**KLASIFIKASI PASIEN TERKENA BREAST CANCER MENGGUNAKAN METODE MACHINE LEARNING**

**Adryan Tjengharwidjaja1, Brando Dharma S.2, Yagyu Munenori Michael Emmanuel3**

Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara,

Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia

*E-mail:* *1adryan.535180052@stu.untar.ac.id**,* *2**brando.535180079@stu.untar.ac.id**,* *3yagyu.535180067@stu.untar.ac.id*

**Abstrak**

Kanker payudara adalah salah satu keganasan terbanyak dan memiliki angka kematian cukup tinggi pada wanita. Menurut data World Health Organization (WHO) pada tahun 2020, terdapat 2,3 juta wanita yang terdiagnosis kanker payudara dan 685,000 kematian secara global. Hingga akhir tahun 2020, ada 7,8 juta wanita hidup yang didiagnosis menderita kanker payudara dalam 5 tahun terakhir, menjadikannya kanker paling umum di dunia. Ada lebih banyak tahun hidup yang disesuaikan dengan disabilitas yang hilang (DALYs) oleh wanita karena kanker payudara secara global daripada jenis kanker lainnya. Kanker payudara terjadi di setiap negara di dunia pada wanita pada usia berapa pun setelah pubertas tetapi dengan tingkat yang meningkat di kemudian hari. Kematian akibat kanker payudara tidak banyak berubah dari tahun 1930-an hingga 1970-an. Perbaikan dalam kelangsungan hidup dimulai pada 1980-an di negara-negara dengan program deteksi dini yang dikombinasikan dengan berbagai cara pengobatan untuk memberantas penyakit infasif. Untuk mencari solusi supaya bisa mengurangi korban jiwa yang diakibatkan oleh kanker payudara, maka kami ingin membuat sebuah sistem yang bisa mengklasifikasi apakah seseorang terkena kanker payudara yang termasuk Benign (tumbuh perlahan, tidak menyebar) atau Malignant (ganas, dan bisa menyebar) dengan beberapa metode machine learning. Data yang digunakan untuk penelitian ini berasal dari Wisconsin yang didapat dari UCI Machine Leaning Repository. Metode yang akan digunakan adalah Logistic Regression, Random Forest dan Decision Tree. Dari ketiga metode tersebut akan dibandingkan manakah model yang memberikan nilai akurasi paling bagus. Setelah melakukan uji coba terhadap dataset test, hasil penelitian ini menunjukkan bahwa dari ketiga metode tersebut, Decision Tree memberikan nilai akurasi paling tinggi, posisi kedua adalah Random Forest, dan hasil paling kecil adalah Logistic Regression. Kesimpulan yang didapat adalah ketiga metode machine learning ini dapat digunakan dan diterapkan ke dalam kasus prediksi klasifikasi kanker payudara benign atau malignant, dan decision tree memberikan hasil paling tinggi. Dengan pengetahuan ini, orang yang terkena kanker payudara bisa mengambil keputusan mengenai cara mengatasi kanker tersebut.

**Kata kunci***-Breast Cancer, Logistic Regression, Decision Tree, Random Forest Classifier*

***Abstract***

*Breast cancer is one of the most common malignancies and has a high mortality rate in women. According to the World Health Organization (WHO) in 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. By the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the last 5 years, making it the most common cancer in the world. There are more disability-adjusted life years lost (DALYs) by women due to breast cancer globally than any other type of cancer. Breast cancer occurs in every country in the world in women at any age after puberty but with increasing rates later in life. Breast cancer mortality did not change much from the 1930s to the 1970s. Improvements in* *survival began in the 1980s in countries with early detection programs combined with various means of treatment to eradicate the invasive disease. To find a solution to reduce the death toll caused by breast cancer, we want to create a system that can classify whether someone has breast cancer that is Benign (grows slowly, does not spread) or Malignant (malignant, and can spread) with several machine learning methods. The data used for this research comes from Wisconsin obtained from the UCI Machine Leaning Repository. The methods that will be used are Logistic Regression, Random Forest, and Decision Tree. Of the three methods will be compared which model provides the best accuracy value. After testing the test dataset, the results of this study show that of the three methods, Decision Tree provides the highest accuracy value, the second position is Random Forest, and the smallest result is Logistic Regression. The conclusion is that these three machine learning methods can be used and applied to the case of predicting benign or malignant breast cancer classification, and decision tree gives the highest result. With this knowledge, people affected by breast cancer can make decisions about how to overcome the cancer.*

