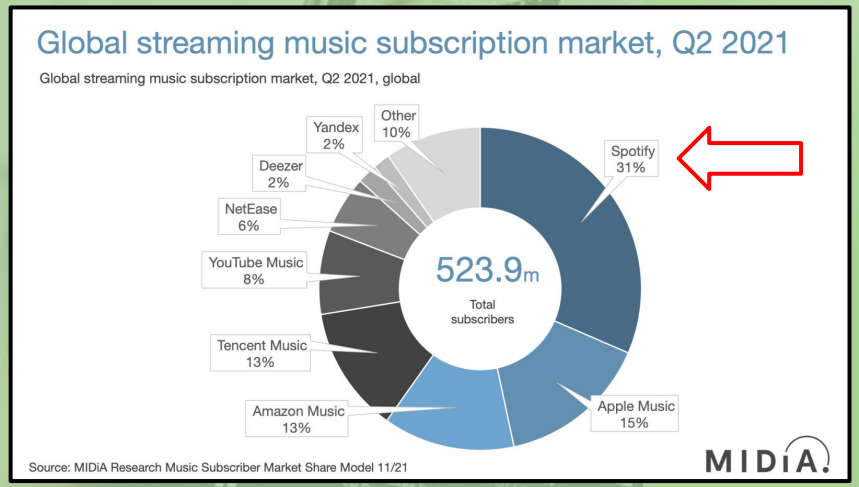
Analisis Klasifikasi Genre Musik Pop Menggunakan Regresi Logistik dan *K-Nearest Neighbors* pada Platform *Streaming* Spotify

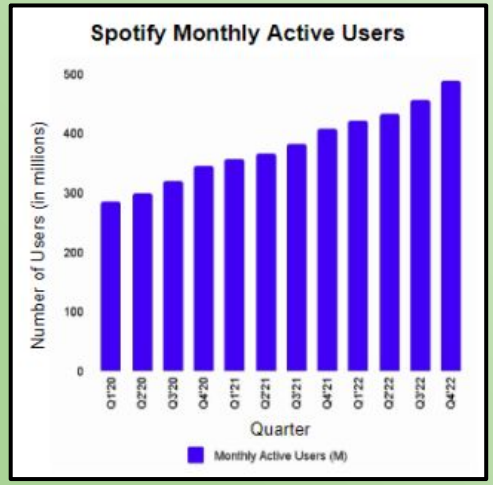
Seminar Tugas Akhir
Katlyn Kenisha - 23101910080
Dosen Pembimbing 1: Maria Zefanya Sampe, M.Si., M.M.
Dosen Pembimbing 2: Maydison Ginting, Ph.D.

Latar Belakang

Teknologi

- Teknologi terus dikembangkan dan mengubah kehidupan manusia dalam berbagai aspek.
- Kini, musik dapat diakses dengan mudah melalui platform *streaming* musik seperti **Spotify**.





Spotify

- Unggul dengan *market share* sebesar 31% sampai dengan Q2-2021.
- Berhasil meningkatkan jumlah *subscriber* berbayar hingga 487 juta pengguna pada akhir tahun 2022.

User

- Menikmati berbagai jenis musik.
- Mendengarkan lagu baru yang mirip dengan lagu yang sudah mereka sukai.
- Tidak ingin bersusah payah untuk mendengarkan lagu.

Spotify

- Pengelompokkan musik ke berbagai genre.
- Mengidentifikasi perilaku pengguna.
- Memberikan rekomendasi lagu.
- Membuat *playlist*.

Rumusan Masalah

1. Bagaimana penerapan model klasifikasi dengan menggunakan regresi logistik dan KNN?
2. Bagaimana performa dari model klasifikasi menggunakan metode regresi logistik dan KNN berdasarkan metrik *precision*, *recall*, dan *F1-score*?

Batasan Penelitian

1. **Data**: Lagu-lagu dari 1 Januari 2010 - 31 Desember 2019 yang ada di Spotify.
2. **Variabel Independen**: *Acousticness*, *danceability*, *duration*, *energy*, *instrumentalness*, *key*, *liveness*, *loudness*, *mode*, *speechiness*, *tempo*, dan *valence*. **Variabel dependen**: Genre.
3. **Metode**: Regresi logistik dan KNN.
4. ***Software***: Google Colab (Python).

