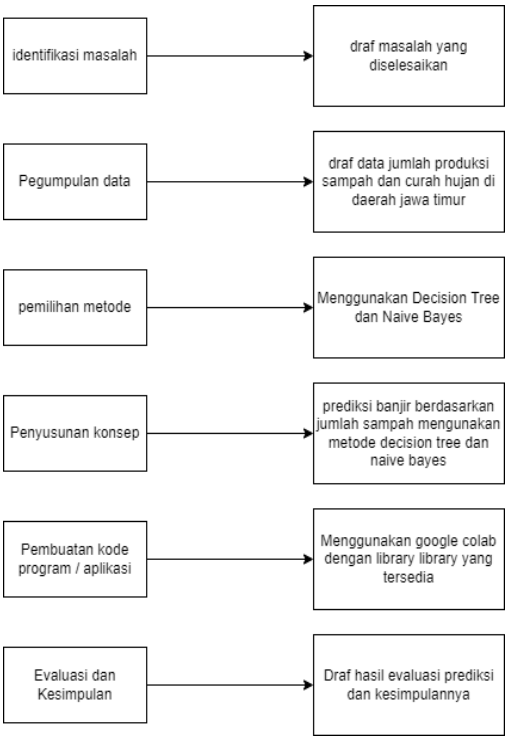
Langkah 3: Masukkan nilai ini dalam Formula Bayes dan hitung probabilitas posterior.

Langkah 4: Lihat kelas mana yang memiliki probabilitas lebih tinggi, mengingat input milik kelas probabilitas lebih tinggi.

2.2 Diagram Alur

2.2.1. Tahapan Pelaksanaan

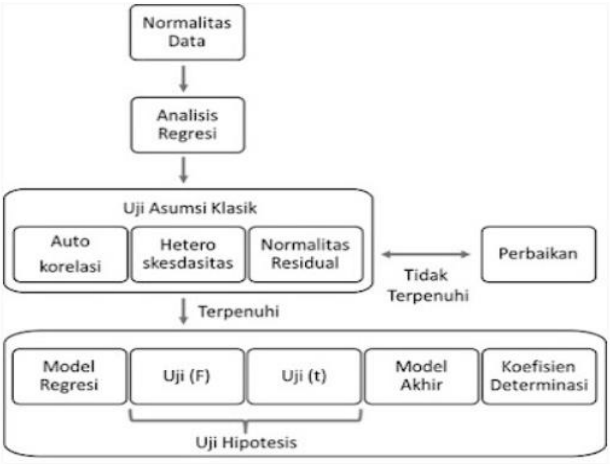
Tahapan pelaksanaan dari penelitian ini memuat lima tahapan dimulai dari identifikasi masalah hingga tahap pengujian dan evaluasi yang digambarkan pada diagram berikut ini :



Gambar 2.1. Diagram Pelaksanaan

2.2.2. Diagram Alur Regresi Linier Sederhana

Berikut merupakan alur dari metode regresi linier sederhana.



Gambar 2.2. Diagram Alur Regresi Linier

Alur dari Regresi linier sederhana ini dimulai dari memasukkan dataset dan lalu melakukan analisis regresi. Setelahnya melakukan uji asumsi klasik sebagai berikut :

1. Auto Korelasi :
   - Sebuah analisis statistik yang dilakukan untuk mengetahui adakah korelasi variabel yang ada di dalam model prediksi dengan perubahan waktu.
   - Uji autokorelasi di dalam model regresi linear, harus dilakukan apabila data merupakan data time series atau runtut waktu. Sebab yang dimaksud dengan autokorelasi sebenarnya adalah: sebuah nilai pada sampel atau observasi tertentu sangat dipengaruhi oleh nilai observasi sebelumnya.
2. Hetero Skesdasitas :
   - Uji yang menilai apakah ada ketidaksamaan varian dari residual untuk semua pengamatan pada model regresi linear. Uji ini merupakan salah satu dari uji asumsi klasik yang harus dilakukan pada regresi linear.
   - Apabila asumsi heteroskedastisitas tidak terpenuhi, maka model regresi dinyatakan tidak valid sebagai alat peramalan.
3. Normalitas Residual :

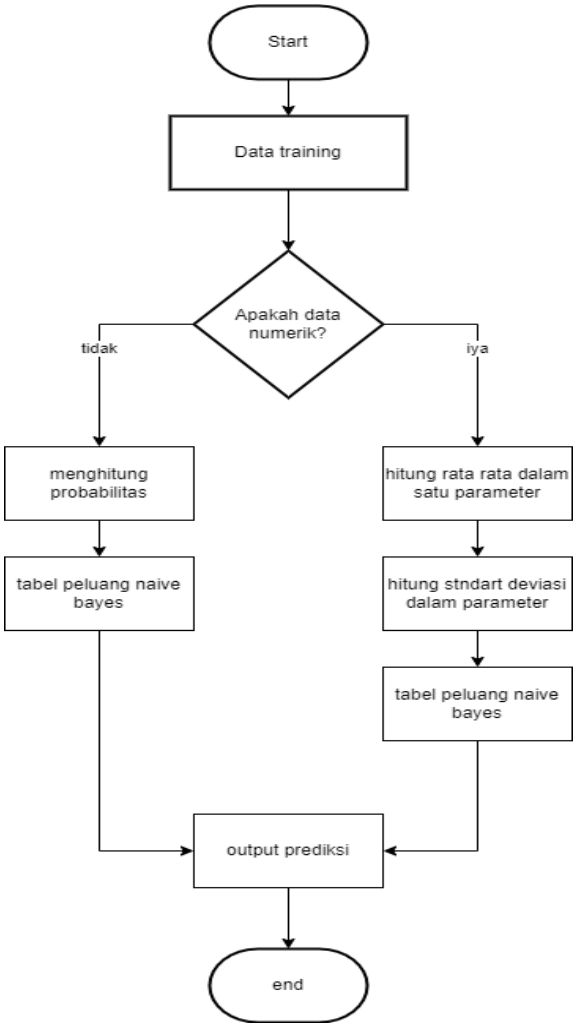
Jika uji asumsi klasik tidak terpenuhi maka data akan melalui perbaikan. Jika terpenuhi, maka dilanjutkan dengan pembuatan model regresi, dua uji hipotesis, model akhir dan koefisien determinasi.

Uji Hipotesis terdiri dari :

1. Uji (F) :
   - Uji F bertujuan untuk mencari apakah variabel independen secara bersama – sama (simultan) mempengaruhi variabel dependen.
2. Uji (T) :

- Uji t dilakukan untuk menguji hipotesis penelitian mengenai pengaruh dari masing-masing variabel bebas secara parsial terhadap variabel terikat.
- Uji T (Test T) adalah salah satu test statistik yang dipergunakan untuk menguji kebenaran atau kepalsuan hipotesis yang menyatakan bahwa diantara dua buah mean sampel yang diambil secara random dari populasi yang sama, tidak terdapat perbedaan yang signifikan

2.2.3 Diagram Alur Naive Bayes



Gambar 2.3. Diagram Alur Naive Bayes

Pertama - tama menentukan data training yang diambil dari dataset dan menentukan apakah data tersebut berbentuk numerik atau tidak. Jika tidak, maka pada data training dilakukan penghitungan probabilitas dan dibuat tabel peluangnya. Jika ya, maka pada data training menghitung rata-rata dalam satu parameter dan menghitung standar deviasi dalam parameter. Setelah itu hasilnya dimasukkan ke dalam tabel peluang.

