BAB II

TINJAUAN PUSTAKA

II.1. Penelitian Terkait

Penelitian terdahulu yang dilakukan oleh M. Rangga Ramadhan Saelan, et al. (2020) Komparasi Algoritma Klasifikasi untuk Prediksi Minat Sekolah Tinggi Pelajar pada *Students Alcohol Consumption*, yang berkesimpulan bahwa Teknik *Cross Validation* didapatkan statistik yang menunjukan bahwa algoritma *Decision Tree* memiliki kinerja lebih baik jika diabandingkan dengan *Naïve Bayes*. Algoritma *Decision Tree* memiliki tingkat akurasi sebesar 86.44% sedangkan *Naïve Bayes* hanya memiliki tingkat akurasi sebesar 82.60%. Dan berdasarkan statistic ROC, bisa dikatakan bahwa *Naïve Bayes* memiliki kinerja yang cukup buruk dengan tingkat *Equal Error Rate* (EER) sebesar 65%, sedangkan *Decision Tree* memiliki tingkat EER lebih rendah yaitu sebesar 55%. Dengan begitu algoritma *Decision Tree* memiliki kinerja lebih baik dalam mengidentifikasi kinerja pelajar dengan pengaruh berbagai faktor salah satunya alkohol.

Penelitian yang dilakukan oleh Noviyanti dan Hendrik (2018), Komparasi Kinerja Algoritma *Data Mining* pada Dataset Konsumsi Alkohol Siswa berkesimpulan bahwa model klasifikasi yang paling efisien dikelola dengan membandingkan ratarata akurasi menggunakan 10-*fold cross validation*. N*aive Bayes* dengan 5 fitur terbaik hasil *Gain Ratio* menghasilkan kinerja terbaik dengan akurasi 100%. Sementara, *Decision Tree C5.0* dengan fitur lengkap menghasilkan akurasi model prediksi terendah di antara model lainnya.

Penelitian yang dilakukan oleh Maulana Anggai Pribadi (2021), *Classification of Enchepalo Graph* (EEG) *Signals For Epilepsy Using Discreate Wavelet Transform and K-Nearest Neighbor Methods*, berkesimpulan bahwa n hasil nilai akurasi data bahwa K=1 memiliki persen terbesar yakni 100% dan persen terkecil terdapat pada K=7 dan K=11 yakni 14,2% dan 18,2% hal tersebut disebabkan karena adanya kelas yang tidak sesuai dengan data uji sehingga dapat mengurangi persentasi dari akurai pada K tersebut

Ahmad Azhari, et al (2015) yang melakukan penelitian berjudul Deteksi Kualitas Dan Kesegaran Telur Berdasarkan Deteksi Objek Transparam Menggunakan Metode DWT Dengan Klasifikasi KNN berkesimpulan bahwa akurasi terbaik sebesar 90.1% dan waktu komputasi 0,5682s dengan menggunakan metode DWT (*Discrete Wavelet Transform*) dengan level dekomposisi level 2 pada subband LL dengan klasifikasi KNN (*K-Neirest Neighbor*) menggunakan jarak euclidean pada K= 1.

Penelitian terdahulu yang dilakukan oleh Setyo Nugroho Wibowo, et al. (2017), Identifikasi Jenis Batuan Beku Melihan Bentuk Pola Batuan Menggunakan Metode DWT Dan KNN yang berkesimpulan bahwa sistem yang telah dibangun mampu mendeteksi batuan dengan level DWT yang digunakan adalah level 1, dengan akurasi terbaik adalah 98.33%. Dalam sistem ini, perubahan jenis *mother wavelet* tidak terlalu berpengaruh besar. Dan untuk komponen terbaik dalam sistem ini adalah komponen LL pada proses DWT dengan akurasi 98.33%. Sedangkan pada proses klasifikasi KNN jenis distance

terbaik yang bisa digunakan adalah jenis *Euclidean* dan *Cityblock* dengan akurasi terbaik 98.33% dengan parameter yang terbaik ada pada k=1 dan k=3.