Tinjauan Pustaka

Genre Musik

Pengelompokkan musik sesuai dengan kemiripannya satu sama lain.

Regresi Logistik

Metode *supervised learning* yang digunakan untuk menganalisa hubungan antara variabel independen dengan variabel dependen yang bersifat kategorik.

$$\pi(x_i) = \frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_i x_i)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_i x_i)}$$

dimana

$\pi(x)$	= peluang sukses suatu kejadian
β_0	= konstanta
x_i	= variabel independen ke-i
β_i	= koefisien dari variabel independen ke-i
i	= banyaknya variabel independen x

KNN

Metode *supervised learning* yang digunakan untuk mengelompokkan data berdasarkan beberapa tetangga terdekat.

$$d(x_i, y_i) = \sum_{i=1}^{k} |x_i - y_i|$$

dimana

$d(x_i, y_i)$	= jarak *Manhattan* antar objek
i	= variabel data
x_i	= data *training*
y_i	= data *testing*
k	= dimensi data

Tinjauan Pustaka

Confusion Matrix

Metode untuk mengevaluasi performa model dengan membandingkan hasil klasifikasi yang dilakukan oleh model (*predicted*) dengan hasil sebenarnya (*actual*).

		Actual Class	
		Class 1	*Class* 0
Predicted Class	*Class* 1	*True Positive* (TP)	*False Positive* (FP)
	Class 0	*False Negative* (FN)	*True Negative* (TN)

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall}$$

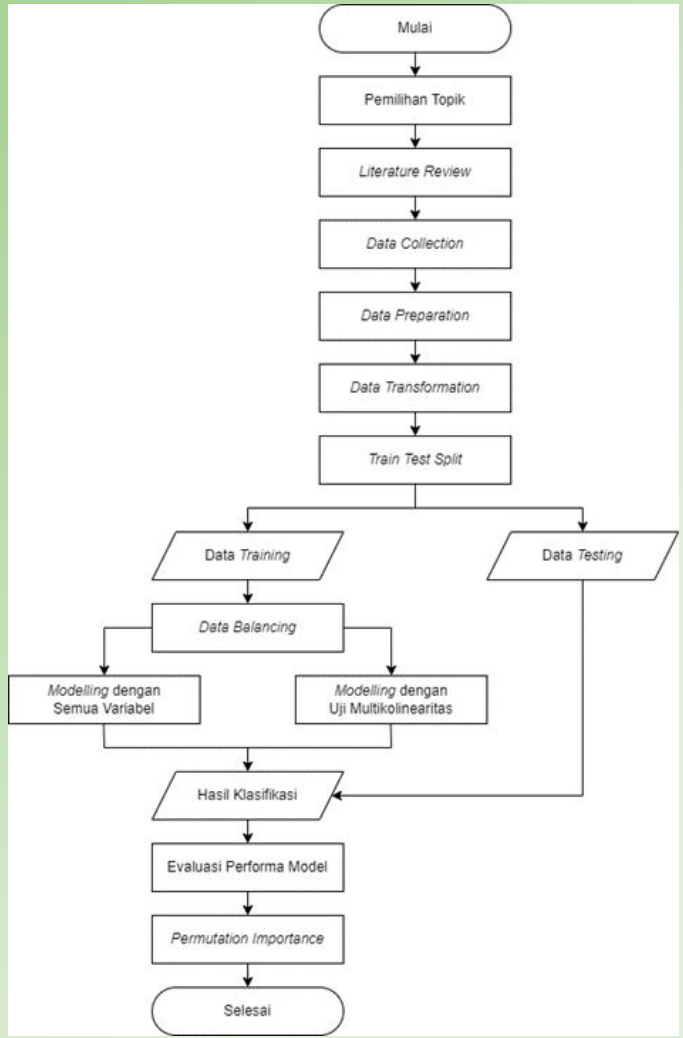
Permutation Importance

Metode untuk menilai kontribusi setiap variabel independen terhadap variabel dependen dengan cara melakukan permutasi acak pada suatu variabel independen.

