**KLASIFIKASI *POLYCYSTIC OVARY SYNDROME* BERDASARKAN CITRA USG MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* DAN *NAÏVE BAYES* UNTUK MEMBANTU MENDETEKSI KESUBURAN WANITA**

**Tugas Akhir**

**Diajukan untuk memenuhi salah satu syarat memperoleh gelar sarjana   
dari Program Studi Teknik Informatika  
Fakultas Informatika  
Universitas Telkom**

**Nanda Budi Prayuga**

**1103134395**



**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**2017**

# **LEMBAR PERNYATAAN**

Dengan ini saya menyatakan bahwa Tugas Akhir dengan judul “**Klasifikasi *Polycystic Ovary Syndrome* berdasarkan Citra USG menggunakan *Principal Component Analysis* dan *Naïve Bayes* untuk Membantu Mendeteksi Kesuburan Wanita**” beserta seluruh isinya adalah murni karya saya sendiri dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, saya siap menanggung resiko atau sanksi yang dijatuhkan kepada saya apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini.

Bandung, Agustus 2017

Yang membuat pernyataan,

Nanda Budi Prayuga

# **LEMBAR PENGESAHAN**

**KLASIFIKASI *POLYCYSTIC OVARY SYNDROME* BERDASARKAN CITRA USG MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* DAN  *NAÏVE BAYES* UNTUK MEMBANTU MENDETEKSI KESUBURAN WANITA**

**POLYCYSTIC OVARY SYNDROME CLASSIFICATION BASED ON USG IMAGE USING PRINCIPAL COMPONENT ANALYSIS AND NAÏVE BAYES TO HELP WOMAN FERTILITY DETECTION**

**Nanda Budi Prayuga**

**NIM: 1103134395**

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian dari syarat memperoleh gelar pada Program Studi Sarjana Teknik Informatika  
Fakultas Informatika  
Universitas Telkom

Bandung, Agustus 2017

Menyetujui

Pembimbing 1

Prof. Adiwijaya  
NIP. 00740196-1

Ketua Program Studi  
Sarjana Teknik Informatika

Dr. Moch. Arif Bijaksana, Ir. M. Tech.   
NIP. 03650312-4

Pembimbing 2

Mohamad Syahrul Mubarok  
NIP. 10830757-3

# **ABSTRAK**

*Polycystic Ovary Syndrome* (PCOS) adalah kelainan sindrom yang diderita wanita di sistem reproduksinya, seseorang dikatakan menderita *Polycystic Ovary Syndrome* (PCOS)jika ada lebih dari 12 follicle berukuran 2-9 mm atau bertambah besarnya volume follicle di ovarium hingga lebih dari 10 cm3[3]. Saat ini untuk mendeteksi *Polycystic Ovary Syndrome* (PCOS)dokter harus melakukan scan USG, dan secara manual menghitung jumlah follicle yang ditandai dengan area hitam di gambar. Pada penelitian sebelumnya [1, 3, 5] hanya berfokus pada peningkatan kualitas citra dan juga pendeteksian ukuran dan jumlah *follicle* untuk mempermudah tenaga medis melihat *follicle* dan menentukan diagnosis pasien. Sehingga saat ini dokter membutuhkan suatu sistem yang dapat membantunya dalam mendiagnosis *Polycystic Ovary Syndrome* (PCOS)secara otomatis berdasarkan citra USG untuk pendeteksian kesuburan wanita.

Pada tugas akhir ini dibangun sebuah sistem klasifikasi dengan menggunakan kombinasi metode *Principal Component Analysis* (PCA) yang berfungsi sebagai dimensi reduksi dan *Naïve Bayes* yang merupakan salah satu turunan dari *Bayesian Network* sebagai *classifier*nya. *Principal Component Analysis* (PCA) dipilih karena dapat mereduksi dimensi dari data set tanpa menghilangkan informasi yang ada didalamnya, sedangkan *Naïve Bayes* dipilih karena keefektifannya dalam melakukan klasifikasi, walaupun dalam teorinya diasumsikan tiap fitur independen (naïve).

Dari hasil pengujian menggunakan metode *k-fold* *cross validation* dengan k=8 dan pengujian dilakukan sebanyak 50x pengujian, dapat dilihat sistem yang dibangun dengan menggunakan metode *Principal Component Analysis* (PCA) dan *Naïve Bayes,* memiliki performansi rata-rata *F1 Score* tertinggi sebesar 84.76%, dengan parameter uji jumlah distribusi data ditiap kelas pada data training masing-masing 40 gambar, dan jumlah *principal component* sebanyak 53.

Kata Kunci : *Polycystic Ovary Syndrome*, ovarium, citra USG, *follicle*, *Naïve Bayes*, *Principal Component Analysis*, *Cross Validation, Imbalanced Data*, Normalisasi.

# **ABSTRACT**

Polycystic Ovary Syndrome (PCOS) is a disorder syndrome suffered by women in the reproductive system, a woman diagnosed to have Polycystic Ovary Syndrome (PCOS) if there are more than 12 follicles of 2-9 mm or increase in volume of follicles in the ovaries up to more than 10 cm3[3]. Nowadays to detect Polycystic Ovary Syndrome (PCOS), the doctor should perform an ultrasound scan, and manually count the number of follicles marked with the black area in the image, the previous paper [1, 3, 5] only focused on improving image quality and the detection of the size and number of follicles to facilitate medical personnel to see the follicle and determine the patient's diagnosis. Otherwise doctors need a system that can help in diagnosing Polycystic Ovary Syndrome (PCOS) automatically based on ultrasound images for female fertility detection.

In this paper, we propose to build a classification system that help in diagnosing Polycystic Ovary Syndrome (PCOS) automatically based on ultrasound image, using a combination of Principal Component Analysis (PCA) method that serves as reduction dimension and Naïve Bayes which is one derivative of Bayesian Network as its classifier. Principal Component Analysis (PCA) is chosen because it reduces the dimensions of the data set without losing the information it contains, while Naïve Bayes is chosen because of its effectiveness in classification, although in theory it is assumed to be an independent feature (naïve).

 From the test result using k-fold cross validation method with k=8 and 50x testing, showed that the system we build using Principal Component Analysis (PCA) and Naïve Bayes method was succesfully implemented with highest performance of Average F1 Score is 84.76%, with testing parameter: the data amount in each class on the training dataset is 40, and the number of principal component is 53.

Keywords: Polycystic Ovary Syndrome, ovarium, citra USG, follicle, Naïve Bayes, Principal Component Analysis, Cross Validation, Imbalanced Data, Normalization.