II.2. Minuman Beralkohol

Indonesia (2012). "Peraturan Menteri Perindustrian." 71/MInd/PER/7/2012 tentang pengendalian dan pengawasan industri minuman beralkohol mendefinisikan minuman beralkohol adalah minuman yang mengandung etil alkohol atau etanol (C2H5OH), diproses dari bahan hasil pertanian yang mengandung karbohidrat dengan cara fermentasi dan destilasi atau fermentasi tanpa destilasi. Definisi ini terlihat jelas berdasarkan batas maksimum etanol yang diizinkan adalah 55%. Etanol dapat dikonsumsi karena diproses dari bahan hasil pertanian melalui fermentasi gula menjadi etanol, yang merupakan salah satu reaksi organik. Jika menggunakan bahan baku pati/karbohidrat, seperti beras, ketan, tape, singkong maka pati diubah terlebih dahulu menjadi gula oleh amylase untuk kemudian diubah menjadi etanol.

Penggolongan Minuman Beralkohol Menurut Peraturan Presiden Nomor 74 Tahun 2013, Minuman Beralkohol adalah minuman yang mengandung etil alkohol atau etanol (C2H5OH) yang diproses dari bahan hasil pertanian yang mengandung karbohidrat dengan cara fermentasi dan destilasi atau fermentasi tanpa destilasi. Minuman beralkohol yang berasal dari produksi dalam negeri atau asal impor dikelompokan dalam golongan sebagai berikut :

1. Minuman beralkohol golongan A adalah minuman yang mengandung etil alkohol atau etanol dengan kadar sampai dengan 5% (lima persen), Jenis 8

minuman ini paling banyak dijual di minimarket atau supermarket yaitu bir. Minuman tradisional yang termasuk minuman golongan A yaitu tuak dengan kadar alkohol 4% (Ilyas, 2013). Konsumsi alkohol golongan A dengan kadar 1 – 5% seseorang belum mengalami mabuk, tetapi tetap memiliki efek kurang baik bagi tubuh.

2. Minuman beralkohol golongan B adalah minuman yang mengandung etil alkohol atau etanol dengan kadar lebih dari 5% (lima persen) sampai dengan 20% (dua puluh persen). Jenis minuman yang termasuk di golongan ini adalah aneka jenis anggur atau wine. Alkohol pada kadar ini sudah cukup tinggi dan dapat membuat mabuk terutama bila diminum dalam jumlah banyak terutama bagi yang tidak terbiasa mengkonsumsi minuman beralkohol.
3. Minuman beralkohol golongan C adalah minuman yang mengandung etil alkohol atau etanol dengan kadar lebih dari 20% (dua puluh persen) sampai dengan 55% (lima puluh lima persen). Jenis minuman yang termasuk dalam golongan ini antara lain whisky, liquor, vodka, Johny Walker, dan lain-lain.

II.3. *Analysis Big Data*

Big Data merupakan istilah untuk menggambarkan data set yang besar baik *Structured, SemiStructured* maupun *Unstructured* data. Prosedur kerja *big data* dibagi menjadi empat diantaranya : Integritas data, Manage dan Analisis Data. Analisis data adalah proses pengolahan data untuk menemukan informasi

yang berguna sehingga data tersebut dapat digunakan dalam pengambilan keputusan untuk solusi suatu masalah. Dengan demikian pengertiand analisis *big data* adalah proses meneliti, mengolah data set besar (*Big Data*) untuk mengetahui pola tersembunyi, korelasi yang tidak diketahui, tren pasar, preferensi pelanggan dan informasi bisnis berguna lainnya

Peran implementasi *big data* salah satunya terciptanya *machine learning*. *Big data* diperlukan karena machine learning membutukan inpt berupa data yang besar, keberadaan *big data* membuat model machine learning dapat dibentuk karena input data berukuran besar.

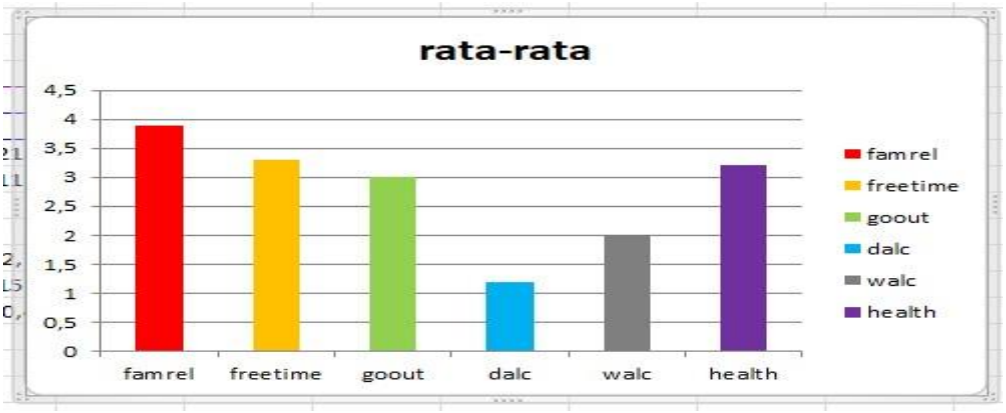
II.4. Metode DWT (*Descrate Wavelet Transform*)