Penelitian Terdahulu

1. *K-Nearest Neighbor Algorithm for Feature Reinforcement and Music Genre Prediction in Mobile Applications: Comparison to Decision* oleh Bharath dan Saraswathi (2023).
2. *Automatic Categorization of Electronic Music Genres* oleh Krebbers (2020).
3. *Music Genre Classification using Machine Learning* oleh Seethal, Vijayakumar (2021).
4. *Single-labelled Music Genre Classification Using Content-Based Features* oleh Ajoodha, Klein, dan Rosman (2015).

Tahapan Penelitian



Pengolahan Data

Data Collection

- *Web* API Spotify
- *Library*: Spotipy
- Membuat fungsi untuk meng-*import playlist*

	artist	track_title	track_id	release_date	popularity	acousticness	danceability	duration_ms	energy
0	Charlie Puth	Dangerously	3qonjOrhFCfTnaaMruHzxW	2016-01-29	72	0.364	0.696	199133	0.517
1	Rizky Febian	Hingga Tua Bersama	5b0NpyYAwW2dUGL08Ir7Bg	2021-05-12	72	0.796	0.579	270926	0.459

Data Preparation

Data Cleaning

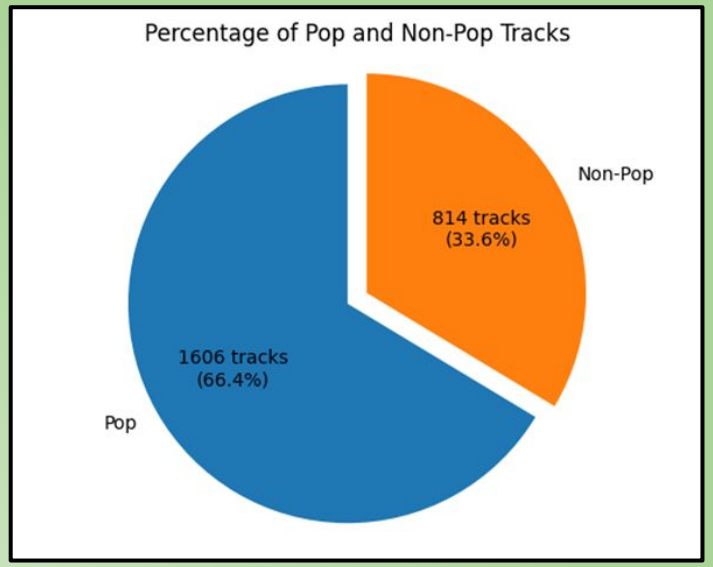
- Menghapus data duplikat, data dengan variabel kosong
- *Filtering release date* dari lagu (2010-2019)

Data Preprocessing

- Mengubah variabel *duration* dari milidetik menjadi menit
- Mengubah variabel genre menjadi hanya ‘pop’ atau ‘non-pop’

Data Preparation (continued)

Exploratory Data Analysis (EDA)



Pengolahan Data

Data Transformation

- Encoding variabel genre dimana pop = 1 dan non-pop = 0
- Normalisasi dengan *Min-Max* terhadap variabel *duration*, *key*, *loudness*, dan *tempo*

Train Test Split

- 80% *training* dan 20% *testing*

Data Balancing

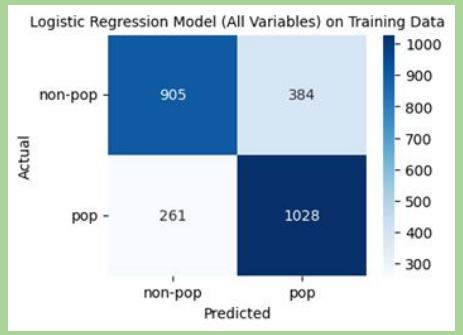
- *Balancing* terhadap data *training* dengan SMOTE

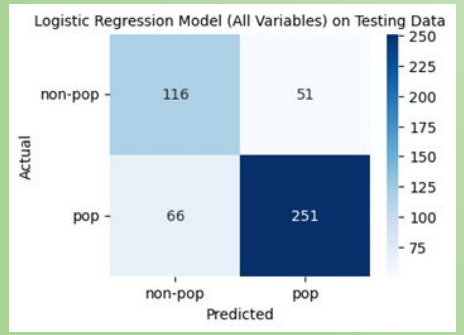
Modelling

- 2 jenis klasifikasi; regresi logistik dan KNN
- 2 jenis model; model dengan semua variabel dan model dengan variabel bebas multikolinearitas
- 5-*Fold cross validation*