# **LEMBAR PERSEMBAHAN**

Puji syukur tidak henti hentinya penulis panjatkan kehadirat Ilahi Rabbi, karena hanya dengan ridhanya tugas akhir ini bisa penulis selesaikan. Pun tak lupa terima kasih penulis ucapkan kepada semua pihak yang telah memberikan semangat, perhatian, doa, dan bantuan dalam proses pengerjaan tugas akhir ini, terlebih khusus penulis sampaikan kepada:

1. Allah SWT Yang Maha Pengasih lagi Maha Penyayang.
2. Nabi Muhammad SAW rahmatan lil alamin yang dengan ajarannya memberikan keteladan kepada seluruh umat manusia, semoga kita termasuk golongan yang beruntung sehingga bisa mendapatkan syafa’at beliau.
3. Kedua orang tua dan keluarga penulis yang selalu memberikan dukungan secara materi dan non materi serta doa yang tiada putusnya kepada penulis. Doa teriring kepada kedua orang tua semoga selalu diberikan kesehatan dan dilancarkan segala urusannya, semoga dengan terselesaikannya tugas akhir ini bisa membuat beliau bangga.
4. Bapak Prof. Adiwijaya sebagai dosen pembimbing 1 yang tak henti-hentinya memberi semangat, membimbing, dan membantu penulis dalam mengerjakan tugas akhir ini.
5. Bapak Mohamad Syahrul Mubarok selaku dosen pembimbing 2 yang dengan sabar membimbing, memberi semangat dan membantu penulis dalam mengerjakan Tugas Akhir ini. Tanpa beliau mungkin tugas akhir ini tidak akan terselesaikan tepat waktu. Terima kasih bapak.
6. Seluruh dosen Universitas Telkom yang telah memberikan ilmunya kepada penulis selama menempuh perkuliahan.
7. Teman teman FIXIS yang selalu memberikan dukungan moral dan menemani perjalan penulis selama berkuliah di Univesitas Telkom. Who we are? We are FIXIS.
8. Keluarga besar ProClub yang menjadi tempat bernaung, diskusi, dan belajar sehingga mendapatkan ilmu yang bisa penulis manfaatkan di dunia kerja.
9. Keluarga besar MyAbuy yang membantu pengerjaan tugas akhir ini, ucapan khusus saya sampaikan kepada Mas Sakti yang dengan sabar memberikan penjelasan saat penulis mendapatkan masalah.
10. Teman teman Percukongan MasBob dan Jos Gandos yang gokil abis, terkhusus kepada David sebagai teman YOLO.
11. Ayaa yang selalu memberikan semangat kepada penulis saat pengerjaan tugas akhir ini.
12. Teman teman lainnya yang tidak bisa penulis tulis satu per satu.

# **KATA PENGANTAR**

Puji syukur tidak henti hentinya penulis panjatkan kehadirat Allah SWT Yang Maha Pengasih lagi Maha Penyayang, karena hanya dengan hidayah dan ridhanya penulis bisa menyelesaikan Tugas Akhir dengan judul “**Klasifikasi *Polycystic Ovary Syndrome* berdasarkan Citra USG menggunakan *Principal Component Analysis* dan *Naïve Bayes* untuk Membantu Mendeteksi Kesuburan Wanita**”.

Penulis menyadari bahwa pengerjaan tugas akhir ini masih jauh dari kata sempurna, baik dari tata bahasa ataupun analisisnya. Oleh karea itu dengan kerendahan hati penulis menerima segala kritik dan saran yang membangun untuk perbaikan di masa mendatang. Penulis juga berharap tugas akhir ini bisa memberi manfaat ataupun inspirasi terhadap pembaca.

Bandung, Agusutus 2017

Nanda Budi Prayuga

# **DAFTAR ISI**

[**LEMBAR PERNYATAAN** ii](#_Toc491111370)

[**LEMBAR PENGESAHAN** iii](#_Toc491111371)

[**ABSTRAK** iv](#_Toc491111372)

[**ABSTRACT** v](#_Toc491111373)

[**LEMBAR PERSEMBAHAN** vi](#_Toc491111374)

[**KATA PENGANTAR** vii](#_Toc491111375)

[**DAFTAR ISI** viii](#_Toc491111376)

[**DAFTAR GAMBAR** x](#_Toc491111377)

[**DAFTAR TABEL** xi](#_Toc491111378)

[**DAFTAR ISTILAH** xii](#_Toc491111379)

[**BAB I PENDAHULUAN** 1](#_Toc491111380)

[**1.1** **Latar Belakang** 1](#_Toc491111381)

[**1.2** **Perumusan Masalah** 1](#_Toc491111382)

[**1.3** **Tujuan** 2](#_Toc491111383)

[**1.4** **Hipotesa** 2](#_Toc491111384)

[**1.5** **Rencana Kegiatan** 2](#_Toc491111385)

[**1.6** **Jadwal Kegiatan** 3](#_Toc491111386)

[**BAB II TINJAUAN PUSTAKA** 4](#_Toc491111408)

[**2.1** **Ovarium** 4](#_Toc491111409)

[**2.2** **Naïve Bayes** 4](#_Toc491111410)

[**2.3** **Imbalance Data** 6](#_Toc491111411)

[**2.4** **Principal Component Analysis (PCA)** 6](#_Toc491111412)

[**2.5** **Normalization** 7](#_Toc491111413)

[**2.6** **Literatur Terkait** 8](#_Toc491111414)

[**BAB III PEMBAHASAN** 9](#_Toc491111415)

[**3.1** **Data Set** 9](#_Toc491111416)

[**3.2** **Gambaran Umum Sistem** 9](#_Toc491111417)

[**3.3** **Perancangan Sistem Klasifikasi** 12](#_Toc491111418)

[3.3.1 Perancangan *Balancing* Data 12](#_Toc491111419)

[3.3.2 Perancangan Normalisasi Data 13](#_Toc491111420)

[*3.3.3* Perancangan *Principal Component Analysis (PCA)* 13](#_Toc491111421)

[3.3.4 Perancangan  *Naïve Bayes* 14](#_Toc491111422)

[3.3.5 Proses Pengujian 15](#_Toc491111423)

[**3.4** **Spesifikasi Kebutuhan Sistem** 15](#_Toc491111424)

[**BAB 4 PENGUJIAN DAN ANALISIS** 17](#_Toc491111425)

[**4.1** **Tujuan Pengujian** 17](#_Toc491111430)

[**4.2** **Pengujian Sistem** 17](#_Toc491111431)

[4.2.1 Skenario Pengujian Sistem 17](#_Toc491111432)

[**4.3** **Analisis** 18](#_Toc491111433)

[**4.3.1** **Pengujian Skenario 1** 18](#_Toc491111434)

[**4.3.2** **Pengujian Skenario 2** 19](#_Toc491111435)

[**4.3.3** **Pengujian Skenario 3** 21](#_Toc491111436)

[**BAB 5 KESIMPULAN DAN SARAN** 23](#_Toc491111437)

[**5.1** **Kesimpulan** 23](#_Toc491111443)

[**5.2** **Saran** 23](#_Toc491111444)

[**DAFTAR PUSTAKA** 24](#_Toc491111445)

[**LAMPIRAN** 25](#_Toc491111446)

[**Contoh Dataset** 25](#_Toc491111447)

[**Tabel Hasil Pengujian** 25](#_Toc491111448)

# **DAFTAR GAMBAR**

[Gambar 2. 1. Ilustrasi model Naive Bayes 12](#_Toc490143882)  
[Gambar 3. 1. Data Set 18](#_Toc490143868)

