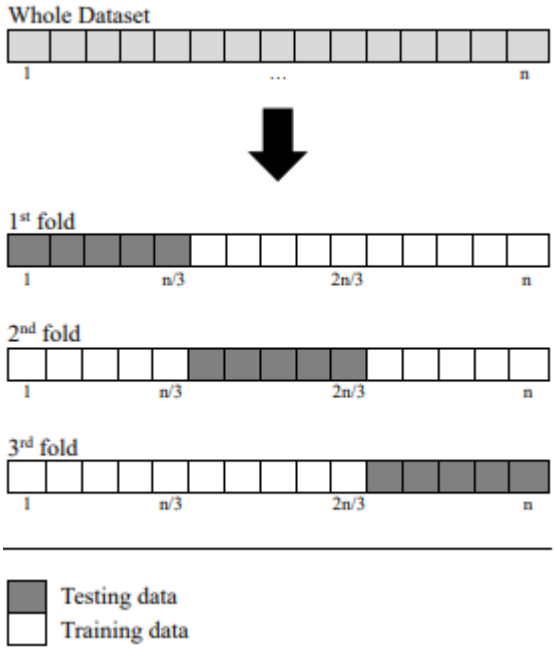
BAB II

TINJAUAN PUSTAKA

2.1 Kajian Pustaka

Penelitian yang berkaitan dengan klasifikasi penyakit sudah banyak dilakukan dan banyak diterapkan dalam berbagai bidang kehidupan. Dalam penelitian yang sudah dilakukan sebelumnya menunjukkan bahwa klasifikasi pada penyakit sangat diperlukan untuk dijadikan sebagai landasan bagi penelitian selanjutnya agar lebih baik lagi. Hal tersebut juga dapat memudahkan dan membantu penelitian berikutnya.

Penelitian dengan judul *"Machine Learning For Data Classification In Indonesia Regional Elections Based On Political Parties Support"* dilakukan oleh M. Fachrie pada tahun 2020. Pada penelitian ini menggunakan dan membandingkan algoritma *K-Nearest Neighbors, Naïve Bayes Classifier, Decision Tree (C4.5), and Neural Networks (Multilayer Perceptron).* Prediksi dilakukan dengan menggunakan data yang diambil dari KPU dan klasifikasi antara dua kelas data yaitu menang dan kalah. Datanya terdiri dari komposisi partai politik yang mendukung masing-masing kandidat. Semua algoritma divalidasi menggunakan teknik *10- fold Cross Validation. Cross Validation* (X-Val) adalah metode yang digunakan untuk mengevaluasi dan memvalidasi kinerja sistem kami dalam konteks klasifikasi. Metode ini melatih dan menguji model menggunakan beberapa kombinasi dataset yang berbeda. X-Val membuat beberapa kombinasi data pelatihan dan pengujian yang terpisah. Jumlah kombinasi ditentukan oleh nilai k-fold, mis. 3, 5, atau 10. Berikut ilustrasi cara kerja X-Val untuk k-fold= 3 pada Gambar 2.1.



Gambar 2. 1 Ilustrasi bagaimana metode Cross Validation dibuat beberapa kombinasi dataset menggunakan 3-fold.

Penelitian dengan judul *"Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python"* dilakukan oleh F. Sodik, B. Dwi, and I. Kharisudin pada tahun 2020. Pada penelitian ini menggunakan dan membandingkan algoritma Regresi logistik, K-nearest neighboor, naive bayes, super vector machine, dan random forest. Prediksi dilakukan menggunakan *data churn modelling* dari Kaggle. Perbandingan 9:1 pada data training dan testing. Kemudian dilakukannya *confusion matrix* tiap model algoritma. Model Algoritma klasifikasi random forest lebih baik dari model algoritma lainnya yaitu nilai akurasi 86,2%, nilai *precision* 0,740, nilai *recall* 0,482, dan nilai *f1* adalah 0,584.[31]