BAB III
HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan pada penelitian ini terdiri dari 5 kolom, yaitu :

1. Kabupaten / kota di Jawa Timur
2. Tahun dengan Rentang 2017 - 2021
3. Jumlah Produksi Sampah per Tahun
4. Terjadi Banjir (dengan Ya = 1, Tidak = 0)
5. Jumlah Banjir per Tahun

Dalam dataset tersebut dapat dilihat pada tabel berikut :

Tabel 3.1. Dataset Banjir dan Sampah

nama_kabupaten_kota	tahun	jumlah_produksi_sampah_per_tahun	jumlah_banjir
Kabupaten Pacitan	2017	1511	21
Kabupaten Ponorogo	2017	419	10
Kabupaten Trenggalek	2017	100	0
Kabupaten Tulungagung	2017	1895	17
Kabupaten Blitar	2017	464	7
Kabupaten Kediri	2017	464	10
Kabupaten Malang	2017	2000	25
Kabupaten Lumajang	2017	2200	22
Kabupaten Jember	2017	465	10
Kabupaten Banyuwangi	2017	125	0
Kabupaten Bondowoso	2017	237	4
Kabupaten Situbondo	2017	161	0
Kabupaten Probolinggo	2017	487	4
Kabupaten Pasuruan	2017	308	2
Kabupaten Sidoarjo	2017	186	0
Kabupaten Mojokerto	2017	123	0
Kabupaten Jombang	2017	616	6
Kabupaten Nganjuk	2017	185	0
Kabupaten Madiun	2017	199	0
Kabupaten Magetan	2017	309	5
Kabupaten Ngawi	2017	244	0
Kabupaten Bojonegoro	2017	304	0
Kabupaten Tuban	2017	238	0
Kabupaten Lamongan	2017	1844	11
Kabupaten Gresik	2017	263	0
Kabupaten Bangkalan	2017	472	6
Kabupaten Sampang	2017	130	0
Kabupaten Pamekasan	2017	151	0
Kabupaten Sumenep	2017	219	0
Kota Kediri	2017	981	9
Kota Blitar	2017	189	0
Kota Malang	2017	264	0
Kota Probolinggo	2017	264	0
Kota Pasuruan	2017	260	0
Kota Mojokerto	2017	200	0
Kota Madiun	2017	278	10
Kota Surabaya	2017	659	7

3.2 Perhitungan Manual

3.2.1 Perhitungan Manual Regresi Linier

a. Menganalisa dan Menentukan Descriptive Statistic

Dengan memakai data analytic pada Excel, maka didapatkan *descriptive statistic* sebagai berikut ini :

Tabel 3.2. *Descriptive Statistic*

Column1		*Column2*	
Mean	701,8947368	Mean	6,357894737
Standard Error	52,03974283	Standard Error	0,59976936
Median	294	Median	0
Mode	161	Mode	0
Standard Deviation	717,3183523	Standard Deviation	8,267250094
Sample Variance	514545,6185	Sample Variance	68,34742412
Kurtosis	-0,260627401	Kurtosis	-0,046350927
Skewness	1,122438549	Skewness	1,058696171
Range	2284	Range	29
Minimum	100	Minimum	0
Maximum	2384	Maximum	29
Sum	133360	Sum	1208
Count	190	Count	190

Berdasarkan hasil analisis di atas, terdapat dua kolom yaitu column 1 yang mewakili jumlah produksi sampah per tahun dan column 2 mewakili jumlah banjir. Berdasarkan kedua hasil analisis dapat diketahui bahwa jumlah atau count data dari dua kolom adalah 190 data.

Berdasarkan data di atas dapat diketahui bahwa jumlah produksi sampah per tahun dan jumlah banjir memiliki masing-masing rata-rata 701,8947368 dan 6,357894737. Dimana rata-rata atau mean merupakan ukuran pemusatan yang sangat sering digunakan. Keuntungan dari menghitung rata-rata adalah angka tersebut dapat digunakan sebagai gambaran atau wakil dari data yang diamati.

Standar error mencerminkan keakuratan sampel yang dipilih terhadap populasinya. Semakin kecil nilai standard error, semakin mengindikasikan bahwa sampling yang diambil bagus, atau cukup mewakili populasi yang sedang diteliti. Dan sebaliknya, makanya nilai standard error akan mengecil saat jumlah sampel diperbanyak. Pada data di atas standar error dari jumlah produksi sampah per tahun dan jumlah banjir adalah 52,03974283 dan 0,59976936. Standar error jumlah produksi sampah per tahun lebih besar dari jumlah banjir menunjukkan bahwa tingkat kesalahan pada pengambilan sampel kecil.

Selanjutnya adalah nilai tengah atau sering disebut median yang merupakan suatu nilai ukuran pemusatan yang menempati posisi tengah setelah data diurutkan. Dimana ketika data sudah diurutkan dari terkecil ke terbesar akan memberikan median jika genap maka mediannya adalah data ke-n ditambah data ke-n+1 dibagi dua. Sedangkan jika data tersebut berjumlah ganjil, maka mediannya tepat di tengah. pada data di atas median dari jumlah produksi sampah dan jumlah banjir masing-masing adalah 294 dan 0.

Kemudian untuk mencari data yang paling banyak muncul dalam sampel yang ada dalam statistika disebut dengan mode. Mode merupakan nilai yang paling sering muncul dari serangkaian data. Mode tidak dapat digunakan sebagai gambaran mengenai data. Dalam data di atas, mode dari jumlah produksi sampah dan jumlah banjir masing-masing adalah 161 dan 0.

Simpangan baku (standard deviation) dinotasikan sebagai s atau σ, menunjukkan rata-rata penyimpangan data dari harga rata-ratanya. Simpangan baku merupakan akar pangkat dua dari variansi. Simpangan baku dalam data jumlah produksi sampah dan jumlah banjir berdasarkan analisis di atas adalah 171,3183523 dan 8,267250094

Variansi (variance) atau sample variance dinotasikan sebagai S2 atau σ2 adalah ukuran penyebaran data yang mengukur rata-rata kuadrat jarak seluruh titik pengamatan dari nilai tengah (meannya). Variansi ini menunjukkan keberagaman pada data yang ada. Dalam analisis data di atas variansi dari jumlah jumlah produksi sampah dan jumlah banjir adalah 514545,618490671 dan 68,3474241158452. Adanya perbedaan nilai menunjukkan kedua data tersebut memiliki keberagaman yang berbeda.

Skewness adalah derajat ketidaksimetrisan suatu distribusi. Jika kurva frekuensi suatu distribusi memiliki ekor yang lebih memanjang ke kanan (dilihat dari meannya) maka dikatakan menceng kanan (positif) dan jika sebaliknya maka menceng kiri (negatif). Secara perhitungan, skewness adalah momen ketiga terhadap mean. Distribusi normal (dan distribusi simetris lainnya, misalnya distribusi t atau Cauchy) memiliki skewness 0 (nol). Sedangkan kurtosis adalah derajat keruncingan suatu distribusi (biasanya diukur relatif terhadap distribusi normal). Kurva yang lebih lebih runcing dari distribusi normal dinamakan leptokurtik, yang lebih datar platikurtik dan distribusi normal disebut mesokurtik. Kurtosis dihitung dari momen keempat terhadap mean. Distribusi normal memiliki kurtosis = 3, sementara distribusi yang leptokurtik biasanya kurtosisnya > 3 dan platikurtik <>. Data di jumlah jumlah produksi sampah dan jumlah banjir masing-masing memiliki nilai kurtosis yaitu -0,260627400639476 dan -0,0463509269649047. sedangkan, nilai skewness dari jumlah jumlah produksi sampah dan jumlah banjir adalah 1,1224385485477 dan 1,05869617082326.

Kemudian jumlah jumlah produksi sampah dan jumlah banjir memiliki range masing-masing adalah 2284 dan 29. Rentang (Range) yang biasanya dinotasikan sebagai R, menyatakan ukuran yang menunjukkan selisih nilai antara maksimum dan minimum. Rentang cukup baik digunakan untuk mengukur penyebaran data yang simetrik dan nilai datanya menyebar merata. Kemudian kedua data tersebut memiliki nilai maksimum dan minimum. Dimana nilai maksimum merupakan nilai terbesar yang ada dalam sekumpulan data tersebut. Sedangkan nilai minimum merupakan data terkecil yang ada pada data tersebut. Berdasarkan hasil di atas nilai maksimum jumlah produksi

sampah dan jumlah banjir secara berturut-turut adalah 2384 dan 29. Sedangkan nilai minimum dari jumlah produksi sampah dan jumlah banjir yaitu 100 dan 0.