Wavelet merupakan salah satu metode ekstraksi yang biasa digunakan pada sinyal. *Wavelet* mempunyai kemampuan untuk menganalisis sinyal *single* dan multidimensional, terutama jika sinyal tersebut memiliki informasi yang berbeda beda di tiap waktunya. Representasi *wavelet* adalah multiscale dari dekomposisi sinyal, yang dapat kita anggap sebagai pohon, dimana di tiap level menyimpan proyeksi sinyal ke dalam fungsi basisnya ke dalam resolusi tertentu atau dengan kata lain *wavelet* mengubah nilai menjadi serangkaian koefisien tertentu. DWT dihitung dengan sebuah kaskade filter dan diikuti oleh 2 subsampling. Berikut parameter yang digunakan pada DWT :

II.4.1. *Mean*

Mean merupakan parameter variable array untuk mencari nilai rata-rata input dari array data N menunjukan banyak *sample* dari input data n yang dimulai dari array 1 sampai N… .

$$\mu = \frac{1}{N}\sum_{i=1}^{N} n$$

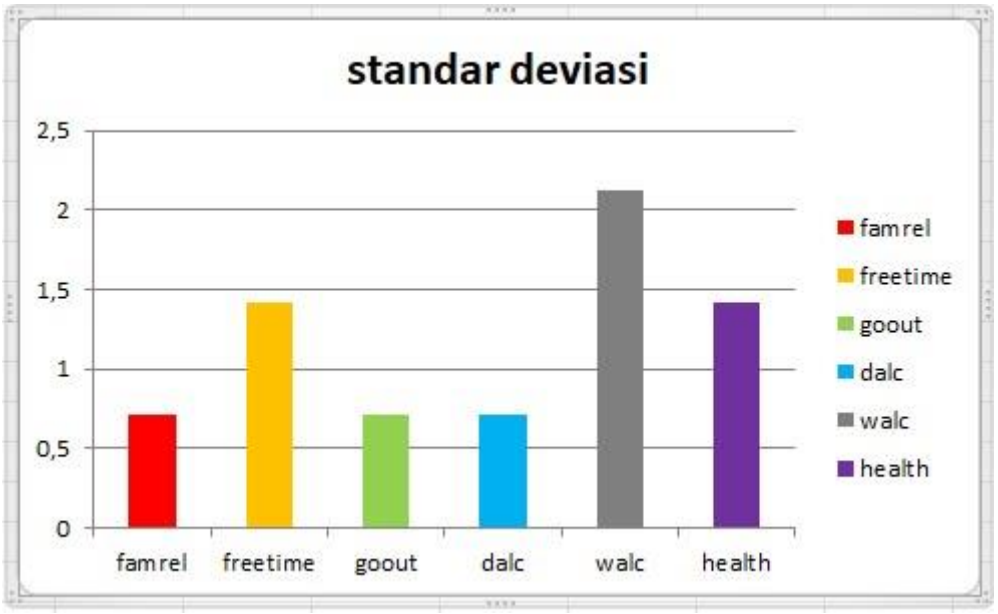


Gambar I1.I. Grafik Nilai *Mean* Pada *Sample*

II.4.2. Standar Devisiasi

Standar deviasi adalah rata-rata jarak dari *mean* ke suatu titik data. Digunakan untuk mengetahui seberapa tersebarnya suatu nilai dari data tersebut. Berikut adalah rumus dari standar deviasi.