Penelitian dengan judul *"PREDIKSI JUMLAH PENDERITA PENYAKIT TUBERKULOSIS DI KOTA BANDAR LAMPUNG MENGGUNAKAN METODE SVM (Support Vector Machine)"* dilakukan oleh F. R. Lumbanraja, I. H. B. Sitepu, D. Kurniawan, and A. Aristoteles pada tahun 2020. Pada penelitian ini menggunakan algoritma *Support Vector Machine* dengan 3 kernel yaitu, *Linear, Gaussian, dan Polynomial*. Prediksi menggunakan data penderita tuberkulosis di kota tersebut, data cuaca dan

matrix jarak antar penderita di lingkup kecamatan. Pada penelitian ini menggunakan 600 data dengan 44 variable. Kemudian melakukan percobaan tanpa dan dengan feature selection serta matrix jarak pada tiap kernel. Dari hasil percobaan, R2 terbesar pada *Kernel Gaussian* tanpa menggunakan *Feature Selection* dan tanpa matriks jarak yang masing-masing nilainya adalah 67 % dan 58.53 %.[33]

Penelitian dengan judul *"Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation"* dilakukan oleh M. F. Faruque, Asaduzzaman, S. M. M. Hossain, M. H. Furhad, and I. H. Sarker pada tahun 2020. Pada penelitian ini menggunakan algoritma *Support Vector Machine Naive Bayes*, *K-Nearest Neighbors* (KNN) dan *C4.5 Decision Tree.* Prediksi menggunakan dataset dari *Medical Centre Chittagong, Bangladesh* dengan 200 pasien diabetes mengggunakan 16 attributes data. Kemudian mencari model terbaik dengan korelasi tertinggi dengan beberapa variable atau attributes data yang ada. Model *Decision Tree* lebih baik dari model lainnya dengan hasil akurasi 73.5%, *F-measure* 72%, dan *AUC* 0.69. Ada korelasi positif untuk memprediksi komplikasi ginjal (Nefropati) dan tekanan darah (Hipertensi) komplikasi dan korelasi negatif dalam memprediksi gangguan pendengaran dan komplikasi kulit (diabetes dermopathy) dari pasien diabetes. Ini akan membantu pasien untuk menyadari faktor risiko yang berhubungan dengan diabetes. [7]

Penelitian dengan judul *"Predictive Supervised Machine Learning Models for Diabetes Mellitus"* dilakukan oleh L. J. Muhammad, E. A. Algehyne, and S. S. Usman pada tahun 2020. Pada penelitian ini menggunakan dan membandingkan algoritma supervised learning yang terdiri dari *Logistic regression, support vector machine, K-nearest neighbor, random forest, naive Bayes* dan *gradient booting.* Prediksi menggunakan dataset diabetes type 2 dari rumah sakit *Murtala Mohammed Specialist, kano.* Kemudian penggunaan *confusion matrix*, dsb. Selanjutnya ditemukan sebagai model terbaik di antara model adalah *random forest* yang dikembangkan

dengan akurasi 88,76%, sedangkan dalam hal kurva karakteristik pengoperasian penerima, *random forest dan gradient boosting* tampaknya menjadi model klasifikasi prediktif terbaik dengan 86,28%.[10]

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Perbandingan	Kontras	Mengkritik	Mempersatukan	Meringkaskan
1	*Machine Learning For Data Classification In Indonesia Regional Elections Based Political Parties Support* [18]	Penelitian dilakukan untuk memprediksi kemenangan kandidat pada Pemilihan Kepala Daerah di Indonesia dengan membandingkan beberapa model yaitu *K-Nearest Neighbors, Naïve Bayes Classifier, Decision Tree (C4.5), and Neural Networks (Multilayer Perceptron).* Prediksi dilakukan dengan	Penelitian ini dilakukan untuk mengembangkan model Machine Learning yang berbasis pada data yang telah terverifikasi oleh lembaga resmi untuk memprediksi kemenangan masing-masing kandidat dalam pemilihan daerah menggunakan data media sosial.	Belum melakukan percobaan dengan jumlah data yang lebih banyak dengan model yang sesuai.	Data dari penelitian ini terdiri dari komposisi partai politik yang mendukung masing-masing kandidat. Kemudian divalidasi menggunakan teknik *10- fold Cross Validation.*	Penelitian dengan menggunakan *Neural Networks* dengan arsitektur *Multilayer Perceptron* memiliki prediksi yang lebih baik. Hasil akurasinya adalah 74,20%.