Kemudian kedua data tersebut dapat dicari jumlah data secara keseluruhan. Berdasarkan perhitungan di atas jumlah data produksi sampah adalah 133360 dan jumlah banjir adalah 1208. Jumlah ini biasanya dalam statistika deskriptif disebut sum. Selain jumlah, rata-rata juga dapat dicari dengan menggunakan rumus dimana jumlah data dibagi dengan banyaknya data.

b. Menghitung dan Menganalisis Korelasi

Dengan memakai data analytic pada Excel, maka didapatkan perhitungan korelasi sebagai berikut ini :

Tabel 3.3. Hasil Perhitungan Korelasi

	Column 1	*Column 2*
Column 1	1	
Column 2	0,90104	1

Di atas merupakan hasil analisis korelasi antara jumlah banjir dengan jumlah sampah. Pada dasarnya korelasi merupakan sebuah analisis yang berfungsi untuk mengetahui hubungan antara variabel yang satu dengan variabel yang lainnya, yang berarti ketika satu variabel terjadi variabel yang lain dapat mempengaruhinya.

Nilai korelasi (r) berkisar antara 1 sampai -1, nilai semakin mendekati 1 atau -1 berarti hubungan antara dua variabel semakin kuat, sebaliknya nilai mendekati 0 berarti hubungan antara dua variabel semakin lemah. Nilai positif menunjukkan hubungan searah (X naik maka Y naik) dan nilai negatif menunjukkan hubungan terbalik (X naik maka Y turun).

Menurut Sugiyono (2007) pedoman untuk memberikan interpretasi koefisien korelasi adalah :

1. 0,00-0,199 : sangat rendah
2. 0,20-0,399 : rendah
3. 0,40-0,599 : sedang
4. 0,60-0,799 : kuat
5. 0,80 -1,000 : sangat kuat.

Jadi berdasarkan keterangan diatas, maka dapat disimpulkan bahwa korelasi antara jumlah produksi sampah per tahun dengan jumlah banjir memiliki hubungan yang sangat kuat yaitu 0,90104. Pada hasil korelasi tersebut memiliki hasil positif. Hal itu disebut Korelasi positif yang berarti korelasi antara dua variabel dalam hal ini jumlah produksi sampah per tahun dengan jumlah banjir berjalan dengan arah yang searah atau sama. Korelasi positif terjadi jika antara dua variabel berjalan searah yang berarti jika variabel X mengalami kenaikan maka variabel Y mengalami kenaikan.

c. Menganalisa dan Menghitung Regresi

Dengan memakai data analytic pada Excel, maka didapatkan perhitungan regresi sebagai berikut ini :

1. Summary Output

Tabel 3.4. *Summary Output*

Regression Statistics	
Multiple R	0,901044413
R Square	0,811881034
Adjusted R Square	0,810880402
Standard Error	3,595252063
Observations	190

Tabel *Summary output* ini melaporkan kekuatan hubungan antara model (variabel bebas) dengan variabel terikat.

Pada *regression statistics*, *multiple R* (R majemuk) adalah suatu ukuran untuk mengukur tingkat (keeratan) hubungan linear antara variabel terikat dengan seluruh variabel bebas secara bersama-sama. Pada kasus dua variabel (satu variabel terikat dan satu variabel bebas), besaran r (biasa dituliskan dengan huruf kecil untuk dua variabel) dapat bernilai positif maupun negatif (antara -1 – 1), tetapi untuk lebih dari dua variabel, besaran R selalu bernilai positif (antara 0 – 1). Nilai R yang lebih besar (+ atau -) menunjukkan hubungan yang lebih kuat. Pada hasil analisis dua variabel di atas bahwa nilai R adalah 0. 90 yang menunjukkan bahwa kedua hubungan variabel tersebut sangat kuat.

R Square (R^2) sering disebut dengan koefisien determinasi, adalah mengukur kebaikan suai *(goodness of fit)* dari persamaan regresi yaitu memberikan proporsi atau persentase variasi total dalam variabel terikat yang dijelaskan oleh variabel bebas. Nilai R^2 terletak antara 0% sampai 100%, dan kecocokan model dikatakan lebih baik kalau R^2 semakin mendekati 100%. Berdasarkan hasil di atas nilai dari koefisien determinasinya adalah 0,81 atau 81%. Artinya 81% keragaman y mampu dijelaskan oleh x dalam model 81%, sedangkan sisanya dijelaskan oleh peubah lain yang diluar model.

Adjusted R Square merupakan suatu sifat penting R^2 adalah nilainya merupakan fungsi yang tidak pernah menurun dari banyaknya variabel bebas yang ada dalam model. Oleh karenanya, untuk membandingkan dua R^2 dari dua model, harus memperhitungkan banyaknya variabel bebas yang ada dalam model. Hal ini dapat dilakukan dengan menggunakan *adjusted R square*. Istilah penyesuaian berarti nilai R^2 sudah disesuaikan dengan banyaknya variabel (derajat bebas) dalam model. Pada dasarnya R^2 yang disesuaikan ini juga akan meningkat bersamaan meningkatnya jumlah variabel, tetapi peningkatannya relatif kecil. Pada hasil analisis di atas, nilai dari *adjusted R square* cukup besar yaitu 0,810.

Standard Error merupakan standar error dari estimasi variabel terikat (dalam kasus ini adalah jumlah penduduk miskin). Angka ini dibandingkan

dengan standar deviasi dari jumlah penduduk miskin. Semakin kecil angka *standar error* ini dibandingkan angka standar deviasi dari jumlah penduduk miskin maka model regresi semakin tepat dalam memprediksi jumlah penduduk miskin. Nilai *standar error* regresi adalah 3,5

2. Tabel ANOVA (Analysis of Variance)

Tabel 3.5. Tabel ANOVA

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10487,60573	10487,60573	811,3676047	4,03341E-70
Residual	188	2430,057431	12,9258374		
Total	189	12917,66316			

Tabel ANOVA (Analysis of Variance) menguji penerimaan (acceptability) model dari perspektif statistik dalam bentuk analisis sumber keragaman. ANOVA ini sering juga diterjemahkan sebagai analisis ragam.

Dari tabel ANOVA tersebut diungkapkan bahwa keragaman data aktual variabel terikat (jumlah hujan) bersumber dari model regresi dan dari residual. Dalam pengertian sederhana untuk kasus ini adalah variasi (turun-naiknya atau besar kecilnya) jumlah banjir disebabkan oleh variasi dari jumlah produksi sampah per tahun (model regresi) serta dari faktor-faktor lainnya yang mempengaruhi jumlah banjir yang tidak dimasukkan dalam model regresi (residual).

Degree of Freedom (df) atau derajat bebas dari total adalah n-1, dimana n adalah banyaknya observasi. Dalam hal ini banyaknya observasi adalah 190 maka derajat bebas total adalah 189. Derajat bebas dari model regresi adalah 1, karena ada satu variabel bebas dalam model ini (luas daerah). Derajat bebas untuk residual adalah sisanya yaitu derajat bebas total – derajat bebas regresi = 190 – 1 = 189.

Kolom SS (Sum of Square) atau jumlah kuadrat untuk regression diperoleh dari penjumlahan kuadrat dari prediksi variabel terikat (Jumlah penduduk miskin) dikurangi dengan nilai rata-rata jumlah banjir dari data sebenarnya. Jadi secara manual mencari terlebih dahulu rata-rata permintaan dari data asli. Kemudian masing-masing prediksi jumlah banjir dikurangi dengan rata-rata tersebut kemudian dikuadratkan. Selanjutnya, seluruh hasil perhitungan tersebut dijumlahkan. Pada data di atas nilai dari SS regression adalah 10487.60573

Kolom SS untuk residual diperoleh dari jumlah pengkuadratan dari residual. Pada hasil output di atas dapat diketahui bahwa nilai SS untuk residual adalah 2430.057431. Kolom SS untuk total adalah penjumlahan dari SS untuk regresi dengan dengan SS untuk residual. Sebenarnya SS total ini adalah variasi (besar-kecil,naik-turun) dari jumlah banjir. Hal ini diukur dengan mengurangi nilai masing-masing permintaan aktual dengan rata-ratanya, kemudian dikuadratkan. Hasil perhitungan tersebut kemudian dijumlahkan. Berdasarkan output di atas maka nilai SS total adalah 12917.66316

Hasil ketika SS tersebut memiliki arti dimana apabila SS total yang diperoleh adalah adalah 12917.66316 yang memiliki arti, variasi dari jumlah penduduk miskin yang dikuadratkan adalah sebesar nilai tersebut. Bervariasinya jumlah banjir disebabkan oleh sebagian berasal dari variabel bebas (jumlah produksi sampah) yaitu sebesar 10487.60573 (regresi). Kemudian sisanya sebesar 2430.057431 disebabkan oleh variabel lain yang juga mempengaruhi jumlah banjir tetapi tidak dimasukkan dalam model (residual).