[Gambar 3. 2 Gambaran Umum Sistem 19](#_Toc490143869)

[Gambar 3. 3. Alur Diagram Balancing Data 21](#_Toc490143870)

[Gambar 3. 4. Alur Diagram Normalisasi Data 22](#_Toc490143871)

[Gambar 3. 5. Alur Diagram PCA 23](#_Toc490143872)

[Gambar 3. 6. Alur Diagram Naïve Bayes 23](#_Toc490143873)

[Gambar 4. 1. Recall, Precision dan F1 Score 28](#_Toc490143983)

[Gambar 4. 2. Recall 29](#_Toc490143984)

[Gambar 4. 3. Precision 29](#_Toc490143985)

[Gambar 4. 4. F1 Score 30](#_Toc490143986)

[Gambar 4. 5. Recall, Precision dan F1 Score 31](#_Toc490143987)

# **DAFTAR TABEL**

[Tabel 3. 1. Confusion Matrix 20](#_Toc490144036)

[Tabel 4. 1. Skenario Pengujian 27](#_Toc490144042)

# **DAFTAR ISTILAH**

|  |  |
| --- | --- |
| **Istilah, Akronim, Singkatan** | **Keterangan** |
| Fitur | Ciri suatu objek |
| *Imbalanced* | Tidak seimbang |
| Dimensi | Aspek yang meliputi atribut yang membentuk suatu entitas |
| *Naïve* | Lugu/bodoh |
| Numpy | Library tambahan untuk python, digunakan untuk mempermudah proses perhitang |
| *Recall* | Tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem |
| *Precision* | Tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. |
| *F1 Score* | Bobot harmonic mean dari recall dan precision |
| Gambar USG | Gambar ultrasonografi |
| *Follicle* | Lubang kecil pada bagian tubuh yang berfungsi untuk menampung sesuatu sesuai dengan tempatn |
| *Variance* | Ukuran seberapa jauh sebuah kumpulan bilangan tersebar |
| *Standar Deviasi* | Ukuran yang digunakan untuk mengukur jumlah variasi atau sebaran sejumlah nilai data |
| *Classifier* | Metode yang digunakan untuk melakukan klasifikasi terhadap suatu objek |

# **BAB I PENDAHULUAN**

## **Latar Belakang**

*Polycystic Ovary Syndrome* (PCOS) adalah kelainan sindrom yang diderita wanita di sistem reproduksinya, seseorang dikatakan menderita *Polycystic Ovary Syndrome* (PCOS) jika ada lebih dari 12 *follicle* berukuran 2-9 mm atau bertambah besarnya volume *follicle* di ovarium hingga lebih dari 10 cm3[2]. Menurut data dari *National Institutes of Health* (NIH) lebih dari 5 juta di USA menderita PCOS. Seseorang yang menderita *Polycystic Ovary Syndrome* (PCOS) akan mengalami masalah pada kesuburannya serta mempunyai hormon androgen dan insulin yang tinggi. Efeknya orang tersebut akan beresiko menderita diabetes tipe 2, kolestrol tinggi, dan tekanan darah tinggi. Sehingga dari data departemen kesehatan Amerika Serikat dibutuhkan lebih dari $4 juta dolar per tahun untuk menganangi permasalahan ini[2]. Karena prosedur penanganan *Polycystic Ovary Syndrome* (PCOS) belum di tentukan, maka penderita *Polycystic Ovary Syndrome* (PCOS) ditangani berdasarkan gejala penyakitnya, bukan menangani penyebab *Polycystic Ovary Syndrome* (PCOS) tersebut.

Saat ini untuk mendeteksi PCOS dokter harus melakukan scan USG dan secara manual menghitung jumlah *follicle* yang di tandai dengan area hitam di gambar. Tentu saja hal ini membutuhkan ketelitian dan pengamatan yang jeli selain ukuran *follicle* yang kecil, *follicle* biasanya tersamarkan dengan obyek lainnya seperti usus atau pembuluh darah. Pada penelitian sebelumnya [1, 3, 5] berfokus pada peningkatan kualitas citra dan juga pendeteksian ukuran dan jumlah *follicle* untuk mempermudah tenaga medis melihat *follicle* dan menentukan diagnosis pasien. Selain itu hasil output dari sistem bersifat implisit, sehingga sistem tidak bisa secara mandiri menentukan deteksi penderita *Polycystic Ovary Syndrome* (PCOS) berdasarkan gambar USGnya.

Oleh karena itu dibutuhkan sistem yang dapat mengklasifikasian *Polycystic Ovary Syndrome* (PCOS) melalui citra USG secara otomatis, sistem yang dibuat diharapkan memiliki performansi yang bagus dengan tingkat performansi yang tinggi, sehingga membantu dokter dalam mendiagnosis *Polycystic Ovary Syndrome* (PCOS) untuk pendeteksian kesuburan wanita.

## **Perumusan Masalah**

Berdasarkan uraian latar belakang masalah, dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana merancang sistem yang dapat mengklasifikasikan *Polycystic Ovary Syndrome* (PCOS) berdasarkan citra USG citra secara otomatis?
2. Bagaimana mengimplementasikan sistem yang dirancang sehingga dapat mengklasifikasikan *Polycystic Ovary Syndrome* (PCOS) berdasarkan citra USG secara otomatis untuk membantu dokter mendeteksi kesuburan wanita?
3. Bagaimana mengetahui performansi hasil klasifikasi dari sistem yang dibuat?

## **Tujuan**

Adapun tujuan yang ingin dicapai dari tugas akhir ini yaitu:

1. Merancang sistem yang dapat mengklasifikasikan *Polycystic Ovary Syndrome* (PCOS) berdasarkan citra USG citra secara otomatis.
2. Mengimplementasikan sistem yang dirancang sehingga dapat mengklasifikasikan *Polycystic Ovary Syndrome* (PCOS) berdasarkan citra USG secara otomatis untuk membantu dokter mendeteksi kesuburan wanita.
3. Mengetahui performansi hasil klasifikasi dari sistem yang dibuat.

## **Hipotesa**

Penelitian sebelumnya [1, 3, 5] berfokus pada peningkatan kualitas citra dan juga pendeteksian ukuran dan jumlah *follicle* untuk mempermudah tenaga medis melihat *follicle* dan menentukan diagnosis pasien. Selain itu hasil output dari sistem bersifat implisit, sehingga sistem tidak bisa secara mandiri menentukan deteksi penderita *Polycystic Ovary Syndrome* (PCOS) berdasarkan gambar USGnya.

Kombinasi penggunaan *Principal Component Analysys* (PCA) dan *Naïve Bayes* sangat berpotensi digunakan dalam pemecahan masalah deteksi *Polycystic Ovary Syndrome* (PCOS). *Principal Component Analysys* (PCA)dipilih karena dapat mereduksi dimensi dari data set tanpa menghilangkan informasi yang ada didalamnya, sedangkan *Naïve Bayes* di pilih karena keefektifannya dalam melakukan klasifikasi, walaupun dalam teorinya di asumsikan tiap fitur independen (naïve). Selain itu *Naïve Bayes* mudah di optimasi dan digabungkan dengan metode-metode lainnya. Hasil akhirnya diharapkan dengan menggunakan metode  *Naïve Bayes* di dapatkan sistem yang powerful, efisien, dan memiliki tingkat performansi yang tinggi untuk membantu dokter dalam mendiagnosis *Polycystic Ovary Syndrome* (PCOS) untuk pendeteksian kesuburan wanita.