No	Judul	Perbandingan	Kontras	Mengkritik	Mempersatukan	Meringkaskan
		menggunakan data dari KPU.				
2	Perbandingan Metode Klasifikasi *Supervised Learning* pada Data Bank *Customers* Menggunakan Python[19]	Penelitian dilakukan untuk meprediksi bank customers dengan membandingkan beberapa model yaitu *Regresi logistik, K-nearest neighboor, naive bayes, super vector machine, dan random forest.* Prediksi dilakukan menggunakan *data churn modelling* dari Kaggle.	Penelitian dilakukan untuk menganalisis dan membandingkan metode–metode pendekatan *supervised learning.* Kemudian dilakukan analisis data dari *pre-processing* sampai hasil akurasinya.	Hanya membandigkan model *supervised learning* tanpa teknik validasi.	Penelitian dilakukan dengan melakukan pemodelan prosedur untuk mengolah dataset, pembagian data pelatihan dan testing sampai dapat hasil akurasi dari setiap model.	Penelitian dengan menggunakan metode klasifikasi *random forest* lebih baik hasilnya. Dengan hasil diantaranya seperti nilai akurasi 86,2%, nilai *precision* 0,740, nilai *recall* 0,482 dan nilai *f1* adalah 0,584.

No	Judul	Perbandingan	Kontras	Mengkritik	Mempersatukan	Meringkaskan
3	Prediksi Jumlah Penderita Penyakit Tuberkulosis di Kota Bandar Lampung Menggunakan Metode *SVM (Support Vector Machine)*[20]	Penelitian dilakukan untuk memprediksi penyakit tuberkulosis dengan model *Support Vector Machine* menggunakan 3 kernel yaitu, *Linear, Gaussian,* dan *Polynomial.* Prediksi dilakukan dengan menggunakan dataset data penderita, data cuaca, dan matrix jarak.	Penelitian dilakukan untuk mencari kernel terbaik dari *support vector machine.* Dengan menggunakan 600 data dengan 44 variabel.	Tidak melakukan banyak percobaan untuk mendapatkan hasilnya.	Penelitian dilakukan dengan pengujian model dengan percobaan tanpa dan menggunakan matrix jarak dan *feature selection* agar mendapat hasil akurasi yang optimal.	Penelitian ini mendapatkan R2 terbesar pada Kernel *Gaussian* dengan tanpa menggunakan *Feature Selection* dan tanpa matriks jarak. Hasil R2 tanpa *feature selection* dan matrix jarak adalah 47.67 % dan 58.53 %.
4	*Predicting Diabetes Mellitus and Analysing Risk-*	Penelitian dilakukan untuk memprediksi penyakit diabetes dan factor resiko korelasi	Penelitian dilakukan untuk mencari model terbaik dengan korelasi tertinggi dengan	Belum melakukan perbandingan dengan model	Penelitian dilakukan dengan melakukan pemodelan	Model *Decision Tree* lebih baik dari model lainnya dengan hasil akurasi 73.5%, *F-*