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



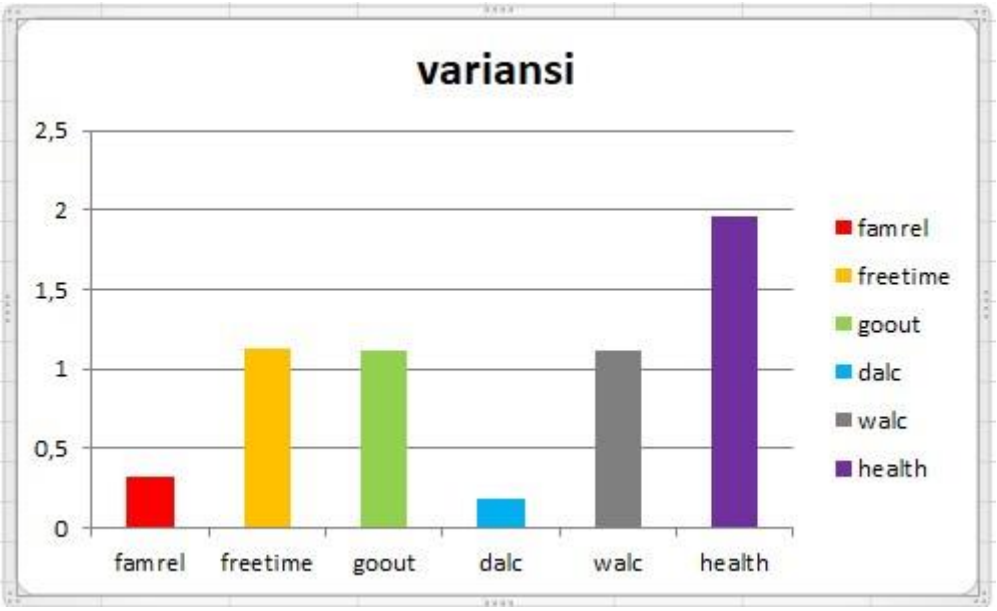
Gambar II.2. Grafik Nilai Standar Deviasi Pada *Sample*

n yang merupakan input, sedangkan N merupakan jumlah banyak *sample* dari data n yang berawal dari 1 dan μ meruapakan *Mean* dari n.

II.4.3. Variansi

Merupakan cara lain mengukur seberapa tersebarnya data dalam suatu kumpulan data. Variansi hampir sama dengan standar deviasi, namun variansi adalah kuadarat dari standar deviasi. Berikut adalah rumus dari variansi.

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



Gambar II.3. Grafik Nilai Variansi Pada *Sample*

n yang merupakan input, sedangkan N merupakan jumlah banyak *sample* dari data n yang berawal dari 1 dan μ meruapakan *Mean* dari n.