Jika membandingkan (bagi) antara SS regresi dengan SS total, maka akan didapatkan proporsi dari total variasi jumlah banjir yang disebabkan oleh variasi jumlah produksi sampah. Praktikan mencoba membagi antara nilai SS regresi dengan SS total yaitu 10487.60573 / 12917.66316= 0,8118810345260621. hasil tersebut sama dengan hasil dari R2 atau koefisien determinasi yang telah dibahas di atas.

Selanjutnya kolom berikutnya dari ANOVA adalah kolom MS (Mean of Square) atau rata-rata jumlah kuadrat. Ini adalah hasil bagi antara kolom SS dengan kolom df. Berdasarkan hasil di atas besarnya MS adalah regresi dan residual berturut-turut adalah 10487.60573 dan 12.9258374.

Dari perhitungan MS ini, selanjutnya dengan membagi antara MS Regresi dengan MS Residual didapatkan nilai F. Nilai F ini yang dikenal dengan F hitung dalam pengujian hipotesis dibandingkan dengan nilai F tabel. Jika F hitung > F tabel, maka dapat dinyatakan bahwa secara simultan (bersama-sama) luas daerah berpengaruh signifikan terhadap jumlah penduduk miskin. Setelah melakukan pembagian antara MS regresi dengan MS residual didapatkan hasil nilai F hitung yaitu 8,113676047015724 seperti pada hasil analisis. Selain itu, dapat juga membandingkan antara taraf nyata dengan p-value (dalam istilah Excel adalah Significance F). Jika taraf nyata > dari p-value maka kesimpulannya sama dengan di atas 3 Tabel Analisis Intercept dan X variabel

Tabel 3.6. Hasil Analisis *Intercept* dan X Variabel

	Coefficients	*Standard Error*	*t Stat*
Intercept	-0,931094645	0,365392905	-2,548201221
X Variable 1	0,010384733	0,000364575	28,48451517

Tabel 3.7. Hasil Analisis *Intercept* dan X Variabel

P-value	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
0,011627491	-1,651891596	-0,210297695	-1,651891596	-0,210297695
4,03341E-70	0,00966555	0,011103916	0,00966555	0,011103916

Tabel berikutnya dari output Excel seperti tabel di atas menampilkan nilai-nilai koefisien, standard error, t Sat, P-value dan selang kepercayaan. Dalam pengujian hipotesis regresi, tahap berikutnya setelah pengujian secara simultan (uji F seperti yang telah disampaikan sebelumnya) adalah pengujian koefisien regresi secara parsial. Pengertian pengujian secara parsial ini dalam kasus yang dibahas praktikan adalah apakah jumlah produksi sampah mempengaruhi jumlah banjir.

Dalam uji parsial, digunakan uji t, yaitu membandingkan antara t-hitung (t Stat) dengan t tabel. Jika t hitung > t tabel pada taraf nyata tertentu, maka dapat disimpulkan variabel tersebut berpengaruh secara signifikan. T hitung ditampilkan pada kolom 4, yang merupakan hasil bagi antara kolom 2 (coefficients) dengan kolom 3 (Standard Error).

Selain membandingkan dengan nilai t-tabel, dapat juga ditarik kesimpulan signifikansinya dengan membandingkan taraf nyata dengan p-value (kolom 5). Jika dimisalkan dengan menggunakan taraf nyata 5 %, maka variabel dengan p-value sama atau lebih kecil dari 5 %, dapat dinyatakan sebagai variabel yang secara parsial berpengaruh signifikan.

Berdasarkan hal tersebut, terlihat bahwa jumlah produksi sampah (x) dengan nilai p-value sebesar 4.0 lebih besar dari 0.05 yang berarti bahwa jumlah produksi sampah tidak terlalu berpengaruh pada jumlah banjir.

Selanjutnya, kolom 6 dan 7 memberikan selang kepercayaan untuk koefisien. Pada hasil output judulnya tertulis Lower 95% dan Upper 95% dimana angka 95% adalah penetapan pada waktu pengolahan dengan Excel dan bisa dirubah sesuai keinginan arti dari selang kepercayaan tersebut adalah nilai koefisien yang diberikan pada output regresi merupakan dugaan titik (point estimate) dari parameter koefisien regresi.

Tingkat kepercayaan sebesar 95% dengan tingkat kesalahan sebesar 5% memberikan dugaan selang (confidence interval) kepercayaan sebesar -1,65189159631736 - -0,21029769463983, dimana nilai parameter sebenarnya diharapkan berada dalam selang tersebut dengan tingkat kepercayaan tertentu.

Berdasarkan hal tersebut, dari output Excel terlihat bahwa dengan tingkat kepercayaan 95%, maka koefisien regresi untuk x atau luas daerah adalah 0,0103847329232224. selanjutnya adalah membuat persamaan regresi atau model regresi dari permasalahan diatas.

Tabel 3.8. Model Regresi

Jumlah Banjir (y) = -0,931094645478596 x 0,0103847329232224**x**

Berdasarkan model regresi di atas bahwa konstanta yang sebesar -0,931094645478596 secara matematis berarti bahwa ketika variabel bebas nilainya 0, maka variabel terikat nilainya adalah sebesar konstanta tersebut.

3.2.2 Perhitungan Manual Naive Bayes

a. Menentukan jumlah data banjir dan jumlah data tidak banjir

Data di filter untuk menunjukkan data daerah yang terdampak banjir atau yang tidak terdampak banjir pada kolom data “Terjadi Banjir(1=Ya, 0=Tidak)”

Tabel 3.9. Data daerah yang banjir pada rentang waktu 2017-2021

	A	B	C	D	E
1	nama_kabupaten_kota	tahun	jumlah_produksi_sampah_per_tahun	Terjadi Banjir (1=Ya, 0=Tidak)	jumlah_banjir
2	Kabupaten Pacitan	2017	1511	1	21
3	Kabupaten Ponorogo	2017	419	1	10
5	Kabupaten Tulungagung	2017	1895	1	17
6	Kabupaten Blitar	2017	464	1	7
7	Kabupaten Kediri	2017	464	1	10
8	Kabupaten Malang	2017	2000	1	25
9	Kabupaten Lumajang	2017	2200	1	22
10	Kabupaten Jember	2017	465	1	10
12	Kabupaten Bondowoso	2017	237	1	4
14	Kabupaten Probolinggo	2017	487	1	4
15	Kabupaten Pasuruan	2017	308	1	2
18	Kabupaten Jombang	2017	616	1	6
21	Kabupaten Magetan	2017	309	1	5
25	Kabupaten Lamongan	2017	1844	1	11
27	Kabupaten Bangkalan	2017	472	1	6
31	Kota Kediri	2017	981	1	9
37	Kota Madiun	2017	278	1	10
38	Kota Surabaya	2017	659	1	7
42	Kabupaten Trenggalek	2018	2001	1	22
44	Kabupaten Blitar	2018	351	1	1
48	Kabupaten Jember	2018	2015	1	25
52	Kabupaten Probolinggo	2018	1377	1	11
55	Kabupaten Mojokerto	2018	1063	1	13

Tabel 3.10. Data daerah yang tidak banjir pada rentang waktu 2017-2021

	A	B	C	D	E
1	nama_kabupaten_kota	tahun	jumlah_produksi_sampah_per_tahun	Terjadi Banjir (1=Ya, 0=Tidak)	jumlah_banjir
4	Kabupaten Trenggalek	2017	100	0	0
11	Kabupaten Banyuwangi	2017	125	0	0
13	Kabupaten Situbondo	2017	161	0	0
16	Kabupaten Sidoarjo	2017	186	0	0
17	Kabupaten Mojokerto	2017	123	0	0
19	Kabupaten Nganjuk	2017	185	0	0
20	Kabupaten Madiun	2017	199	0	0
22	Kabupaten Ngawi	2017	244	0	0
23	Kabupaten Bojonegoro	2017	304	0	0
24	Kabupaten Tuban	2017	238	0	0
26	Kabupaten Gresik	2017	263	0	0
28	Kabupaten Sampang	2017	130	0	0
29	Kabupaten Pamekasan	2017	151	0	0
30	Kabupaten Sumenep	2017	219	0	0
32	Kota Blitar	2017	189	0	0
33	Kota Malang	2017	264	0	0
34	Kota Probolinggo	2017	264	0	0
35	Kota Pasuruan	2017	260	0	0
36	Kota Mojokerto	2017	200	0	0
39	Kota Batu	2017	159	0	0

b. Menentukan total keseluruhan data banjir

Tabel 3.11. Total Data Banjir

	Ya	Tidak			
	1	0			
Banyak Data	89	101			
Harus Sama					
Jumlah banyak data	190	=	Total	190	

- Rumus mencari data yang berdata banjir kondisi “Iya” : =COUNTIF(D2:D191,H2), yang mana menghasilkan nilai 89 data.