Model dengan Semua Variabel

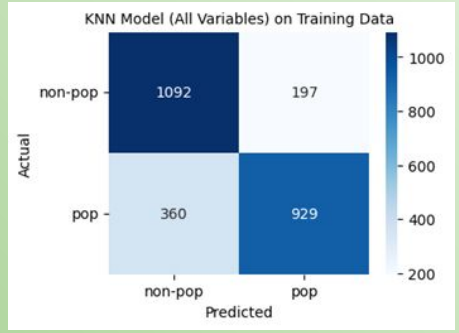
Regresi Logistik

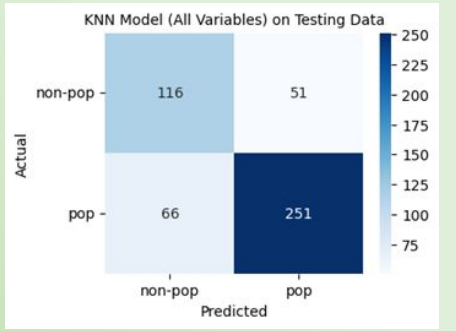




	Training	*Testing*
Precision	72.8%	83.1%
Recall	79.8%	79.2%
F1-Score	76.1%	81.1%

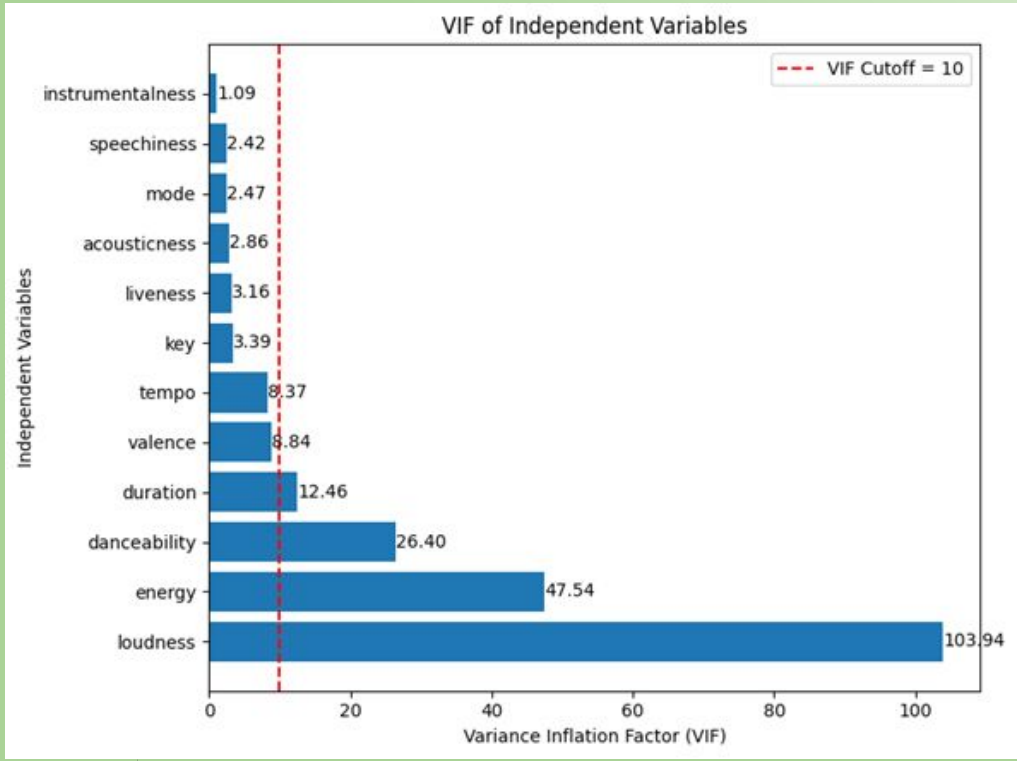
KNN

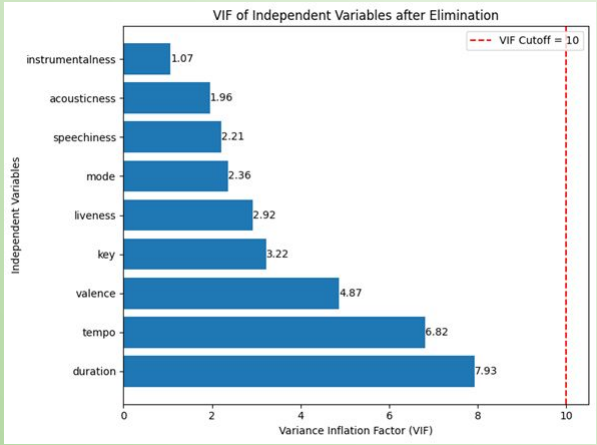




	Training	*Testing*
Precision	82.5%	78.7%
Recall	72.1%	73.5%
F1-Score	76.9%	76%

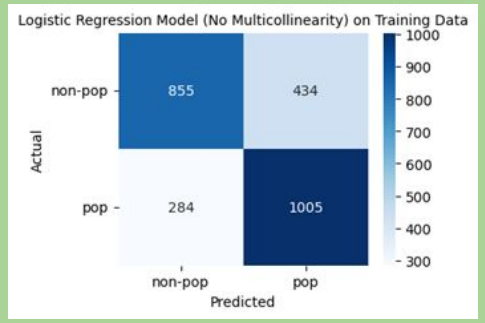
Uji Multikolinearitas

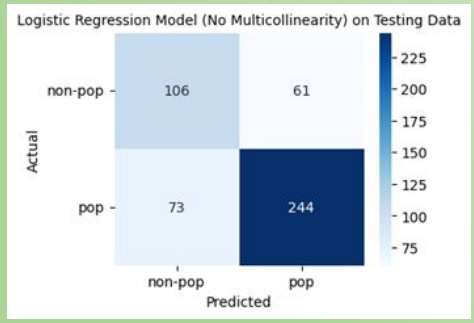




Model dengan Variabel Bebas Multikolinearitas

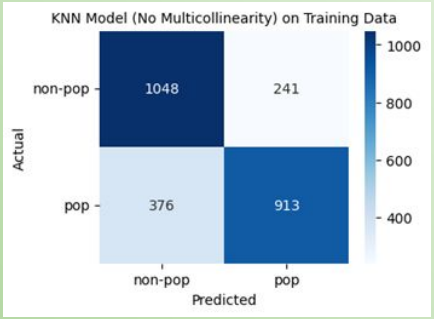
Regresi Logistik

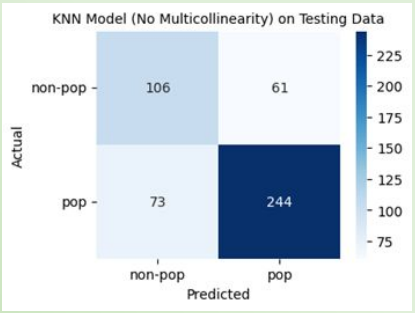




	Training	*Testing*
Precision	69.8%	80%
Recall	78%	77%
F1-Score	73.7%	78.5%

KNN



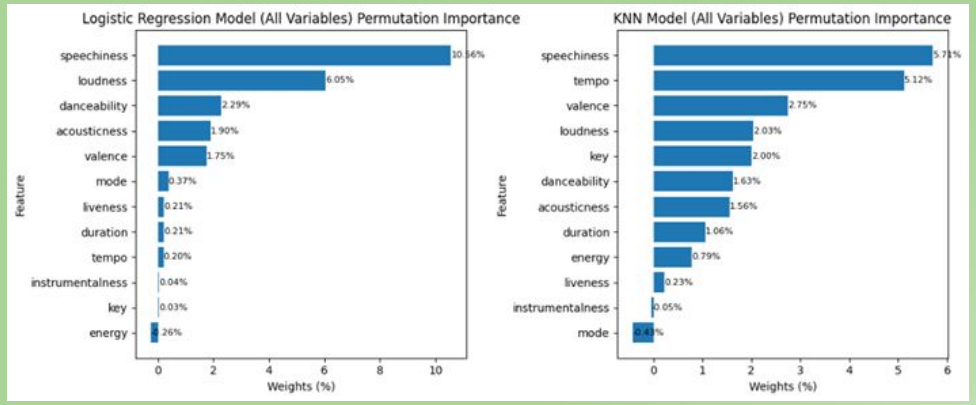


	Training	*Testing*
Precision	79.1%	78.3%
Recall	70.8%	74.1%
F1-Score	74.7%	76.2%

Perbandingan Performa Model

Model	Variabel	*Precision*		*Recall*		***F1-Score***	
		Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
Regresi Logistik	Semua	72.8%	83.1%	79.8%	79.2%	76.1%	**81.1%**