No	Judul	Perbandingan	Kontras	Mengkritik	Mempersatukan	Meringkaskan
	Factors Correlation [3]	variable data dengan membandingkan beberapa model yaitu *Support Vector Machine Naive Bayes, K-Nearest Neighbour (KNN) dan C4.5 Decision Tree (DT).*	beberapa variable data yang ada	lain dengan korelasi variable yang ada.	prosedur untuk mendapatkan hasil akurasi dan korelasi tertinggi dari berbagai variable data.	*measure* 72%, dan *AUC* 0.69. Ada korelasi positif untuk memprediksi komplikasi ginjal (Nefropati) dan tekanan darah (Hipertensi) komplikasi dan korelasi negatif dalam memprediksi gangguan pendengaran dan komplikasi kulit *(diabetes dermopathy)* dari pasien diabetes. Ini akan membantu pasien untuk

No	Judul	Perbandingan	Kontras	Mengkritik	Mempersatukan	Meringkaskan
						menyadari faktor risiko yang berhubungan dengan diabetes.
5	*Predictive Supervised Machine Learning Models for Diabetes Mellitus*[11]	Penelitian dilakukan untuk memprediksi penyakit diabetes dengan membandingkan beberapa model yaitu *Logistic regression, support vector machine, K-nearest neighbor, random forest, naive Bayes dan gradient booting algorithms.*	Penelitian dilakukan untuk mencari model terbaik dalam memprediksi penyakit dengan melakukan evaluasi matrix dengan perhitungan akurasi yang berbeda tiap model.	Hanya membandingkan model algoritma dan menghasilkan akuasi kurang dari 90%	Penelitian dilakukan dengan melakukan pemodelan prosedur untuk mendapatkan hasil akurasi yang optimal.	Model berbasis pembelajaran prediktif hutan acak ditemukan sebagai model terbaik di antara model yang dikembangkan dengan akurasi 88,76%, sedangkan dalam hal kurva karakteristik pengoperasian penerima, hutan acak dan booting gradien tampaknya menjadi model pembelajaran

No	Judul	Perbandingan	Kontras	Mengkritik	Mempersatukan	Meringkaskan
						prediktif terbaik dengan 86,28%.
6	Optimasi Algoritma K-Nearest Neighbors Dengan Teknik Cross Validation Dengan Streamlit (Studi Data: Penyakit Diabetes)	Penelitian ini dilakukan untuk memprediksi penyakit diabetes dengan model algoritma K-Nearest Neighbors dengan dan tanpa Teknik cross validation	Penelitian ini dilakukan untuk optimasi algoritma K-Nearest Neighbors dalam memprediksi penyakit diabetes dengan Teknik cross validation dilihat dari confusion matrix dan classification report keduanya.	Belum melakukan optimasi algoritma lain dengan Teknik yang serupa.	Penelitian dilakukan dengan melakukan pengujian model algoritma dengan dan tanpa Teknik cross validation untuk meminimalkan overfitting dan hasil akurasi yang optimal	*Classification report* yang memiliki nilai sebesar 92% lebih akurat daripada akurasi yang bernilai 94% karena ada penggunaan Teknik *cross validation* yang bisa meminimalkan overfitting.

2.2 Dasar Teori

2.2.1 Diabetes

Diabetes mellitus adalah salah satu penyebab utama morbiditas di seluruh dunia dan diperkirakan akan meningkat secara substansial selama beberapa dekade mendatang.[21] Ada tiga kategori utama diabetes:

a. Diabetes tipe 1 [1] kebanyakan terjadi pada anak-anak dan remaja. Dalam hal ini, tubuh memproduksi insulin sangat sedikit atau tidak sama sekali. Akibatnya, suntikan insulin setiap hari diperlukan untuk menjaga kadar glukosa tetap terkendali. Sering buang air kecil, penurunan berat badan secara tiba-tiba, rasa haus yang tidak normal, rasa lapar yang terus-menerus, penglihatan kabur, dan kelelahan adalah gejala umum dari diabetes jenis ini. Ini dapat diobati dengan bantuan terapi insulin.