- Rumus mencari data yang berdata banjir kondisi “Tidak” : =COUNTIF(D2:D191,I2), yang mana menghasilkan nilai 101 data.
- Rumus total keseluruhan dengan cara menggunakan formula ‘SUM’ untuk data pada kolom 89 dan 101 serta mencocokkannya dengan *real* data yang sebenarnya pada keseluruhan tabel data. Maka hasilnya adalah terdapat 190 data (banjir dan tidak).

c. Menentukan probabilitas banjir dengan kondisi iya

Tabel 3.12. Probabilitas Terjadi Banjir

Probabilitas Terjadi Banjir			
Terjadi Banjir		Nilai	
Ya	1	0.468421053	
Tidak	0	0.531578947	
		1	(sudah sesuai)

- Rumus mencari nilai probabilitas banjir kondisi “Iya” : =COUNTIF(D2:D191,H10)/COUNTA(D2:D191), dimana dalam menghitung banyaknya data yang berkondisi banjir “Iya” kemudian dibagi dengan banyak data keseluruhan.
- Rumus mencari nilai probabilitas banjir dengan kondisi “Tidak” : =COUNTIF(D2:D191,H11)/COUNTA(D2:D191), dimana dalam menghitung banyaknya data yang berkondisi banjir “Tidak” kemudian dibagi dengan banyak data keseluruhan.

d. Menghitung mean atau rata-rata banjir dengan kondisi “iya”

- Menentukan jumlah dari total sampah yang berdampak banjir berkondisi “Iya”.

Tabel 3.13. Jumlah Keseluruhan Sampah pada Kondisi Banjir “Iya”

193	Sampah (Ya)	
194	No	Sampah
195	1	1511
196	2	419
197	3	1895
198	4	464
199	5	464
200	6	2000
201	7	2200
202	8	465
203	9	237
204	10	487
275	81	1416
276	82	1610
277	83	712
278	84	2191
279	85	803
280	86	148
281	87	1700
282	88	2197
283	89	1596
284	Σ	112746

- Menghitung mean atau rata-rata dari data sampah berstatus iya pada munculnya banjir.

Tabel 3.14. Mean Banjir dengan Kondisi "Iya"

Menghitung Mean	
[μC]_1, Banjir	jumlah sampah(iya)/ banyak data sampah(iya)
$\mu = \frac{1}{n}\sum_{i=1}^{n} x_1$	=B284/COUNTA(B195:B283)
	mengasilkan :
	1266.808989

e. Menghitung standar deviasi banjir dengan kondisi “iya”

- Menghitung sampah-√sampah untuk nantinya dijadikan perhitungan dalam standar deviasi.

Tabel 3.15. Menghitung Standar Deviasi (1)

C195 | =B195-I$198

	A	B	C
193	Sampah (Ya)		
194	No	Sampah	Sampah-√Sampah
195	1	1511	244.1910112
196	2	419	-847.8089888
197	3	1895	628.1910112
198	4	464	-802.8089888
199	5	464	-802.8089888
200	6	2000	733.1910112
201	7	2200	933.1910112
202	8	465	-801.8089888
203	9	237	-1029.808989
204	10	487	-779.8089888
205	11	308	-958.8089888
206	12	616	-650.8089888
207	13	309	-957.8089888
208	14	1844	577.1910112

- Kemudian hitungan tadi di pangkat 2, dengan formula “Power” pada excel

Tabel 3.16. Menghitung Standar Deviasi Kondisi Banjir “Iya” (2)

D195 | =POWER(C195,2)

	A	B	C	D
193	Sampah (Ya)			
194	No	Sampah	Sampah-√Sampah	(Sampah-√Sampah)^2
195	1	1511	244.1910112	59629.24997
196	2	419	-847.8089888	718780.0814
197	3	1895	628.1910112	394623.9466
198	4	464	-802.8089888	644502.2724
199	5	464	-802.8089888	644502.2724
200	6	2000	733.1910112	537569.059
201	7	2200	933.1910112	870845.4635
202	8	465	-801.8089888	642897.6545
203	9	237	-1029.808989	1060506.553
204	10	487	-779.8089888	608102.059
205	11	308	-958.8089888	919314.6769
206	12	616	-650.8089888	423552.3399
207	13	309	-957.8089888	917398.059
208	14	1844	577.1910112	333149.4635
209	15	472	-794.8089888	631721.3286

- Lalu hitung jumlah dari data sampah yang telah dilakukan perpangkatan tadi.

Tabel 3.17. Menghitung Standar Deviasi Kondisi Banjir “Iya” (3)

279	85	803	-463.8089888	215118.7781
280	86	148	-1118.808989	1251733.553
281	87	1700	433.1910112	187654.4522
282	88	2197	930.1910112	865255.3174
283	89	1596	329.1910112	108366.7219
284	**Σ**	**112746**		**43357433.75**

- Menghitung standar deviasi keseluruhan data sampah dengan kondisi banjir (Iya)

Tabel 3.18. Menghitung Standar Deviasi Kondisi Banjir “Iya” (4)

Menghitung Standar Deviasi	
[σC]_1, Banjir	akar dari hasil data sampah yang telah data tersebut
	dikuadratkan untuk kemudian dibagi dengan banyak
$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1 - \mu)^2}{n-1}}$	=SQRT(D284/(COUNTA(D195:D283)-1))
	menghasilkan :
	701.9245763

f. Menghitung mean atau rata-rata banjir dengan kondisi “tidak”

- Menentukan jumlah dari total sampah yang berdampak banjir berkondisi “Tidak”

Tabel 3.19. Menghitung Standar Deviasi Kondisi Banjir “Tidak” (1)

Sampah (Tidak)	
No	**Sampah**
1	100
2	125
3	161
4	186
5	123
6	185
7	199
8	244
9	304
10	238
11	263
12	130
95	161
96	166
97	131
98	227
99	357
100	152
101	180
Σ	**20614**

- Menghitung mean atau rata-rata dari data sampah berstatus tidak pada munculnya banjir

Tabel 3.20. Menghitung Mean Banjir dengan Kondisi “Tidak”

Menghitung Mean	
[μC]_1, Banjir	jumlah sampah(iya)/ banyak data sampah(iya)
	=M296/COUNTA(M195:M295)
$\mu = \frac{1}{n}\sum_{i=1}^{n} x_1$	mengasilkan :
	204.099

g. Menghitung standar deviasi banjir dengan kondisi "tidak"

- Menghitung sampah-√sampah untuk nantinya dijadikan perhitungan dalam standar deviasi

Tabel 3.21. Menghitung Standar Deviasi Kondisi Banjir "Tidak" (2)

N195 =M195-T$198

	L	M	N
193	Sampah (Ya)		
194	No	Sampah	Sampah-√Sampah
195	1	100	-104.0990099
196	2	125	-79.0990099
197	3	161	-43.0990099
198	4	186	-18.0990099
199	5	123	-81.0990099
200	6	185	-19.0990099
201	7	199	-5.099009901
202	8	244	39.9009901
203	9	304	99.9009901
204	10	238	33.9009901
205	11	263	58.9009901

- Kemudian hitungan tadi di pangkat 2, dengan formula "Power" pada excel

Tabel 3.22. Menghitung Standar Deviasi Kondisi Banjir "Tidak" (3)

O195 =POWER(N195,2)

	L	M	N	O
193	Sampah (Ya)			
194	No	Sampah	Sampah-√Sampah	(Sampah-√Sampah)^2
195	1	100	-104.0990099	10836.60386
196	2	125	-79.0990099	6256.653367
197	3	161	-43.0990099	1857.524654
198	4	186	-18.0990099	327.5741594
199	5	123	-81.0990099	6577.049407
200	6	185	-19.0990099	364.7721792
201	7	199	-5.099009901	25.99990197
202	8	244	39.9009901	1592.089011
203	9	304	99.9009901	9980.207823
204	10	238	33.9009901	1149.27713
205	11	263	58.9009901	3469.326635
206	12	130	-74.0990099	5490.663268
207	13	151	-53.0990099	2819.504852
208	14	219	14.9009901	222.0395059

- Lalu hitung jumlah dari data sampah yang telah dilakukan perpangkatan tadi.