## **Rencana Kegiatan**

Rencana kegiatan yang akan dilakukan pada penulisan tugas akhir ini adalah:

* **Kajian Pustaka**

Pada tahap kajian pustaka dilakukan pencarian materi-materi berupa paper, jurnal, buku dan artikel yang berhubungan dengan permasalahan yang diangkat di tugas akhir ini untuk dijadikan referensi. Tujuan dari dilakukannya kajian pustaka adalah untuk memahami permasalahan dan menemukan metode yang tepat untuk menyelesaikan permasalahan yang ditemukan.

* **Pengumpulan Data**

Pada tahap ini data set yang akan digunakan di kumpulkan. Data yang digunakan pada tugas akhir kali ini adalah foto USG rahim dari pasien normal dan pasien penderita *Polycystic Ovary Syndrome* (PCOS). Data set yang digunakan berjumlah 54 buah yang dibagi menjadi dua, yaitu : gambar ovarium positif terkena *Polycystic Ovary Syndrome* (PCOS) dengan jumlah 14 gambar dan negative *Polycystic Ovary Syndrome* (PCOS) dengan jumlah 40 gambar.

* **Analisis dan Perancangan Sistem**

Pada tahap ini dilakukan analisis dan perancangan sistem yang dibuat dalam bentuk flowchart dan block diagram agar mudah untuk dipahami. Pada tahap ini pula dijelaskan gambaran tentang metode yang digunakan dalam menyelesaikan masalah yang didefiniskan di atas.

* **Pengujian**

Pada tahap pengujian, model sistem yang sudah dibangun di uji untuk mengetahui performansinya. Sedangkan model sistem yang dibangun pada tugas akhir ini menggunakan Python.

* **Analisis Hasil Pengujian**

Setelah melalui tahap implementasi dan analisis, maka tahap selanjutnya dilakukan analisis hasil dari metode yang digunakan. Berdasarkan hasil pengujian model sistem dapat dilihat performansi dari model sistem, sehingga hasil akhirnya dapat ditarik kesimpulan.

* **Pembuatan Laporan Tugas Akhir**

Tahap terakhir yatiu pembuatan laporan tugas akhir, dimana hasil implementasi dan analisis didokumentasikan menjadi sebuah dokumen Laporan Tugas Akhir. Tujuan dari adanya dokumentasi agar karya yang dihasilkan dapat tersimpan dan nantinya dapat dijadikan sumber untuk pengembangan selanjutnya.

## **Jadwal Kegiatan**

Tabel 1.1 Jadwal Kegiatan

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Bulan | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 |
| 1. | Kajian Pustaka |  |  |  |  |  |  |
| 2. | Pengumpulan Data |  |  |  |  |  |  |
| 3. | Analisis dan Perancangan Sistem |  |  |  |  |  |  |
| 4. | Pengujian Sistem |  |  |  |  |  |  |
| 5. | Analisis Hasil Pengujian |  |  |  |  |  |  |
| 6. | Pembuatan Laporan |  |  |  |  |  |  |

# **BAB II TINJAUAN PUSTAKA**

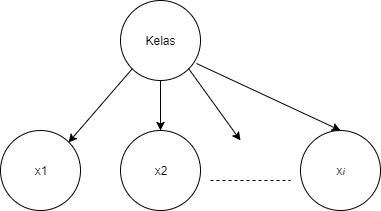
## **Ovarium**

Untuk mengetahui ukuran ovarium normal dapat dilihat saat masa menstruasi, melalui sebuah tes telah dilakukan di Creighton University School Of Medicine didapat hasil dimana ciri-ciri ovarium normal yaitu : ukuran awal *follicle* 2-4 mm, lalu akan terus tumbuh mencapai 10 mm dihari 8-9 dan mencapai ukuran 18-24mm pada hari ke 14[4]. Selain itu ovarium dikatakan normal jika volume *follicle* di ovarium kurang dari 10 cm3.

Sedangkan salah satu kriteria seseorang menderita *Polycystic Ovary Syndrome* (PCOS) adalah ada lebih dari 12 *follicle* berukuran 2-9 mm atau bertambah besarnya volume *follicle* di ovarium hingga lebih dari 10 cm3[2]. Namun jika ada *follicle* dominan dengan ukuran lebih dari 10 mm, harus dilakukan pemindaian dan pengecekan lebih lanjut. Perbedaan paling signifikan antara ovarium normal dengan ovarium penderita *Polycystic Ovary Syndrome* (PCOS) adalah volume dari ovarium. Ovarium normal memiliki volume kurang dari 10 cm3 (7.94 ± 2.34 cm3), sedangkan ovarium penderita *Polycystic Ovary Syndrome* (PCOS) memiliki volume melebihi 10 cm3 (14.04 ± 7.36 cm3)[8, 9].

## ***Naïve Bayes***

*Naïve Bayes* adalah salah satu metode *supervised learning* turunan dari teorema *Bayes* yang digunakan untuk mengklasifikasikan data set berbentuk kontinu. Ciri utama dari algoritma *Naïve Bayes* yaitu adanya asumsi bahwa tiap fitur yang ada di data adalah independen. Untuk mengilustrasikan asumsi tersebut dapat dilihat di Gambar 2.1.



Gambar 2. 1. Ilustrasi model Naive Bayes

Persamaan yang digunakan pada model Naïve Bayes bisa dilihat di bawah ini:

(1)

atau jika dijabarkan seperti di bawah ini

(2)

Penjelasan dari rumus di atas :

adalah *Posterior Probability* data terhadap kelas

adalah *Prior Probability* suatu kelas

= adalah *Likelihood Probability* sebuah fitur terhadap   
kelas C

Dalam prosesnya seperti yang bisa dilihat di persamaan (1), setiap data akan dicari posterior probabilitinya yang kemudian ditentukan label kelas dari data tersebut. Penentuan label dilakukan menggunakan *maximum a posterior probability* (MAP). Penjelasan dari *maximum a posterior probability* (MAP) adalah sebagai berikut.

(3)

Pada persamaan (3) di atas akan di lakukan perhitungan MAP dengan cara menjumlah perhitungan dari *conditional probability* yang di asumsikan independen sebanyak *i* fitur data yang kemudian dikalikan dengan *prior* dari kelas. Nilai tertinggi dari hasil perhitungan akan menentukan label dari data tersebut.

Dari persamaan (3) di butuhkan prior untuk mencari *maximum a posterior probability* (MAP) dari suatu data, sedangkan *prior* sendiri adalah frekuensi kemunculan data dari suatu kelas terhadap jumlah keseluruhan data set. Rumus persamaan *prior* adalah seperti berikut ini.

(5)

dimana adalah jumlah data di tiap kelas c dan *N* adalah jumlah data pada data set.