b. Diabetes tipe 2 [1] lebih banyak terjadi pada orang dewasa (90% kasus). Tubuh tidak sepenuhnya merespon insulin yang mengakibatkan kadar glukosa lebih tinggi. Obesitas, pola makan yang tidak sehat, tekanan darah tinggi, dan kurangnya aktivitas fisik dianggap sebagai faktor risiko utama yang menyebabkan diabetes tipe 2. Suntikan insulin diperlukan ketika obat oral tidak cukup untuk mengontrol kadar gula darah.

c. Diabetes Melitus Gestasional (GDM), atau hanya diabetes gestasional terdiri dari tekanan darah tinggi selama kehamilan dan dapat menyebabkan komplikasi kesehatan bagi ibu dan anak. Biasanya menghilang selama tahap kehamilan tetapi yang terkena bersama dengan anak-anak mereka memiliki risiko terkena diabetes tipe 2 di kemudian hari. Menurut survei tahun 2017 [1], sekitar 204 juta wanita menderita GDM. Sekitar 21,3 juta kelahiran hidup memiliki beberapa bentuk hiperglikemia dalam kehamilan, di antaranya sekitar 85,1%

terjadi karena diabetes gestasional. GDM biasanya mempengaruhi sekitar satu dari tujuh kelahiran.

2.2.2 Machine Learning

Machine Learning adalah pembelajaran mengenai algoritma dalam mempelajari sesuatu untuk melakukan beberapa hal tertentu yang secara otomatis dilakukan manusia. Machine Learning adalah salah satu bidang kecerdasan buatan yang bisa mempengaruhi berbagai aspek lainnya, yaitu matematika, statistika, dan berbagai aspek dari komputer sains. Machine Learning bertujuan untuk mempelajari sebuah algoritma dalam melakukan sistem belajar secara otomatis dengan kontribusi yang sangat minimal yang pada umumnya dilakukan oleh manusia. [20] Machine Learning dapat dibagi menjadi empat bentuk, *Un-supervised learning; Supervised learning; Semi-supervised learning; and Reinforcement learning.* [22]

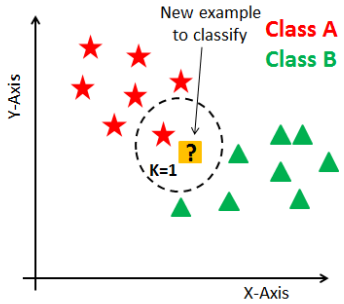
2.2.3 Klasifikasi

Machine learning memiliki beberapa bentuk dan salah satunya adalah supervised learning. Penggunaan machine learning dalam regresi dan klasifikasi, komunitas peneliti yang lebih luas berfokus pada *supervised learning.* [22] Klasifikasi dalam machine learning ada berbagai macam jenis algoritma machine learning dan akan dipakai pada penelitian ini yaitu *KNeighborsClassifier*. Klasifikasi berarti pengelompokan, penggolongan, menyusun data secara sistematis atau sesuai kaidah yang telah ditetapkan.

2.2.4 K nearest neighbors

Metode KNN merupakan salah satu metode klasifikasi dalam machine learning. Metode ini bekerja dengan mencari k pola (diantara semua pola latih disemua kelas) yang terdekat dengan pola masukan kemudian menentukan kelas keputusan berdasarkan jumlah pola terbanyak. [19] Proses pelatihan KNN menghasilkan k yang memberikan akurasi tertinggi dalam mengeneralisasi data yang akan datang. Masalahnya, sampai saat ini k tidak dapat ditentukan secara matematik. Jadi proses pelatihan proses pelatihan

pada dasarnya adalah melakukan observasi terhadap sejumlah k sampai dihasilkan k yang paling optimum. [19] Berikut visualisasi *K-Nearest Neighbors* pada gambar 2.2 dibawah ini.