Tabel 3.23. Menghitung Standar Deviasi Kondisi Banjir "Tidak" (4)

288	94	358	153.9009901	23685.51475
289	95	161	-43.0990099	1857.524654
290	96	166	-38.0990099	1451.534555
291	97	131	-73.0990099	5343.465249
292	98	227	22.9009901	524.4553475
293	99	357	152.9009901	23378.71277
294	100	152	-52.0990099	2714.306833
295	101	180	-24.0990099	580.7622782
296	**Σ**	**20614**		**461427.0099**

- Menghitung standar deviasi keseluruhan data sampah dengan kondisi banjir (Tidak)

Tabel 3.24. Menghitung Standar Deviasi Kondisi Banjir "Tidak" (5)

Menghitung Standar Deviasi	
[σC]_1, Banjir	akar dari hasil data sampah yang telah data tersebut dikuadratkan untuk kemudian dibagi dengan banyak =SQRT(D284/(COUNTA(D195:D283)-1)) menghasilkan : 67.92842

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1 - \mu)^2}{n - 1}}$$

3.3 Implementasi Kode Program

3.3.1 Kode Program Regresi Linier

1. Mempersiapkan Data

```
[ ] # import package
    import numpy as np
    import pandas as pd
```

```
[ ] # baca data
    data = pd.read_csv('data tubes.csv')
    data.head()
```

	nama_kabupaten_kota	tahun	jumlah_produksi_sampah_per_tahun	Terjadi Banjir (1=Ya, 0=Tidak)	jumlah_banjir
0	Kabupaten Pacitan	2017	1511	1	21
1	Kabupaten Ponorogo	2017	419	1	10
2	Kabupaten Trenggalek	2017	100	0	0
3	Kabupaten Tulungagung	2017	1895	1	17
4	Kabupaten Blitar	2017	464	1	7

Gambar 3.1. Persiapan Data

Melakukan import library dan dataset “data tubes.CSV”. Lalu cek data dengan menampilkan urutan teratas dataset.

2. Memahami Data

```
[ ] # pemahaman terhadap data
    # ukuran data
    data.shape

    # info data
    data.info()

    # deskripsi data
    data.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 190 entries, 0 to 189
Data columns (total 5 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   nama_kabupaten_kota              190 non-null    object
 1   tahun                            190 non-null    int64
 2   jumlah_produksi_sampah_per_tahun 190 non-null    int64
 3   Terjadi Banjir (1=Ya, 0=Tidak)   190 non-null    int64
 4   jumlah_banjir                    190 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.5+ KB
```

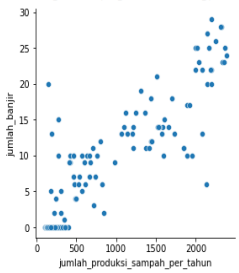
	tahun	jumlah_produksi_sampah_per_tahun	Terjadi Banjir (1=Ya, 0=Tidak)	jumlah_banjir
count	190.00000	190.000000	190.000000	190.000000
mean	2019.00000	701.894737	0.468421	6.357895
std	1.41795	717.318352	0.500320	8.267250
min	2017.00000	100.000000	0.000000	0.000000
25%	2018.00000	186.750000	0.000000	0.000000
50%	2019.00000	294.000000	0.000000	0.000000
75%	2020.00000	1213.750000	1.000000	11.750000
max	2021.00000	2384.000000	1.000000	29.000000

Gambar 3.2. Check Dataset

Memahami dataset yang tersedia dengan mengecheck ukuran data, info data dan deskripsi data.

3. Melakukan Visualisasi Data

```
[ ] # import library untuk visualisasi
    import matplotlib.pyplot as plt
    import seaborn as sns

[ ] # operasi visualisasi
    sns.pairplot(data, x_vars=['jumlah_produksi_sampah_per_tahun'],
                 y_vars='jumlah_banjir', size=4, aspect=1, kind='scatter')
    plt.show()

/usr/local/lib/python3.8/dist-packages/seaborn/axisgrid.py:2076: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

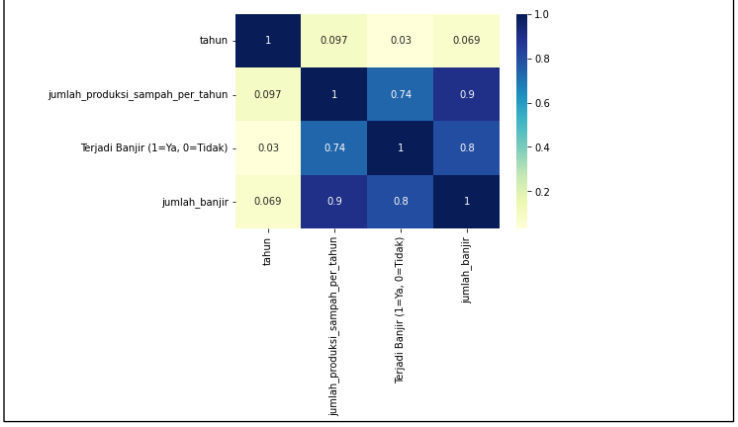


Gambar 3.3. Visualisasi Data dengan Scatterplot

Melakukan visualisasi data dengan mengimport library visualisasi dan menentukan x dan y untuk visualisasi dengan scatterplot

4. Visualisasi Data Dengan Heatmap

```
[ ] sns.heatmap(data.corr(), cmap="YlGnBu", annot = True)
    plt.show()
```



Gambar 3.4. Visualisasi Data dengan Heatmap

Pada visualisasi heatmap tersebut, jumlah banjir berkorelasi kuat dengan jumlah produksi sampah per tahun dengan angka 0,9.

5. Tahapan Regresi Linier

a) Membuat variabel x dan y

```
[ ] # Buat variabel bebas X dan Y, sebagai contoh ambil dari hasil analisis korelasi dari kegaitan sebelumnya
    X = data['jumlah_produksi_sampah_per_tahun']
    y = data['jumlah_banjir']
```

Gambar 3.5. Pembuatan Variabel x dan y

Membuat variabel x dan y berdasarkan analisis korelasi dari kegiatan sebelumnya yaitu :

x : kolom jumlah produksi sampah per tahun

y : kolom jumlah banjir

b) Pemisahan Data Testing dengan Data Training

```
[ ] # Buat pemisahan data uji dan data latih dengan proporsi 7:3
    from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.7,
                                                        test_size = 0.3, random_state = 100)
```

Gambar 3.6. Pemisahan Data Test dan Data Train

Memisahkan data train dan data test dengan proporsi data train : 0,7 dan data test : 0,3.

c) Hasil Training Data

```
[ ] # hasil training dtaset
    X_train
    y_train

    165    19
    65      0
    74      0
    89      0
    81      0
           ..
    87     10
    103     0
    67      0
    24      0
    8      10
    Name: jumlah_banjir, Length: 133, dtype: int64
```

Gambar 3.7. Hasil Training

Berikut merupakan hasil training dataset dengan x train dan y train.

d) Training Model

```
[ ] # training model
    import statsmodels.api as sm

    X_train_sm = sm.add_constant(X_train)

    /usr/local/lib/python3.8/dist-packages/statsmodels/tsa/tsatools.py:142: FutureWarning: I
      x = pd.concat(x[::order], 1)
```

Gambar 3.8. Training Model

Mengimport Library yang diperlukan dan membuat training model untuk variabel x.

e) Fitting Garis Regresi