Metode *supervised learning* menggunakan  *Naïve Bayes* sangat berpotensi digunakan dalam pemecahan masalah deteksi *Polycystic Ovary Syndrome* (PCOS).  *Naïve Bayes* dipilih karena keefektifannya dalam melakukan klasifikasi, walaupun dalam teorinya diasumsikan tiap fitur independen (*naïve*). Selain itu *Naïve Bayes* mudah dioptimasi dan digabungkan dengan metode-metode lainnya. Hasil akhirnya diharapkan dengan menggunakan metode  *Naïve Bayes* didapatkan sistem yang powerful, efisien, dan memiliki tingkat performansi yang tinggi untuk membantu dokter dalam mendiagnosis *Polycystic Ovary Syndrome* (PCOS) untuk pendeteksian kesuburan wanita.

## **Imbalance Data**

*Imbalance Data* adalah permasalah dalam *machine learning* dimana total data antar kelas tidak sama besarnya[10]. Data set kelas yang jumlahnya lebih besar dibandingkan jumlah data set di kelas lainnya disebut *major class,* sedangkan data set kelas yang jumlahnya paling sedikit disebut *minor class*[10]. Permasalahan yang biasanya dihadapi saat membangun sistem klasifikasi dengan data yang tidak seimbang adalah kesalahan klasifikasi terhadap *minor class*.

Untuk itu di butuhkan metode untuk melakukan *balancing data.* Menurut Haibo He, dan Edwardo A. Garcia [7] beberapa solusi yang di dapat dari permasalah tersebut yaitu :

* Random Oversampling
* Random Undersampling
* Synthetic Sampling with Data Generation (SMOTE)
* Adaptive Synthetic Sampling (ADAYS)
* Dll

Proses balancing data sangat penting untuk dilakukan agar performansi dari sistem dapat meningkat, agar kesalahan klasifikasi terhadap *minor class* dapat di turunkan.

## ***Principal Component Analysis* (PCA)**

Menurut J.Shlens [9], *Principal Component Analysis* (PCA) adalah metode non-parametric yang digunakan untuk mengekstrak informasi yang revelan dari dataset yang besar. *Principal Component Analysis* (PCA) juga dapat digunakan untuk mereduksi data yang kompleks ke data yang lebih kecil dimensinya.

Secara garis besar ada empat proses yang garus di lakukan dalam proses PCA, proses itu yaitu sebagai berikut [9]:

1. Lakukan normalisasi data dengan mengurangi semua dimensi dalam data dengan rata-rata dari dimensi data tersebut.
2. Lakukan perhitungan matriks *covariance* untuk mencari hubungan antara dimensi di data sets (biasanya lebih dari 3 dimensi), dimana itu tidak bisa di visualisasikan. Rumus perhitungan covariance bisa di lihat di bawah ini.

(6)

Penjelasan dari rumus di atas :

elemen ke-i dari data pada dimensi *X*

elemen ke-i dari data pada dimensi *Y*

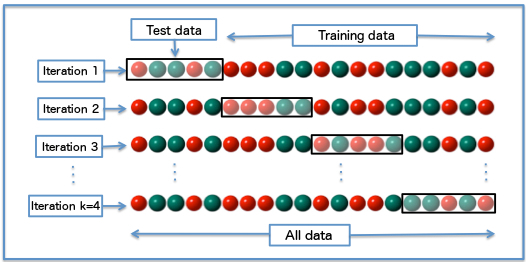
rata-rata yang di dapat dari dimensi *X*

rata-rata yang di dapat dari dimensi *Y*

1. Hitung nilai eigen dan vector eigen dari matriks covariance yang sudah di dapatkan dari langkah sebelumnya
2. Ambil vektor eigen dengan nilai eigen terbesar sebanyak *x principal component* (nilai *x* yang di spesifikasikan oleh user), lalu transformasikan ke domain eigen dengan cara mengalikan dataset dengan vektor eigen. Hasil akhir yang di dapat yaitu dimensi dari data set yang sudah tereduksi dan *feature vector.*

## ***K-Fold Cross Validation***

*K-Fold Cross Validation* adalah metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua bagian, yaitu data *training* dan data *testing* [11]. Seluruh data akan diacak dan dibagi menjadi K buah subset dengan ukuran yang sama. Performansi dari sistem yang dibangun akan dihitung dari rata-rata hasil performansi dari K iterasi. Kelebihan dari metode ini yaitu setiap subset akan menjadi data *testing* sebanyak satu kali dan akan menjadi data training sebanyak K-1 kali sehingga tidak akan ada masalah dalam pembagian data. Sebaliknya kekurangan dari metode ini akan membutuhkan waktu komputasi sebanyak K kali waktu komputasi, karena dilakukan K iterasi untuk mengetahui performansi sistem yang di bangun. Untuk lebih jelasnya dapat dilihat di Gambar 2.2.



Gambar 2. 2 K-Fold Cross Validation (Sumber Wikipedia.org)

## **Normalization**

*Normalization* atau normalisasi adalah proses transformasi dimana sebuah atribut numerik diskalakan dalam range tertentu, seperti -1.0 sampai 1.0, atau 0.0 sampai 255.0 [8]. Ada pun beberapa metode yang di gunakan dalam normalisasi data yaitu:

* Min-Max Normalization

Min-Max Normalization adalah metode normalisasi data yang digunakan dengan cara memetakan value dari data asli ke min-max value yang sudah di definisikan sebelumnya. Rumus perhitungan min-max normalization dapat di lihat di bawah ini.

Penjelasan dari rumus di atas :

nilai baru dari *x*

minimum value dari *x*

maksimum value dari *x*

maksimum value baru dari *x*

maksimum value baru dari *x*

* Z-Score Normalization

Disebut juga *zero-mean normalization*, dimana value dari sebuah atribut *x* dinormalisasi berdasarkan nilai rata-rata dan standar deviasi dari atribut *x* [8]. Rumus perhitungan min-max normalization dapat dilihat di bawah ini.

Penjelasan dari rumus di atas:

nilai baru dari *x*

rata-rata value dari *x*

standar devisasi dari *x*

## **Literatur Terkait**

Berdasarkan penelitian yang dipublikasikan National Institutes of Health (NIH) seseorang dikatakan menderita *Polycystic Ovary Syndrome* (PCOS) jika ada lebih dari 12 *follicle* berukuran 2-9 mm atau bertambah besarnya volume *follicle* di ovarium hingga lebih dari 10 cm3[2]. Saat ini untuk mendeteksi sindrom *Polycystic Ovary Syndrome* (PCOS) dokter harus melakukan scan USG dan secara manual menghitung jumlah *follicle* yang ditandai dengan area hitam di gambar. Metode yang digunakan untuk mendeteksi *Polycystic Ovary Syndrome* (PCOS) bisa menggunakan NIH 1990, Rotterdam 2003, atau AE-PCOS Society 2006. Sedangkan untuk menangani permasalahan *Polycystic Ovary Syndrome* (PCOS) bisa digunakan metode seperti berikut ini:

* + Modifikasi gaya hidup dan mengurangi obesitas
  + Penggunaan metformin, thiazolidinediones, clomiphene dan aromatase inhibitors.
  + Terapi atau operasi
  + Stimulasi ovarian gonadotropin
  + Continuous positive airway pressure (CPAP)

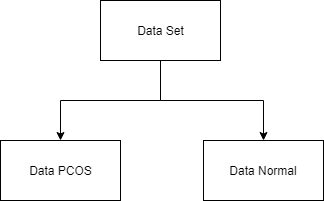
Berbeda lagi dengan paper berjudul *Follicle Detection on the USG Images to Support Determination of Polycystic Ovary Syndrome,* berdasarkan kriteria diatas untuk mendeteksi *Polycystic Ovary Syndrome* (PCOS) digunakan langkah-langkah sebagai berikut: *homogeneous regions, region growing, follicle extraction dan follicle quantification*. Sedangkan di paper berjudul *Particle Swarm Optimization on Follicles Segmentation to Support PCOS Detection* langkah-langkah metode yang digunakan untuk mendeteksi *Polycystic Ovary Syndrome* (PCOS) berdasarkan citra USG yaitu:

* + *Image* *Preprocessing*: *Cropping, Denoising, Contrast Enhancement*.
  + *Image* *segmentation* untuk memisah gambar ke beberapa bagian yang mempunyai area homogen yang sama. Sedangkan metode *Image segmentation* sendiri ada beberapa diantaranya: thresholding, edge-based, region growing and clustering.
  + *Clustering* menggunakan metode *Particle Swarm Optimization* (PSO) yang diimprovisasi dengan *non-parametric fitness function*, untuk meminimalisasi eror nilai *pixel* dan memaksimalkan similaritas gambar berdasarkan persepsi mata manusia.
  + *Follicle Extraction*

Dari percobaan diatas didapat kesimpulan *Particle Swarm Optimization* (PSO) *image clustering* dengan *contrast enhancement preprocessing* akan menghasilkan jarak intracluster dan quantization *error* yang lebih besar daripada metode *Particle Swarm Optimization* (PSO) tanpa contrast enhancement preprocessing. Hasilnya dengan membandingkan koefisien *DICE* dan *ROI*, metode PSO dengan *contrast enhancement preprocessing* mempunyai performansi yang lebih baik karena ukuran *follicle* yang dihasilkan program mendekati ukuran *follicle* aslinya

# **BAB III PEMBAHASAN**

## **Data Set**

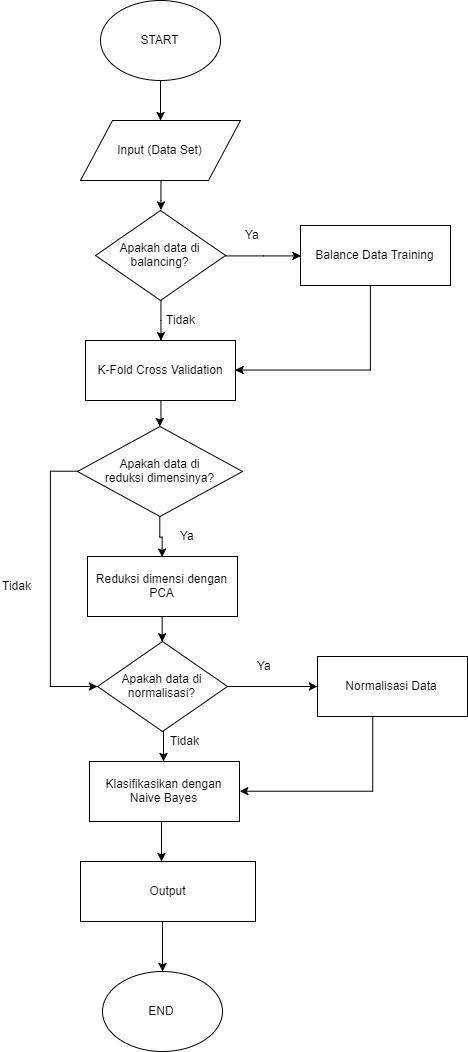


Gambar 3. 1 Data Set

Data set adalah data yang akan digunakan untuk proses pembuatan model dari sistem. Data set berupa gambar USG ovarium yang dibagi menjadi 2 kelas label. Gambar ovarium positif terkena *Polycystic Ovary Syndrome* (PCOS) dengan jumlah 14 gambar dan negatif *Polycystic Ovary Syndrome* (PCOS) dengan jumlah 40 gambar.

## **Gambaran Umum Sistem**

Gambar 3.2 menggambarkan proses secara umum sistem yang akan dibuat

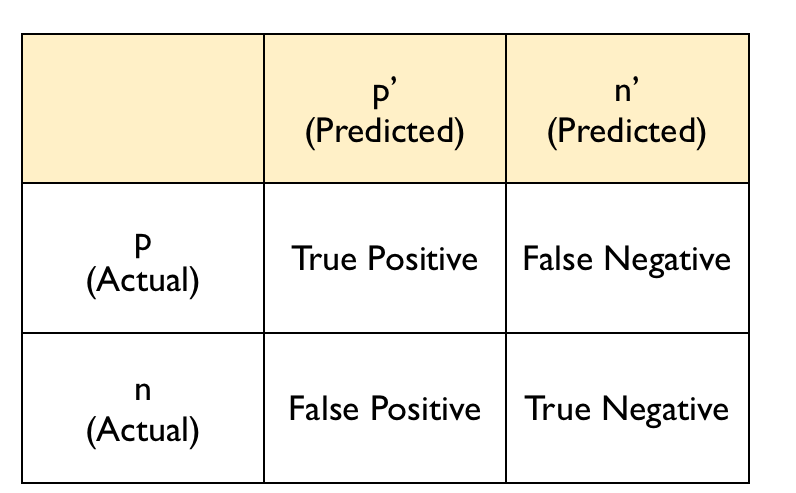


Gambar 3. 2 Gambaran Umum Sistem

Pada dasarnya sistem yang akan dibangun dibagi menjadi lima bagian utama, yaitu: *balancing* data, *K-Fold Cross Validation*, reduksi dimensi, normalisasi data dan klasifikasi. Sebelum melalui kelima proses tersebut data set di bagi menjadi dua bagian yaitu data *Polycystic Ovary Syndrome* (PCOS) dan data Normal di Bab 3.1. Naïve Bayes adalah metode klasifikasi yang sensitif dengan jumlah data set ditiap kelasnya yang harus seimbang. Salah satu masalah yang dihadapi dalam pembuatan sistem ini adalah jumlah data set yang sedikit (54 gambar) dan distribusi data ditiap kelas yang tidak seimbang. Oleh karena itu, agar hasil performansi sistem meningkat, sebelum dilakukan proses PCA dan klasifikasi dengan Naïve Bayes, maka harus dilakukan proses *balancing* data (jumlah data set dalam setiap kelas di seimbangkan).