***Keywords--*** *Breast Cancer, Logistic Regression, Decision Tree, Random Forest Classifier*

**1. PENDAHULUAN**

Latar Belakang

Kanker payudara adalah salah satu keganasan terbanyak dan memiliki angka kematian cukup tinggi pada wanita. Menurut data World Health Organization (WHO) pada tahun 2020, terdapat 2,3 juta wanita yang terdiagnosis kanker payudara dan 685,000 kematian secara global. Hingga akhir tahun 2020, ada 7,8 juta wanita hidup yang didiagnosis menderita kanker payudara dalam 5 tahun terakhir, menjadikannya kanker paling umum di dunia. Ada lebih banyak tahun hidup yang disesuaikan dengan disabilitas yang hilang (DALYs) oleh wanita karena kanker payudara secara global daripada jenis kanker lainnya. Untuk mencari solusi supaya bisa mengurangi korban jiwa yang diakibatkan oleh kanker payudara, maka kami ingin membuat sebuah sistem yang bisa mengklasifikasi apakah seseorang terkena kanker payudara yang termasuk Benign (tumbuh perlahan, tidak menyebar) atau Malignant (ganas, dan bisa menyebar) dengan beberapa metode machine learning.

Tujuan Sistem

Sistem dibuat untuk mengklasifikasi apakah seseorang terkena kanker payudara yang terdiagnosa Benign atau Malignant dengan menggunakan beberapa metode machine learning untuk dibandingkan tingkat akurasi masing-masing metode.

Manfaat Sistem

Sistem dapat menerapkan *machine learning* dalam kasus nyata seperti *predictive analysis* mengklasifikasi orang yang terkena kanker payudara. Selain itu, informasi yang didapat bisa membantu pasien yang terkena kanker payudara dan dapat membantu mengambil keputusan untuk menanganinya.

Penelitian Relevan

Penelitian ini didukung dengan beberapa studi literartur penelitian relevan yang membahas mengenai permasalahan yang di angkat atau menggunakan algoritma yang sama. Berdasarkan hasil eksplorasi, berikut telah ditemukan beberapa penelitian relevan. Penelitian relevan pertama berjudul “*Predicting Football Match Results with Logistic Regression*” oleh Darwin Prasetio dan Dra. Harlili, M.Sc. (2016) [1]. Pada penelitian ini, banyak orang yang masih bingung untuk memprediksi tim mana yang menang berdasarkan performa setiap match, kekuatan tim, kemampuan tim, dll. Penelitian ini menggunakan metode *Logistic Regression* untuk memecahkan masalah ini menggunakan data dari *Barclays’ Premier League* dan soffia.com dengan menggunakan 4 variabel, yaitu *Home Offense, Home Defense, Away Offense* dan *Away Defense*

Penelitian relevan kedua berjudul “*Covid Symptom Severity using Decision Tree*” oleh Naim Rochmawati, dkk. (2020) [2]. Pada penelitian ini, banyak orang yang masih mengganggap virus COVID-19 yang memiliki gejala batuk-pilek sebagai gejala biasa. Oleh karena itu, untuk meningkatkan kewaspadaan, digunakanlah *Decision Tree* untuk membuat solusi dengan menggunakan dataset J48.