```
[ ] # fitting garis regresi
    lr = sm.OLS(y_train, X_train_sm).fit()
    lr.params

    const                            -0.787503
    jumlah_produksi_sampah_per_tahun    0.010484
    dtype: float64
```

Gambar 3.9. Fitting Garis Regresi

Tampil jumlah const dan jumlah produksi sampah per tahun dalam dtype float64.

f) Menganalisis Garis Regresi

```
[ ] # analisis statistika dari garis regresi
    lr.summary()
```

OLS Regression Results

Dep. Variable:	jumlah_banjir	**R-squared:**	0.791
Model:	OLS	**Adj. R-squared:**	0.789
Method:	Least Squares	**F-statistic:**	494.9
Date:	Sun, 18 Dec 2022	**Prob (F-statistic):**	2.53e-46
Time:	06:35:18	**Log-Likelihood:**	-362.23
No. Observations:	133	**AIC:**	728.5
Df Residuals:	131	**BIC:**	734.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7875	0.446	-1.764	0.080	-1.671	0.096
jumlah_produksi_sampah_per_tahun	0.0105	0.000	22.246	0.000	0.010	0.011

Omnibus:	61.461	**Durbin-Watson:**	2.207
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	269.590
Skew:	1.603	**Prob(JB):**	2.88e-59
Kurtosis:	9.195	**Cond. No.**	1.31e+03

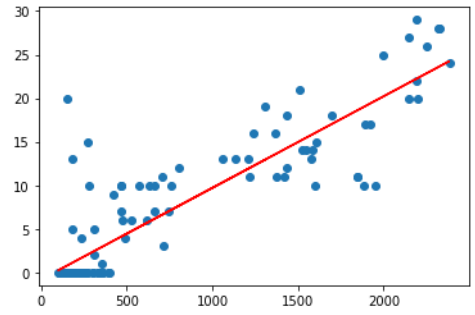
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.31e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Gambar 3.10. Analisis Garis Regresi

Melakukan analisis garis regresi yang terdiri atas beberapa informasi yang muncul seperti R-squared, dll.

g) Melakukan Visualisasi Garis Regresi

```
[ ] # visualisasi garis regresi
    plt.scatter(X_train, y_train)
    plt.plot(X_train, -0.7875 + 0.0105*X_train, 'r')
    plt.show()
```



Gambar 3.11. Visualisasi Garis Regresi

Visualisasi data regresi dengan memasukkan jumlah const dan jumlah produksi sampah per tahun dalam perhitungannya.

6. Tahapan Residual Analysis

a) Prediksi y Value Dari Data x yang Dilatih

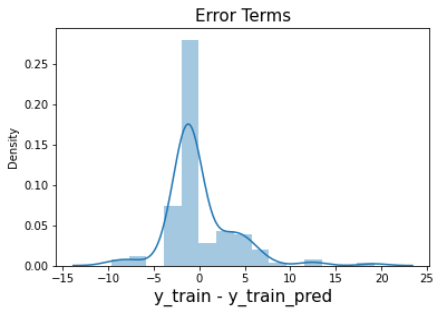
```
[ ] # prediksi y_value dari data x yang telah dilatih
    y_train_pred = lr.predict(X_train_sm)

    res = (y_train - y_train_pred)
```

Gambar 3.12. Prediksi y Value

Prediksi dari y value yang memiliki hubungan dengan data x yang telah dilatih

b) Distribusi Histogram

```
[ ]  # cek histogram apakah berdistribusi normal atau tidak
     fig = plt.figure()
     sns.distplot(res, bins = 15)
     plt.title('Error Terms', fontsize = 15)
     plt.xlabel('y_train - y_train_pred', fontsize = 15)
     plt.show()

     /usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: FutureWa
       warnings.warn(msg, FutureWarning)
```

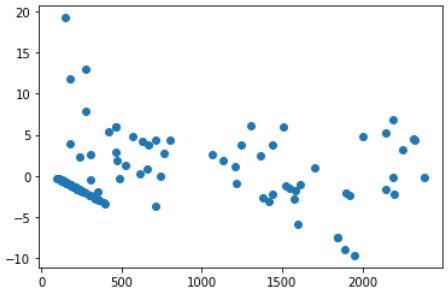


Gambar 3.13. Histogram

Mengecheck histogram apakah terdistribusi dengan normal atau tidak. Jika puncak yang muncul hanya satu puncak, maka data berdistribusi dengan normal.

c) Distribusi dengan scatterplot

```
[ ]  plt.scatter(X_train,res)
     plt.show()
```



Gambar 3.14. Distribusi Scatterplot

Menunjukkan distribusi dalam bentuk visualisasi scatterplot.

d) Prediksi Pada Data Uji dan Prediksi y Value

```
[ ]  # prediksi pada data uji dan evaluasi model
     X_test_sm = sm.add_constant(X_test)

     # prediksi y value yang berkorelasi dengan X_test_sm
     y_test_pred = lr.predict(X_test_sm)

     # cetak 5 data terprediksi teratas
     y_test_pred.head()

     /usr/local/lib/python3.8/dist-packages/statsmodels/tsa/tsatools.py:142
       x = pd.concat(x[::order], 1)
     161    1.529351
     77     1.445483
     11     0.900340
     158    1.959174
     164    0.858406
     dtype: float64
```

Gambar 3.15. Prediksi Pada Data Uji dan Prediksi y Value

e) Menghitung Nilai r^2

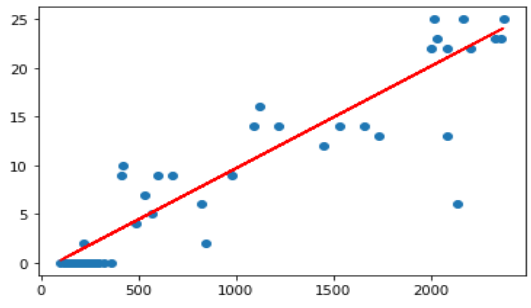
```
[ ]  # hitung nilai r^2
     from sklearn.metrics import r2_score

     r_squared = r2_score(y_test, y_test_pred)
     r_squared

     0.8512563266808444
```

Gambar 3.16. Nilai r^2

Menghitung nilai r^2 yang hasilnya 0,85. Jika dibandingkan dengan data training yang hasilnya 0,79. Hal tersebut menunjukkan bahwa data tidak berselisih jauh dan data merupakan data yang stabil.

f) Melakukan Visualisasi Data r

```
[ ]  # visualisasi data
     plt.scatter(X_test, y_test)
     plt.plot(X_test, y_test_pred, 'r')
     plt.show()
```



Gambar 3.17. visualisasi Data r

Menunjukkan data r dalam bentuk visualisasi scatterplot.

3.3.2 Kode Program Naive Bayes

1. Mempersiapkan Data

```
import numpy as np
import pandas as pd

# Load data CSV
df = pd.read_csv('data tubes.csv')

# Cek data
display(df.head())
```

	nama_kabupaten_kota	tahun	jumlah_produksi_sampah_per_tahun	Terjadi Banjir (1=Ya, 0=Tidak)	jumlah_banjir
0	Kabupaten Pacitan	2017	1511	1	21
1	Kabupaten Ponorogo	2017	419	1	10
2	Kabupaten Trenggalek	2017	100	0	0
3	Kabupaten Tulungagung	2017	1895	1	17
4	Kabupaten Blitar	2017	464	1	7

Gambar 3.18. Persiapan Data

Melakukan import library dan dataset “data tubes.CSV”. Lalu cek data dengan menampilkan urutan teratas dataset.

2. Melakukan Encode pada kolom Dataset

```
[ ] from sklearn.preprocessing import LabelEncoder

    # Inisiasi label encoder
    encode = LabelEncoder()

    # Terpakan label encoder
    df['nama_kabupaten_kota'] = encode.fit_transform(df['nama_kabupaten_kota'])
    df['jumlah_produksi_sampah_per_tahun'] = encode.fit_transform(df['jumlah_produksi_sampah_per_tahun'])

    df
```

	nama_kabupaten_kota	tahun	jumlah_produksi_sampah_per_tahun	Terjadi Banjir (1=Ya, 0=Tidak)	jumlah_banjir
0	17	2017	130	1	21
1	20	2017	91	1	10
2	26	2017	0	0	0
3	28	2017	144	1	17
4	2	2017	93	1	7
...	...	...	...	...	...
185	35	2021	17	1	20
186	34	2021	33	0	0
187	32	2021	139	1	18
188	37	2021	159	1	20
189	29	2021	136	1	10

Gambar 3.19. Encode Dataset

Melakukan encode pada kolom dataset yang berupa string. Yaitu pada kolom nama kabupaten / kota.

3. Memisahkan fitur dengan label