Tahap selanjutnya setelah dilakukan *balancing* data adalah membagi data set menggunakan metode *K-Fold Cross Validation* seperti yang dijelaskan di Bab 2.5. Setelah itu maka akan memasuki tahap reduksi dimensi menggunakan PCA. Reduksi dimensi menggunakan PCA di lakukan, untuk mendapatkan fitur-fitur penting dari data dengan sesedikit mungkin menghilangkan informasi di dalamnya, selain itu juga mempercepat proses komputasi. Data yang sudah direduksi dimensinya kemudian akan masuk ke tahap klasifikasi menggunakan Naïve Bayes untuk menentukan label dari setiap gambar yang di inputkan.

Untuk mengetahui performansi dari sistem yang sudah dibuat maka perlu dihitung recall, precision, dan f1 measure. Untuk melakukan perhitungan ini di butuhkan *confusion matrix*, seperti dapat dilihat di Tabel 3.3:



Tabel 3. 1. Confusion Matrix

Sedangkan rumus perhitungannya sendiri yaitu :

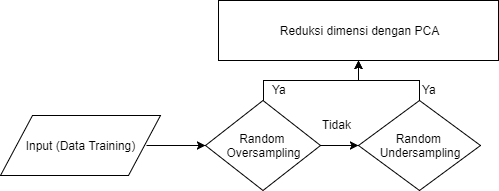
Pada tugas akhir ini akan diobservasi pengaruh penggunaan metode *balancing* data, normalisasi data dan PCA terhadap sistem yang dibangun.

## **Perancangan Sistem Klasifikasi**

Perancangan sistem klasifikasi yang akan dibangun dibagi menjadi empat bagian utama, yaitu: *balancing* data, *K-Fold Cross Validation*, normalisasi data, reduksi dimensi, dan klasifikasi.

### **Perancangan *Balancing* Data**

Perancangan sistem *balancing* datadalam sistem yang akan dibangun dapat dilihat dialur diagram pada Gambar 3.3*.*



Gambar 3. 3. Alur Diagram Balancing Data

Sebelum dilakukan proses reduksi dimensi dengan PCA, input data set akan di*balancing* dengan salah satu dari dua metode di tersebut (Gambar 3.3). Menurut Haibo He, dan Edwardo A. Garcia [7], penjelasan dari masing-masing metode di atas (Gambar 3.3) adalah sebagai berikut.

1. Random Oversampling

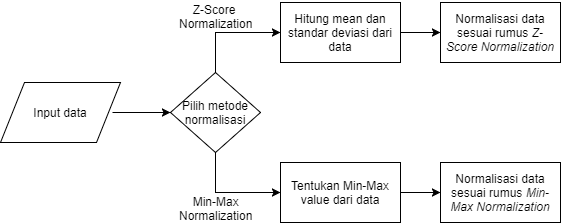
Random Oversampling adalah suatu mekanisme untuk menambah data set *minor class* dengan cara memilih data yang tersedia dari *minor class* secara acak, lalu salin data yang sudah di pilih dan tambahkan ke data set *minor class.*

1. Random Undersampling

Kebalikan dari random oversampling, random undersampling adalah suatu mekanisme untuk mengurangi data set *mayor class* dengan cara memilih data yang tersedia dari *mayor class* secara acak, lalu hilangkan data yang sudah di pilih dari data set *mayor class.*

### **Perancangan Normalisasi Data**

Perancangan sistem normalisasidatadalam sistem yang akan dibangun dapat dilihat dialur diagram pada Gambar 3.4*.*



Gambar 3. 4. Alur Diagram Normalisasi Data

Dilihat dari alur diagram normalisasi di atas (Gambar 3.4) ada dua metode yang dapat diujicoba dalam sistem yang dibangun. Penjelasan dari masing-masing metode di atas (Gambar 3.3) adalah sebagai berikut.

1. *Z-Score Normalization*

Ada dua langkah yang harus di lakukan dalam *Z-Score Normalization.* Pertama hitung mean dan standar deviasi dari data, langkah kedua yaitu menormalisasi sesuai dengan rumus yang sudah di jelaskan di bab 2.

1. *Min-Max Normalization*

Begitu pula dengan *Min-Max Normalization,* ada dua langkah yang harus di lakukan dalam *Min-Max Normalization.* Pertama hitung mean dan standar deviasi dari data, langkah kedua yaitu menormalisasi sesuai dengan rumus yang sudah di jelaskan di bab 2.

### **Perancangan *Principal Component Analysis (PCA)***

Perancangan sistem *Principal Component Analysis (PCA)* dalam sistem yang akan dibangun dapat dilihat dialur diagram pada Gambar 3.5.

C:\Users\Nanda\Downloads\PCA.png

Gambar 3. 5. Alur Diagram PCA

Penjelasan dari masing-masing proses pada diagram PCA di atas (Gambar 3.5) adalah sebagai berikut.

1. Lakukan Normalisasi Data

Tahap pertama yang harus dilakukan dalam proses PCA yaitu melakukan normalisasi data dengan cara mengurangi semua dimensi dalam data dengan rata-rata dari dimensi data tersebut.

1. Hitung Matriks Covariance

Dalam proses ini perhitungan matriks covariance dilakukan dengan menggunakan rumus yang sudah dijelaskan di bab 2.

1. Hitung Nilai Eigen dan Vektoe Eigen

Hasil dari perhitungan matriks covariance dilangkah sebelumnya kemudian diolah lagi untuk mendapatkan nilai eigen dan vektor eigen nya.

1. Tranformasikan ke Domain Eigen

Setelah nilai eigen dan vektor eigen didapat, lalu dilakukan pengurutan vektor eigen berdasarkan nilai eigen terbesar, yang kemudian di ambil *x principal component* dari vektor eigen tersebut (nilai *x* yang di spesifikasikan oleh user). Langkah terakhir transformasikan matriks original ke domain eigen dengan cara mengalikan dataset dengan vektor eigen dari perhitungan sebelumnya. Hasil akhir yang didapat yaitu dimensi dari data set yang sudah tereduksi dan *feature vector.*

### **Perancangan *Naïve Bayes***

Perancangan sistem  *Naïve Bayes* dalam sistem yang akan dibangun dapat dilihat dialur diagram pada Gambar 3.6.

C:\Users\Nanda\Downloads\Gaussian Naive Bayes.png

Gambar 3. 6. Alur Diagram Naïve Bayes

Penjelasan dari masing-masing proses pada diagram PCA di atas (Gambar 3.5) adalah sebagai berikut.

1. Hitung *Prior* tiap Kelas

Tahap pertama yang harus dilakukan dalam adalah menghitung prior dari tiap kelas dengan cara jumlah data disetiap kelas dibagi denganjumlah data.

1. Hitung *Mean* dan *Variance* tiap Fitur dari Data Training

Dalam proses ini akan dicari *mean* dan *variance* tiap fitur data training yang kemudian akan dipakai diproses berikutnya

1. Hitung *Posterior Probability* Data terhadap Kelas

Proses perhitungan *posterior probability* data terhadap kelas yang dilakukan sudah dijabarkan dan dapat di lihat di bab 2.