Penelitian relevan ketiga berjudul “*Automatic Fruit Classification using Random Forest Algorithm*” oleh Hossam M. Zawbaa, Maryam Hazman, dkk. (2014) [3]. Pada penelitian ini, banyak mesin-mesin yang belum bisa mengklasifikasikan secara akurat jenis gambar buah yang akan diuji. Tujuan dari penulisan jurnal ini adalah untuk mengklasifikasi gambar buah secara akurat menggunakan algoritma Random Forest. Pengujian ini menggunakan 178 gambar buah dan menggunakan 2 metode sebagai perbandingan, yaitu KNN dan *Random Forest Classifier.*

Penelitian relevan keempat berjudul “*Sentiment classification on Big Data using Naïve* *Bayes and Logistic Regression*” oleh Prabhat, Anjuman., and Vikat Khullar. (2017) [4]. Penelitian ini dilakukan untuk mengklasifikasikan sentimen atau yang dikenal sebagai ulasan pada *twitter.* Pengujian dilakukan menggunakan *Naïve Bayes* dan Regresi Logistik. Dalam pekerjaan penelitian ini, kami telah menggunakan *twitter* waktu nyata *review*. Kumpulan data ini berisi jumlah yang sama dari ulasan positif dan negatif yang membuat mesin algoritma pembelajaran mudah untuk mengklasifikasikan ulasan dengan tingkat akurasi 66.667% pada *Naïve Bayes* dan 76.767% pada Regresi Logistik.

**2. METODE PENELITIAN**

## 2.1 Data

Data yang digunakan adalah data *Breast Cancer*. Pada dataset ini, target nya adalah atribut Diagnosis yang terbagi menjadi *Benign* dan *Malignant*. Kami memperoleh dataset *Breast Cancer* ini melalui situs web Kaggle, berikut adalah link nya <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>. Data pada web Kaggle juga bisa ditemukan melalui link berikut ini, [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29](https://archive.ics.uci.edu/ml/datasets/Breast%2BCancer%2BWisconsin%2B%28Diagnostic%29) Setelah mendapatkan dataset, harus dilakukan pra-pemrosesan data terlebih dahulu, contohnya seperti melakukan perubahan format data atau mencari apakah terdapat *missing values* dan menentukan atribut apa saja yang akan dipakai untuk dilakukan uji coba. Pada artikel ini, kami tidak menggunakan seluruh atribut yang ada pada dataset. Terdapat beberapa atribut yang tidak perlu digunakan/berguna untuk perhitungan. Atribut yang dibuang adalah id dan Unnamed: 32. Atribut lain yang masih ada pada dataset akan digunakan untuk analisa dan uji coba.

## 2.1.1 Algoritma

Berikut adalah algoritma-algoritma yang digunakan untuk membuat system:

1. *Algoritma Logistic Regression*

*Logistic Regression* adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara variabel terikat yang memiliki dua kategori atau lebih dengan satu atau lebih peubah bebas berskala kategori atau kontinu [7].

*Logistic Regression* dapat dibagi menjadi :

* *Binary Logistic Regression,*
* *Multinomial Logistic Regression*, dan
* *Ordinal Logistic Regression*

Berikut adalah rumus umum dari *Logistic Regression*:

$Y= β0+ β1X+ℇ $ (1)

Keterangan :

Y = Variabel terikat (Nilai yang akan diuji )

𝛽0 = Konstanta

𝛽1 = Koefisien Regresi

X = Variabel bebas

ℇ = Galat acak

1. *Algoritma* *Decision Tree Classifier*

Metode ini merupakan salah satu metode yang ada pada teknik klasifikasi dalam data mining. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga berguna untuk mengekplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target [8].

Biasanya Algoritma *Decision Tree* menggunakan algoritma C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3. Algoritma C4.5 dan ID3 diciptakan oleh seorang peneliti dibidang kecerdasan buatan bernama j. Rose quinlan pada akhir tahun 1970-an. Algoritma C4.5 membuat pohon keputusan dari atas ke bawah, dimana atribut paling atas merupakan akar, dan yang paling bawah dinamakan daun [8].

Secara umum, algoritma C4.5 untuk membangun sebuah pohon keputusan adalah sebagai berikut:

* Hitung jumlah data, jumlah data berdasarkan anggota atribut hasil dengan syarat tertentu. Untuk proses pertama syaratnya masih kosong.
* Pilih atribut sebagai Node.
* Buat cabang untuk tiap-tiap anggota dari Node.
* Periksa apakah nilai *entropy* dari anggota Node ada yang bernilai nol. Jika ada, tentukan daun yang terbentuk. Jika seluruh nilai entropy anggota Node adalah nol, maka proses pun berhenti.
* Jika ada anggota Node yang memiliki nilai *entropy* lebih besar dari nol, ulangi lagi proses dari awal dengan Node sebagai syarat sampai semua anggota dari Node bernilai nol.