Gambar 2. 2 Visualisasi KNN

Terlihat pada gambar ada class A dan class B sebagai output data. Ada pemilihan nilai k yang optimal. Misalnya ada variable y1 dan y2 dan memiliki lima data dengan output data berurutan A,B,B,A,A. Kemudian ada satu data baru yang memiliki dua variable yang sama(x1,x2) tapi tidak memiliki output data, class A atau B. Dalam hal ini, ada perhitungan jarak antara data baru dan masing-masing data lainnya dengan formula sebagai berikut.

$$\mathrm{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

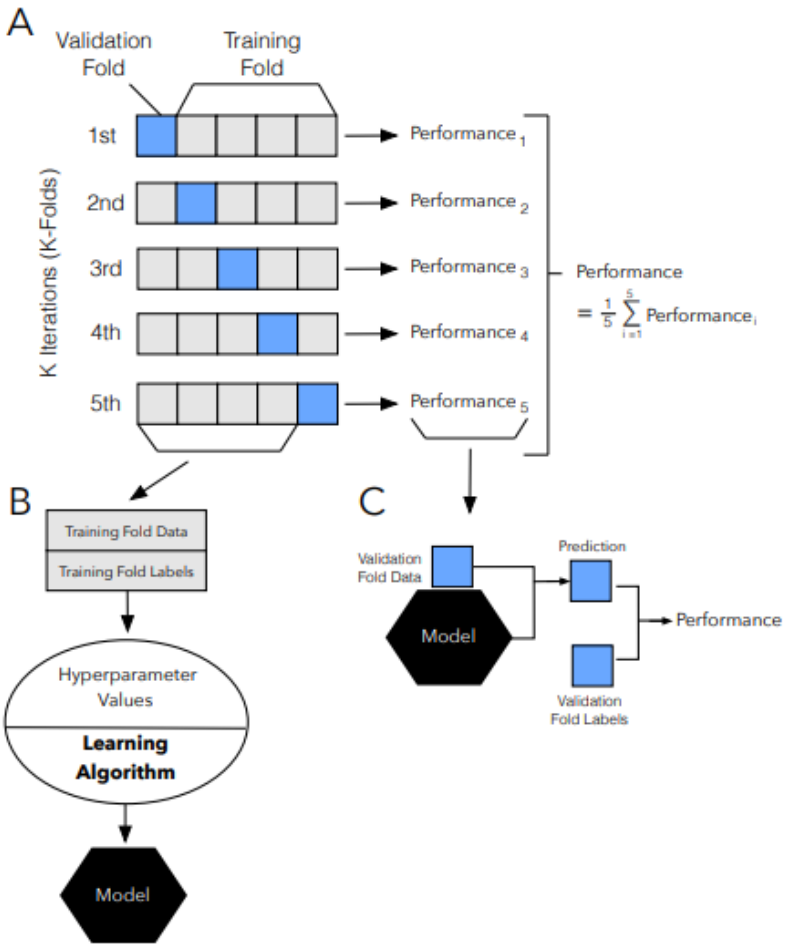
Gambar 2. 3 Formula Perhitungan Jarak

Proses perhitungannya adalah data baru(x) dikurangi data sebelumnya(y) dan sesuaikan formula. Setelah itu, data dirangkum dari jarak terdekat dan pilih jumlah data berdasarkan nilai k yang diambil. Analisis jumlah data yang diambil lebih banyak dari class A dan B. Sehingga data baru memiliki output data sesuai hasil analisis jumlah data yang diambil.

2.2.5 Cross Validation

Istilah validasi silang digunakan secara secara luas dalam literatur, dimana praktisi dan peneliti kadang-kadang merujuk pada metode ketidaksepakatan train/tes sebagai teknik validasi silang.