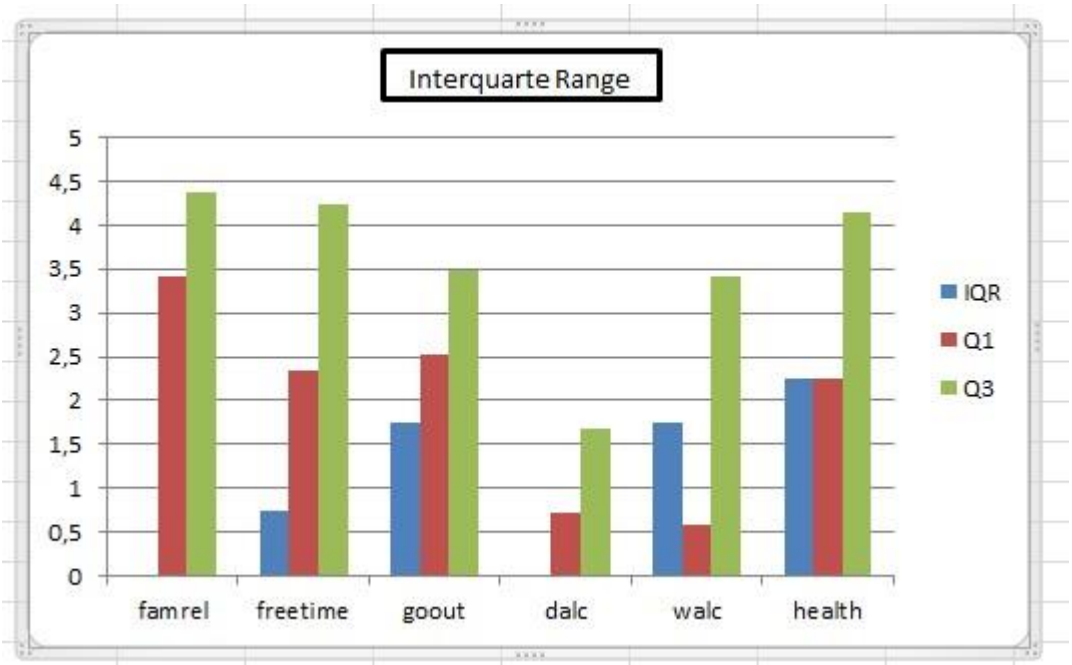
II.4.4.*Interquarte Range*

Interquartile Range adalah parameter statistik yang membagi kumpulan data kedalam kuartil. Kuartil membagi kumpulan data yang diperingkat menjadi tiga bagian yang sama. Nilai-nilai yang memisahkan bagian disebut kuartil pertama, kedua, dan ketiga; dan mereka dilambangkan dengan Q1, Q2, dan Q3, masing-masing.

$$Q3 - Q1 = IQR$$
$$Q1 = (\sigma z1) + \mu$$
$$Q3 = (\sigma z3) + \mu$$

Q1 merupakan kuartil atas dari array dan Q3 merupakan kuartil bawah dari array sedangkan z1 merupakan standar nilai dari kuartil satu (-0.67) dan z3

merupakan standar nilai dari kuartil tiga (+0.67) dan μ sebagai *mean* sedangkan σ merupakan standar deviasi.



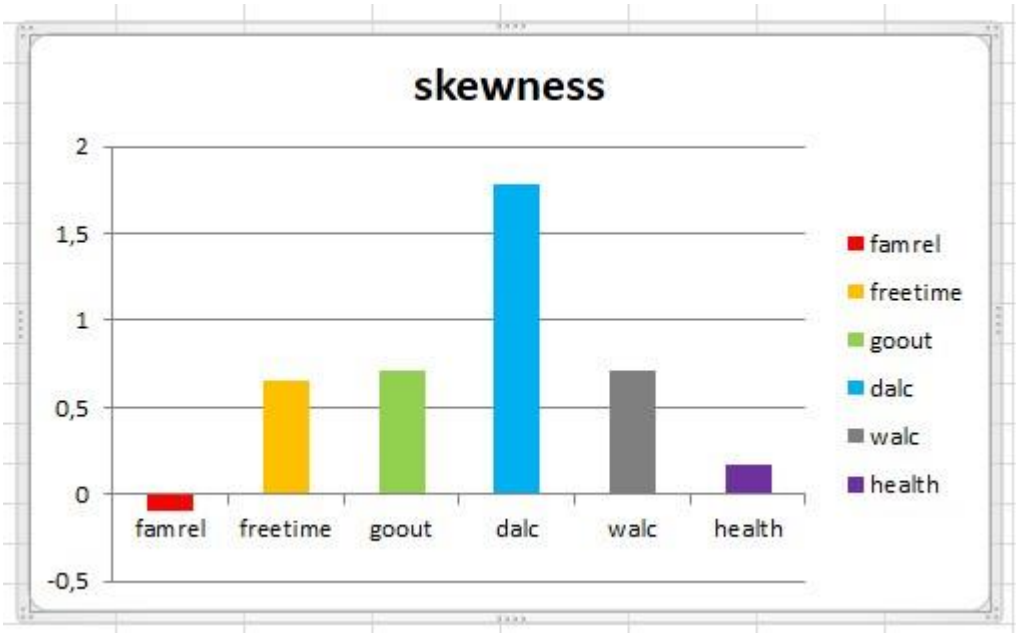
Gambar II.4. Grafik Nilai *Interquarte Range* Pada *Sample*

II.4.5.*Skewness*

Skewness merupakan bentuk pengukuran yang menunjukan tingkat kemiringan kurva yang apabila kurva lebih melenceng ke arah kanan (dilihat dari *mean*) maka ia dikatakan menceng positif jika sebaliknya maka disebut menceng negatif.

$$\beta 2 = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \left(\frac{n-\mu}{\sigma}\right)^3$$

n yang merupakan input, sedangkan N merupakan jumlah banyak *sample* dari data n yang berawal dari 1, μ merupakan *Mean* dari n dan σ merupakan standar deviasi dari n.