Node adalah atribut yang mempunyai nilai gain tertinggi dari atribut-aribut yang ada. Untuk menghitung nilai gain suatu atribut digunakan rumus seperti yang tertera dalam persamaan berikut:

$Gain\left(S,A\right)=Entropy\left(S\right)-(\sum\_{i=1}^{n}\frac{Ai}{S}.Entropy(Ai))$ (2)

Keterangan:

A : Atribut

n : Jumlah partisi atribut A

Ai : Jumlah kasus pada partisi ke-i.

S : Jumlah kasus.

Sementara itu, untuk menghitung nilai *Entropy* dapat dilihat pada persamaan berikut ini:

$Entropy\left(S\right)=\sum\_{i=1}^{n}-pi. log\_{2}pi $ (3)

Keterangan:

S : Himpunan kasus.

n : Jumlah partisi

Pi : Proporsi dari Si ke S

1. *Algoritma Random Forest Classifier*

Metode random forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* (Breiman 2001). Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas n amatan dan p peubah penjelas, random forest dilakukan dengan cara (Breiman 2001; Breiman & Cutler 2003) [9]:

* Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.
* Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan random feature selection.
* Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon

## 2.2 Rancangan eksperimen

Rancangan eksperimen alur prediksi terkena kanker payudara dapat dilihat pada Gambar 1.



**Gambar 1.** Flowchart Rancangan Eksperiment dari Input sampai End

## 2.2.1 Metode evaluasi

 Metode Evaluasi yang digunakan untuk mengolah data sehingga didapatkan prediksi yang akurat berdasarkan data latih dan data real adalah *K-fold Cross Validation. K-fold Cross Validation* adalah metode statistik yang digunakan untuk membandingkan dan mengevaluasi kinerja model *Machine Learning*. Pada metode evaluasi ini, dilakukan *K-fold Cross Validation* dengan *5-fold. K-fold Cross Validation* yang kami gunakan adalah *Shuffle and Split*. Seperti *K-fold Cross Validation*, di mana pengguna mendefinisikan nilai k untuk jumlah *fold. Proses* pertama-tama mengacak sampel dan kemudian membaginya menjadi sepasang set pelatihan dan pengujian untuk setiap lipatan. Pengguna dapat mengontrol keacakan untuk reproduktifitas.

Kelebihan dan Kekurangan *K-fold Cross Validation*

Ada beberapa keuntungan dari *K-fold Cross Validation* dengan teknik *over validation*. Ada beberapa kelemahan juga. Kelebihan dan kekurangan dapat dilihat pada penjelasan berikut :

Kelebihan

* Akurasi Model yang Lebih Baik. Dengan menggunakan *K-fold Cross Validation*, akan didapatkan model yang lebih akurat daripada hanya menggunakan pemisahan acak dari kumpulan data ke dalam rangkaian kereta dan pengujian.
* Mengurangi *Overfitting*. Saat menggunakan *Cross Validation*, model ini dilatih dan diuji secara ketat di sepanjang jalan. Jadi, data yang Anda berikan ke model akan didistribusikan dengan cara yang lebih tepat daripada sekadar metode pelatihan dan pengujian. Ini akan membuat model tidak terlalu pas dengan rangkaian kereta, yang pada akhirnya memberikan peningkatan kinerja pada data yang tidak terlihat.
* Penyetelan *Hyperparameter* yang Lebih Baik. Metode penyetelan hiperparameter seperti pencarian grid dan pencarian acak dengan *K-fold Cross* Validation menjadi berlipat lebih kuat daripada tanpa validasi silang. Untuk itu harus digunakan metode penyetelan hyperparameter dengan validasi silang untuk kinerja yang lebih baik.
* Ekstraksi Fitur yang Lebih Baik. *Cross Validation* dapat digunakan untuk mengekstrak fitur terpenting dari kumpulan data. *Reverse Feature Extraction with Cross-Validation* (RFECV) adalah metode yang menggunakan validasi silang saat mengekstrak fitur terbaik untuk model pembelajaran mesin.
* Model yang Disempurnakan untuk Data yang Tidak Seimbang. Metode *K-fold Validation* sangat berguna untuk data yang tidak seimbang. Karena data dibagi secara merata di antara rangkaian kereta dan pengujian, distribusi data akan seimbang, yang pada akhirnya menghasilkan model yang berkinerja lebih baik.