Namun, mungkin lebih masuk akal untuk menganggap validasi silang sebagai persilangan tahap pelatihan dan validasi dalam putaran yang berurutan. Di sini, ide utama di balik validasi silang adalah bahwa setiap sampel dalam kumpulan data kami memiliki peluang untuk diuji. *K-fold cross-validation* adalah kasus khusus dari cross-validation dimana kita melakukan iterasi pada kumpulan data sebanyak k kali.[23] Di setiap putaran, kami membagi dataset menjadi k bagian: satu bagian digunakan untuk validasi, dan k 1 bagian yang tersisa digabungkan menjadi subset pelatihan untuk evaluasi model seperti yang ditunjukkan pada gambar 2.4, yang menggambarkan proses validasi silang 5 kali lipat.

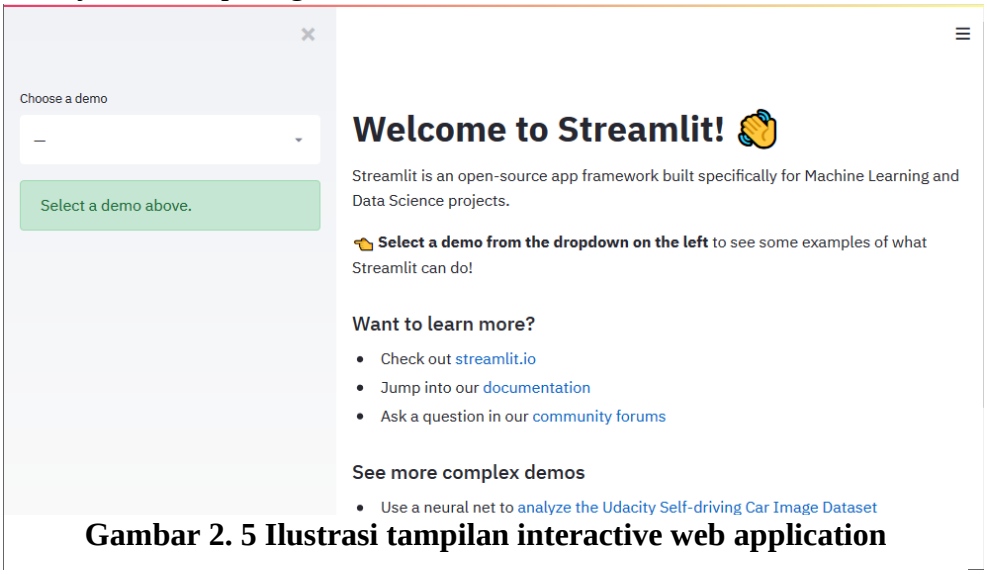


Gambar 2. 4 Ilustrasi prosedur validasi silang k-fold.

Selain itu, kumpulan fitur yang besar dapat menyebabkan masalah, seperti: (1) Waktu yang lama untuk melatih algoritme, (2) waktu yang lama dan banyak sumber daya yang dibutuhkan untuk menghasilkan variabel, dan (3) *overfitting* ketika terlalu banyak fitur yang tidak relevan digunakan.[24] Oleh karena itu, penilaian relevansi fitur merupakan aspek yang sangat penting dari tugas klasifikasi.

2.2.6 Streamlit

Streamlit mudah digunakan karena menggunakan perintah yang telah ditentukan sebelumnya untuk membangun aplikasi web berbasis data interaktif. Perintah sederhana seperti *st.write()* untuk mengimplementasikan berbagai objek langsung dari teks sederhana hingga pandas dataframe atau visualisasi matplotlib menjadi mungkin. [17] Model yang saya terapkan terdiri dari dari beranda arahan bersama dengan bilah sisi navigasi yang dapat berinteraksi dengan pengguna. Beranda terdiri dari informasi dasar tentang proyek dan membiasakan pengguna dengan seluruh pernyataan masalah. Halaman kumpulan data menampilkan kumpulan data proyek beserta deskripsinya dalam bentuk tabel menggunakan fungsi bingkai data pandas. Berikut ilustrasi tampilan dari *interactive web application* menggunakan library streamlit pada gambar 2.5



Gambar 2. 5 Ilustrasi tampilan interactive web application dengan streamlit