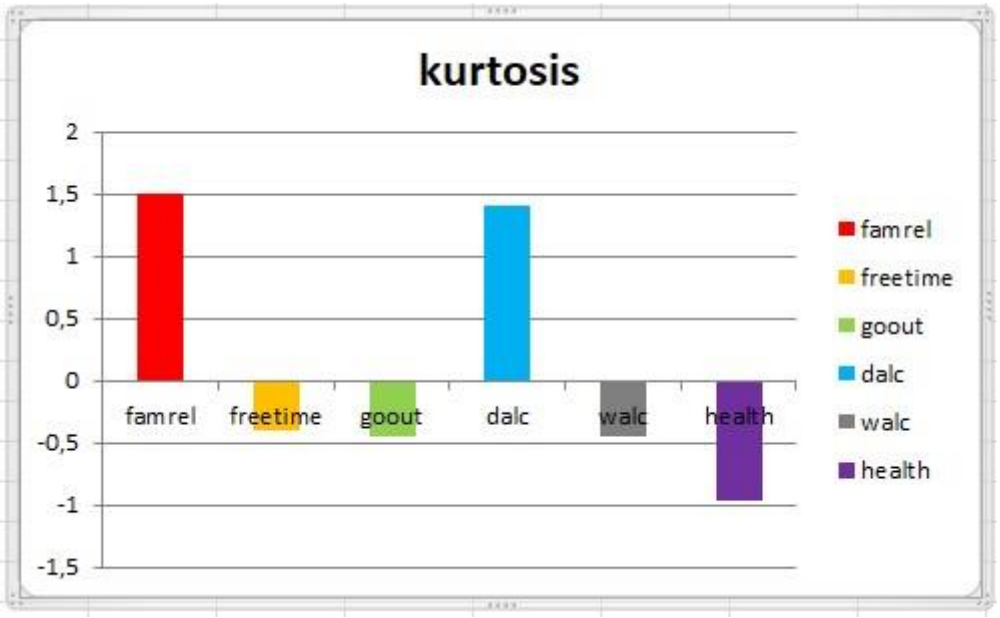


Gambar II.5. Grafik Nilai *Skewness* Pada *Sample*

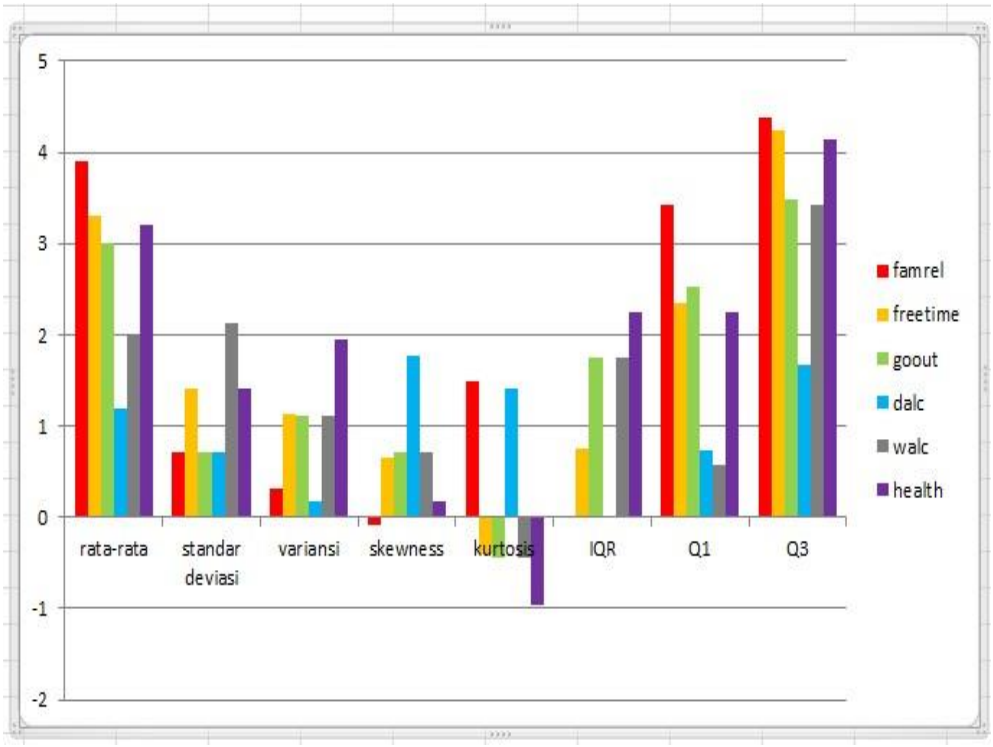
II.4.6.Kurtosis

Kurtosis merupakan bentuk pengukuran yang menunjukan tingkat keruncingan dari relative kurva.