```
[ ] # Memisahkan fitur dengan label
    X = df.iloc[:,:3]
    y = df.iloc[:,3]
```

Gambar 3.20. Memisahkan Fitur dengan Label

Memisahkan fitur pada label sebagai berikut :

- X : 3 baris pada dataset yang dimulai dari nama kabupaten / kota sampai kolom jumlah sampah.
- Y : Baris setelah jumlah sampah, yaitu kolom terjadinya banjir.

4. Split Data Training dan Data Testing

```
[ ] # Split data training dan testing

    from sklearn.model_selection import train_test_split

    X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.3, random_state=30)
```

Gambar 3.21. Split Data Training dan Testing

Memisah data antara data X dan Y training dengan data X dan Y testing.

5. Hasil Akurasi Data Testing dan Training

```
[ ]  from sklearn.naive_bayes import GaussianNB
     from sklearn.metrics import accuracy_score

     # Inisiasi obyek MultinomialNB
     gnb = GaussianNB()

     # Fit model
     # Label y harus dalam bentu 1D atau (n_samples,)
     gnb.fit(X_train, y_train)

     # Prediksi dengan data training
     y_train_pred = gnb.predict(X_train)

     # Evaluasi akurasi training
     acc_train = accuracy_score(y_train, y_train_pred)

     # Prediksi test data
     y_test_pred = gnb.predict(X_test)

     # Evaluasi model dengan metric akurasi
     acc_test = accuracy_score(y_test, y_test_pred)

     # Print hasil evaluasi
     print(f'Hasil akurasi data train: {acc_train}')
     print(f'Hasil akurasi data test: {acc_test}')

     Hasil akurasi data train: 0.8796992481203008
     Hasil akurasi data test: 0.9473684210526315
```

Gambar 3.22. Hasil Akurasi Data Testing dan Training

Berikut merupakan hasil dari data training dan data testing yang tingkat akurasinya berkisar 0,87 dan 0,94 yang menunjukkan bahwa data ini cukup akurat.

BAB IV
KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan penelitian tugas akhir ini telah dilakukan implementasi dengan dua metode machine learning pada dataset jumlah produksi sampah pada seluruh daerah di Jawa Timur dalam tiap tahunnya. Dataset yang telah dipilih berdasarkan topik yang akan dibahas, dan dataset tersebut diperoleh dari Badan Pusat Statistik atau BPS. Dalam penelitian ini diambil topik prediksi banjir untuk memprediksi terjadinya banjir pada daerah Jawa Timur, berikut beberapa kesimpulan yang berhubungan dengan tujuan penelitian :

1. Penelitian ini dapat memprediksi banjir dari jumlah produksi sampah dengan tahapan pelaksanaan seperti berikut :
   a) Identifikasi masalah
   b) Pengumpulan data
   c) Pemilihan metode
   d) Penyusunan konsep
   e) Pembuatan kode program
   f) Evaluasi dan kesimpulan
2. Penelitian ini menggunakan dua metode yaitu regresi linier sederhana dan naive bayes.
3. Penelitian ini berhasil mengimplementasikan dua metode tersebut untuk memprediksi adanya banjir dari produksi sampah.
4. Pada metode pertama yaitu regresi linier sederhana, yang mana menggunakan dua sampel dalam penentuan hasilnya. Sampel tersebut diperoleh dari variabel X dan Y. Pada variabel X ini merupakan variabel bebas yang mempengaruhi variabel Y, dan variabel Y merupakan variabel terikat yang dipengaruhi oleh variabel X.
5. Pada metode kedua yaitu naive bayes, yang mana menggunakan perhitungan mean dan standar deviasi dari data produksi sampah di kedua kondisi yaitu kondisi terjadi banjir dan tidak terjadi banjir. Kemudian hasil mean dan standar deviasi tersebut dilakukan perhitungan pada rumus naive bayes, yang nantinya diperoleh hasil prediksi untuk setiap kondisi banjir iya dan tidak.

4.2 Saran

Saran yang diperlukan untuk pengembangan program lebih lanjut antara lain :

1. Penambahan tampilan proses pelatihan sehingga pengguna dapat melihat proses yang terjadi.
2. Penggunaan metode klasifikasi data yang lain untuk perbandingan.
3. Penggunaan data yang lebih bervariasi.

DAFTAR PUSTAKA

Amalia, I., Suzanna, E., & Dewi, R. 2021. Asesmen Psikologis Korban Bencana Banjir Bandang Aceh Tengah. *Jurnal Penelitian Pendidikan, Psikologi Dan Kesehatan (J-P3K)*, *2*(1), 7–13. doi:10.51849/j-p3k.v2i1.69

Badan Pusat Statistik. 2018. *Banyaknya Desa/Kelurahan Menurut Jenis Bencana Alam dalam Tiga Tahun Terakhir (Desa)*. https://www.bps.go.id/indicator/168/954/1/banyaknya-desa-kelurahan-menurut-jenis-bencana-alam-dalam-tiga-tahun-terakhir.html

Gautama, T. K., Hendrik, A., & Riskadewi. (2016). Pengenalan Objek pada *Computer vision* dengan Pencocokan Fitur Menggunakan Algoritma SIFTStudi Kasus: Deteksi Penyakit Kulit Sederhana. Jurnal Teknik Informatika dan Sistem Informasi, 437-450.

Hutauruk, T. R., Kusuma, A. R., & Ningsih, W. 2020. Estimasi Kerugian Ekonomi Akibat Banjir Pada Kawasan Pemukiman Penduduk Di Bantaran Sungai Karang Mumus Kota Samarinda. *Jurnal Riset Inossa*, *2*(1), 47–59.

Isnawaty, Subardin, & Normawan, L. L. (2022). Penerapan Internet Of Things (Iot) Pada Sistem Monitoring Tempat Sampah Rumah Tangga Menggunakan Metode Haversine Formula. Digital Transformation Technology (Digitech), 35-44.

Muzakky, A., Nurhadi, A., Nurdiansyah, A., Wicaksana, G., & Istiadi. 2018. Perancangan Sistem Deteksi Banjir Berbasis IoT. *Conference on Innovation and Application of Science and Technology (CIASTECH)*, 660–667.

Putra, I. G. N. A. W., & Mandala, I. G. N. P. 2020. Penilaian Kerusakan Dan Kerugian Infrastruktur Publik Akibat Dampak Bencana Banjir Di Kota Semarang. *Wahana Teknik Sipil: Jurnal Pengembangan Teknik Sipil*, *1*(2), 86–97.

Putra, I. S. W., Hermawan, F., & Hatmoko, J. U. D. 2020. Penilaian Kerusakan Dan Kerugian Infrastruktur Publik Akibat Dampak Bencana Banjir Di Kota Semarang. *Wahana Teknik Sipil: Jurnal Pengembangan Teknik Sipil*, *86*(97), 77–84.

Saidi, S., Kattan, L., Jayasinghe, P., Hettiarachi, P., & Taron, J. 2018. Integrated infrastructure systems—A review. *Sustainable Cities and Society*, *36*, 1–11.

Santri, Apriyanto, E., & Utama, S. P. 2020. Dampak Sosial Ekonomi Dan Estimasi Kerugian Ekonomi Akibat Banjir Di Kelurahan Rawa Makmur Kota Bengkulu. *Naturalis: Jurnal Penelitian Pengelolaan Sumber Daya Alam Dan Lingkungan*, *9*(2), 77–84.

Sembiring, S. T. B. 2018. *Peranan Pemerintah Kota Dalam Penanggulangan Sampah Di Tps Kelurahan Padang Bulan Kecamatan Medan Baru Kota Medan*.

Setiawan, H., Jalil, M., Enggi, M. S., Purwadi, F., Adios, C. S., Brata, A. W., & Jufda, A. S. 2020. Analisis Penyebab Banjir Di Kota Samarinda. *Jurnal Geografi Gea*, *20*(1), 39–43.