Kekurangan

* Biaya Komputasi Melakukan *Cross Validation* akan membutuhkan waktu ekstra. Jika memilih metode *Cross Validation* seperti LOOCV untuk sampel data besar, *overhead* komputasi akan tinggi. Tetapi menggunakan validasi silang 5 kali lipat atau 10 kali lipat tidak akan memakan banyak waktu. Dan performanya akan cukup memuaskan.
* Buruk dengan Data Sekuensial. Jika bekerja dengan data sekuensial seperti data deret waktu, validasi silang k kali lipat adalah pilihan yang buruk. Karena tidak bekerja dengan baik dengan data sekuensial karena sifatnya. Dalam deret waktu, perlu memprediksi nilai masa depan berdasarkan serangkaian nilai masa lalu dari data Anda. Di bawah batasan ini, k fold akan gagal bekerja dengan baik. Tetapi dapat menggunakan pemisahan deret waktu, variasi dari k kali lipat, untuk memvalidasi silang model deret waktu.

**3. HASIL DAN PEMBAHASAN**

Dari pra-pemrosesan data yang telah kami lakukan, terdapat 10 *feature* yang bisa digunakan untuk uji coba perhitungan, supaya bisa memilih atribut manakah yang paling tinggi untuk menentukan apakah seseorang terdiagnosa kanker, kami akhirnya membuat sebuah plot nilai rata-rata dari feature seperti Gambar 2 dan Gambar 3 seperti berikut:



**Gambar 2** Plot nilai *mean* *feature.*



**Gambar 3** Plot nilai *mean feature* lanjutan

Berdasarkan observasi plot diatas, nilai means dari cell radius, perimeter, area, compactness, concavity dan concave points bisa digunakan dalam klasifikasi kanker. Nilai yang lebih besar dari parameter-parameter ini cenderung menunjukkan korelasi dengan tumor malignant. Nilai mean dari texture, smoothness, symmetry atau fractual dimension tidak menunjukkan preferensi khusus dari satu diagnosis dari satunya. Dalam histogram tidak ditemukan outlier besar yang terlihat yang memerlukan pembersihan lanjutan. Dari analisa tersebut maka kami memutuskan untuk menggunakan 5 *feature* saja (radius, perimeter, area, compactness, concave point).

Setelah mengetahui *feature* apa saja yang akan digunakan dalam perhitungan, maka kami mulai mempersiapkan data supaya bisa di proses. Pertama kami membagi data menjadi *data train* dan *data test* dengan rasio pembagian data sebesar 70:30 seperti Gambar 4 berikut:



**Gambar 4.** Code membagi data menjadi *data train* dan *data test*

Berikutnya kami membangun model klasifikasi, kode nya seperti Gambar 5 berikut:



**Gambar 5** Membuat model klasifikasi

Untuk mengetahui seberapa baiknya model *Logistic Regression, Decision Tree*, dan *Random Forest*, maka kami menggunakan *K-Fold Cross Validation* untuk mengevaluasi ketiga metode tersebut, untuk artikel ini kami membaginya menjadi 5-fold. Akhirnya, kita bisa melakukan perhitungan nilai akurasi prediksi masing-masing metode dengan 5 *predictor* yang telah ditetapkan sebelumnya. Hasil nilai akurasi prediksi dapat dilihat pada Gambar 6:



**Gambar 6.** Hasil perhitungan akurasi ketiga metode

**4. KESIMPULAN**

Dari uji coba yang didapat, dapat dilihat bahwa ketiga metode *machine learning* ini bisa melakukan prediksi apakah seseorang terdiagnosa kanker payudara Benign atau Malignant. Akurasi yang dihasilkan tidak berbeda jauh antara satu dengan yang lainnya, namun tetap perlu diperhatikan bahwa untuk mengevaluasi bagus atau tidaknya suatu model harus memperhatikan nilai metode evaluasi yang ada. Pada hasil uji coba cross validation pada *data test*, jika diamati dengan seksama, metode *Random Forest* memiliki nilai yang terbaik dibandingkan dua metode lain. Walaupun metode *Decision Tree* memiliki nilai akurasi prediksi yang sangat tinggi hingga 100%, tapi kami percaya ini merupakan kasus *overfitting*. Metode *Logistic Regression* adalah metode yang memberikan hasil paling kecil dibandingkan *Random Forest* dan *Decision Tree*.

Dengan demikian kesimpulan yang dapat kami buat adalah metode *machine learning* *Logistic Regression, Decision Tree* dan *Random Forest* bisa diterapkan dalam kasus nyata dimana pada kasus ini untuk melakukan klasifikasi orang yang terkena kanker payudara, dan metode yang memberikan hasil terbaik adalah *Random Forest*. Kami mengetahui bahwa artikel ini masih jauh dari sempurna karena waktu yang tidak banyak, untuk pengembangan bisa dimulai daripenambahan UI dan dataset lain supaya tidak terjadi *overfitting* dalam perhitungan, hal lain yang dapat dilakukan bisa berupa pembuatan constraint terhadap parameter atau dilakukan pemangkasan (*pruning*).

**DAFTAR PUSTAKA**

[1] Prasetio, Darwin., dan Dra. Harlili, M. Sc., “P*redicting Football Match Results with Logistic Regressio*n”. [*International Conference on Advanced Informatics: Concepts, Theory And*  *Application* (ICAICTA)](https://ieeexplore.ieee.org/xpl/conhome/7786033/proceeding), 2016.

[2] Rochmawati, Naim dkk., “*Covid Symptom Severity using Decision Tree*”. [*Third International Conference on Vocational Education and Electrical Engineering (ICVEE)*](https://ieeexplore.ieee.org/xpl/conhome/9243082/proceeding)*, 2020.*

[3] Zawbaa, Hossam M. dkk., “*Automatic Fruit Classification using Random Forest Algorithm*”. [14th *International Conference on Hybrid Intelligent Systems*](https://ieeexplore.ieee.org/xpl/conhome/7076842/proceeding), 2014.

[4] Prabhat, Anjuman., and Vikat Khullar, “*Sentiment classification on Big Data using Naïve Bayes* and *Logistic Regression”.* [*International Conference on Computer Communication and Informatics (ICCCI)*](https://ieeexplore.ieee.org/xpl/conhome/8106917/proceeding), 2017.

[5] Dr. William H. Wolberg. (2016). Kaggle, UCI. *Breast Cancer Wisconsin* (Diagnostic) Data Set. Diakses pada 25 Oktober 2021, dari [https://www.kaggle.com/uciml/breast-cancer- wisconsindata/code?datasetId=180&searchQuery=Logistic+Regression](https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/code?datasetId=180&searchQuery=Logistic+Regression).

[6] WHO. *Breast Cancer*. (2020). Diakses pada 01 November 2021, dari <https://www.who.int/news->[room/fact-sheets/detail/breast-cancer](https://www.who.int/news-room/fact-sheets/detail/breast-cancer).

[7] Tampil, Yumira Adriani, Komalig, Hanny, Langi, Yohanis, “Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado”. JdC, Vol. 6, No. 2, September 2017.

[8] Achmad, Budanis Dwi Meilani dan Slamat, Fauzi.” Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree”. Jurnal IPTEK Vol 16 No.1 Mei 2012.

[9] Dewi, Nariswari Karina, Syafitri, Utami Dyah, Mulyadi, Soni Yadi. “PENERAPAN METODE RANDOM FOREST DALAM DRIVER ANALYSIS”. Forum Statistika dan Komputasi, April 2011.