$$\beta 1 = \sum_{i=1}^{N} \left(\frac{n-\mu}{n\sigma}\right)^4$$



Gambar II.6. Grafik Nilai Kurtosis Pada *Sample*



Gambar II.7. Grafik Metode DWT Pada *Sample*

II.5. Metode KNN *(K-Nearest Neighbor)*

KNN *(K-Nearest Neighbor)* merupakan salah satu kelompok *instance based learning* yang algoritmanya termasuk katagori *lazy learning*. Prinsip dasar KNN adalah mencari nilai K dimana nilai K adalah jumlah data terdekat (mirip) dari data baru atau *data testing* (Wu et al., 2008). Kedekatan jarak suatu data dengan data lain merupakan kunci dari algoritma *K-Nearest Neighbor* yang biasanya dihitung menggunakan perhitungan *Euclidean distance.*

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2}$$

Dimana: x_i = Data sampel

X_j = Data uji

r = Variabel data

n = Dimensi data

Langkah-langkah yang harus dilakukan pada klasifikasi *K-Nearest Neighbor* yaitu :

1. Tentukan nilai k, dimana k adalah jumlah tetangga terdekat.
2. Hitung jarak antara data citra baru dan data pelatihan menggunakan metode *Euclidean distance.*
3. Urutkan jarak dan tentukan tetangga terdekat berdasarkan jarak k minimum
4. Cek *output* atau label pada masing-masing kelas tetangga terdekat.
5. Klasifikasikan citra baru ke mayoritas kelas terdekat.

II.6. *RapidMiner*

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining, text mining* dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *RapidMiner* memiliki

kurang lebih 500 operator *data mining*, termasuk operator untuk *input, output, data preprocessing* dan visualisasi. *RapidMiner* merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya sendiri.

RapidMiner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi. *RapidMiner* sebelumnya bernama YALE (*Yet Another Learning Environment)*, dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit dari University of Dortmund*. *RapidMine*r didistribusikan di bawah lisensi AGPL *(GNU Affero General Public License*) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan mengunakan *RapidMiner* di lebih dari 40 negara. *RapidMiner* sebagai *software open source* untuk *data mining* tidak perlu diragukan lagi karena *software* ini sudah terkemuka di dunia. *RapidMiner* menempati peringkat pertama sebagai *software data mining* pada polling oleh KDnuggets, sebuah portal data-mining pada 2010-2011.

RapidMiner menyediakan GUI (*Graphic User Interface*) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan file XML (*Extensible Markup Language)* yang mendefenisikan proses analitis keingginan pengguna untuk diterapkan ke data. File ini kemudian dibaca oleh *RapidMiner* untuk menjalankan analis secara otomatis. *RapidMiner* memiliki beberapa sifat sebagai berikut:

1. Ditulis dengan bahasa pemrograman Java sehingga dapat dijalankan di berbagai sistem operasi.

2. Proses penemuan pengetahuan dimodelkan sebagai *operator trees*
3. Representasi XML internal untuk memastikan format standar pertukaran data.
4. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
5. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
6. Memiliki GUI, command line mode, dan Java API yang dapat dipanggil dari program lain.

Beberapa Fitur dari *RapidMiner*, antara lain:

1. Banyaknya algoritma *data mining*, seperti *decision tree* dan *self-organization map.*
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.
3. Banyaknya variasi *plugin*, seperti *text plugin* untuk melakukan analisis teks.
4. Menyediakan prosedur *data mining* dan *machine learning* termasuk: ETL (*extraction, transformation, loading*), data *preprocessing*, visualisasi, *modelling* dan evaluasi.
5. Proses *data mining* tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI.
6. Mengintegrasikan proyek *data mining* Weka dan statistika.