	Bebas Multikolinearitas	69.8%	80%	78%	77%	73.7%	78.5%
KNN	Semua	82.5%	78.7%	72.1%	73.5%	76.9%	76%
	Bebas Multikolinearitas	79.1%	78.3%	70.8%	74.1%	74.7%	76.2%

Permutation Importance

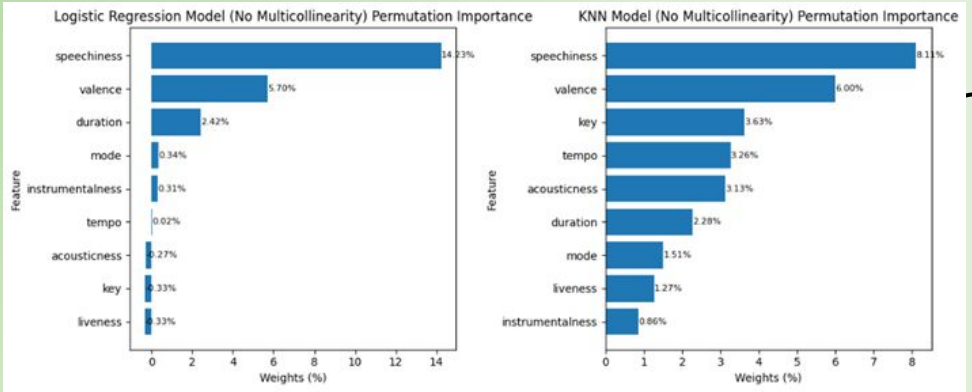


Semua Variabel

←

Variabel Bebas Multikolinearitas

→



Kesimpulan

- Dalam penerapan model klasifikasi, ditemukan bahwa isu *underfitting* muncul dalam regresi logistik. Di sisi lain, parameter optimal untuk model KNN yang ditemukan adalah k = 3 dan metrik jarak = *Manhattan*.
- Model regresi logistik mempunyai performa yang lebih baik dari model KNN. Model terbaik berdasarkan asumsi yang terpenuhi dan metrik yang dihasilkan adalah **model regresi logistik dengan variabel bebas multikolinearitas** dengan nilai *F1-Score* sebesar 78.5%.

Saran

- Mencoba metode klasifikasi lainnya seperti *Random Forest*, SVM, dan lain-lain.
- Menambahkan variabel independen lainnya seperti analisis lirik pada lagu.
- Mengimplementasikan *multiclass classification*.

Referensi

- MIDiA, “Music subscriber market shares Q2 2021,” [Online]. Available: https://www.midiaresearch.com/blog/music-subscriber-market-shares-q2- 2021
- Spotify, “Shareholder Deck Q4 2022 Update,” [Online]. Available: https://s29.q4cdn.com/175625835/files/doc_financials/2022/q4/ShareholderDeck-Q4-2022-FINAL.pdf
- “Spotify - The User Experience,” [Online]. Available: https://henree.me/projects/spotify-user-experience
- C. Chen, “Spotify Questionnaire: Personalizing Auto-Generated Playlists,” [Online]. Available: https://cc2395.medium.com/spotify-questionnaire-personalizing-auto-generated-playlists-a648486f5b0
- E. Jones, “Survey shows students prefer Spotify for streaming music; Apple Music is runner-up,” [Online]. Available: https://kealakai.byuh.edu/survey-shows-students-prefer-spotify-for-streaming-music-apple-music-is-runner-up
- D. W. Hosmer, Jr., S. Lemeshow dan R. X. Sturdivant, Applied Logistic Regression. 2013
- I. José, “KNN (K-Nearest Neighbors) #1,” [Online]. Available: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d