1. Beri Label tiap Data menggunakan *Metode Maxium a Posterior Probability (MAP)*

Setelah nilai *posterior probability* data terhadap kelas didapat, lalu dilakukan *labeling* data dengan dengan menggunakan metode *maximum a posterior probability* (MAP) seperti yang sudah di jelaskan di bab 2*.* Nilai tertinggi dari hasil perhitungan akan menentukan label dari data tersebut.

### **Proses Pengujian**

Skenario proses pengujian dalam sistem yang akan dibangun dapat dilihat di bawah ini.

1. Skenario I – Pengujian berdasarkan jumlah distribusi data disetiap kelasnya terhadap sistem yang dibangun
   1. Tujuan: Mengetahui pengaruh jumlah distribusi data disetiap kelasnya terhadap sistem yang dibangun.
   2. Pengukuran: F1 Score Measure
   3. Variabel Peubah: Jumlah distribusi data disetiap kelasnya
   4. Variabel Tetap: Nilai PCA
   5. Cara Pengukuran: Menghitung F1 Score jumlah distribusi data disetiap kelasnya yang telah ditentukan di skenario
2. Skenario II – Pengujian berdasarkan penggunaan PCA dan jumlah *principal components* yang di gunakan.
3. Tujuan: Mendapatkan jumlah *principal components* paling optimal
4. Pengukuran: F1 Score Measure
5. Variabel Peubah : Jumlah *principal components* yang digunakan
6. Variabel Tetap : Jumlah distribusi data disetiap kelasnya
7. Cara Pengukuran : Menghitung F1 Score berdasarkan jumlah *principal component* yang telah ditentukan di skenario
8. Skenario III – Pengujian berdasarkan penggunaan normalisasi terhadap data
9. Tujuan : Mengetahui pengaruh penggunaan normalisasi data
10. Pengukuran : F1 Score Measure
11. Variabel Peubah : Menggunakan normalisasi atau tidak
12. Variabel Tetap : Jumlah distribusi data disetiap kelasnya, Nilai PCA
13. Cara Pengukuran : Menghitung F1 Score berdasarkan penggunaan normalisasi yang telah ditentukan di skenario

## **Spesifikasi Kebutuhan Sistem**

Dalam pengerjaan tugas akhir ini pasti dibutuhkan perangkat keras dan perangkat lunak. Berikut spesifikasi perangkat keras dan perangkat lunak yang digunakan.

**Spesifikasi Perangkat Keras**

Berikut spesifikasi perangkat keras yang digunakan dalam pengerjaan tugas akhir ini:

* + Merk : Asus ROG
  + Model : Asus ROG GL552VW
  + Prosesor : Intel Core i7-6700HQ
  + RAM : 16 GB 1600 MHz DDR4
  + Storage : SSD 128 GB, HDD 1 TB
  + GPU : NVIDIA GForce GTX960M 4GB DDR5

**Spesifikasi Perangkat Lunak**

Berikut spesifikasi perangkat lunak yang digunakan dalam pengerjaan tugas akhir ini:

* + OS : Arch Linux
  + Aplikasi : Microsoft Word 2013, Microsoft Excel 2013, Python 3, Numpy, PIL, sklearn.

# **BAB 4 PENGUJIAN DAN ANALISIS**

Pada bagian ini dipaparkan hasil pengujian terhadap sistem berdasarkan skenario pengujian yang sudah dibuat beserta analisis terhadap pengujian yang sudah dilakukan.



## **Tujuan Pengujian**

Tujuan pengujian pada sistem yang dibangun adalah untuk mengetahui performansi dari sistem berdasarkan skenario pengujian yang sudah dibuat menggunakan data set yang sudah disediakan. Sistem yang dibuat diharapkan memiliki performansi yang bagus dengan tingkat klasifikasi yang tinggi sehingga membantu dokter dalam mendiagnosis *Polycystic Ovary Syndrome* (PCOS) untuk pendeteksian kesuburan wanita.

## **Skenario Pengujian Sistem**

Pengujian terhadap sistem yang dibangun menggunakan skenario-skenario yang sudah dijelaskan di Bab 3. Performansi sistem dapat diukur dengan dengan menghitung *recall*, *precision*, dan *f1 score* untuk setiap skenario terhadap data set yang ada. Hasil performansi dari setiap skenario yang diukur akan dijelaskan di bab ini. Parameter-parameter yang akan diobservasi di sistem yang dibangun yaitu:

1. Jumlah distribusi data disetiap kelasnya pada tahap *balancing data*
2. Penggunaan PCA dan jumlah *principal components* yang digunakan pada tahap dimensi reduksi
3. Penggunaan normalisasi terhadap data set pada tahap *normalization*

Adapun tujuan dari dilakukannya observasi terhadap parameter-parameter di atas adalah:

1. Mengetahui performansi sistem dan menganalisis pengaruh jumlah distribusi data disetiap kelasnya terhadap sistem yang dibangun.
2. Mengetahui performansi sistem dan menganalisis pengaruh penggunaan PCA dan jumlah *principal components* yang digunakan terhadap sistem yang dibangun.
3. Mengetahui performansi sistem dan menganalisis pengaruh penggunaan normalisasi pada data set terhadap sistem yang dibangun.

### Skenario Pengujian Sistem

Skenario pengujian sistem yang dilakukan pada tugas akhir ini dapat dilihat pada table 4.1.

|  |  |  |  |
| --- | --- | --- | --- |
| Skenario ke- | Tujuan | Parameter Uji | Parameter Konstan |
| 1 | Mengetahui performansi sistem dan menganalisis pengaruh jumlah distribusi data disetiap kelasnya terhadap sistem yang dibangun | Jumlah distribusi data tiap kelas: 40-40 (*Random Oversampling*), 14-40, 14-14 (*Random Undersampling*). | Nilai PCA = 60,  K-Fold = 8. |
| 2 | Mengetahui performansi sistem dan menganalisis pengaruh penggunaan PCA dan jumlah *principal components* yang digunakan ditahap dimensi reduksi terhadap sistem yang dibangun. | Jumlah *principal components* yang digunakan:  (1) Tanpa PCA, 50, 60, 70, .., 200.  (2) 45, 46, 47, 48, .., 65. | Jumlah distribusi data tiap kelas 40-40, K-Fold = 8. |
| 3 | Mengetahui performansi sistem dan menganalisis pengaruh penggunaan normalisasi pada data set yang dilakukan ditahap *normalization* terhadap sistem yang dibangun. | Menggunakan normalisasi atau tidak | Nilai PCA = 53,  K-Fold = 8,  Jumlah distribusi data tiap kelas 40-40. |

Tabel 4. 1. Skenario Pengujian

## **Analisis**

Pada bab ini di jelaskan analisis terhadap skenario sistem yang sudah dibangun dan diuji.

### **Pengujian Skenario I**

Pada skenario ini diukur pengaruh jumlah distribusi data disetiap kelasnya terhadap performansi sistem dengan menghitung *recall*, *precision*, dan *f1 score* terhadap data set. Hasil pengujian pada skenario ini dapat dilihat pada Gambar 4.1 dan Tabel 4.2.

Gambar 4. 1. Grafik Recall, Precision dan F1 Score: Observasi Jumlah Distrubusi Data Setiap Kelas

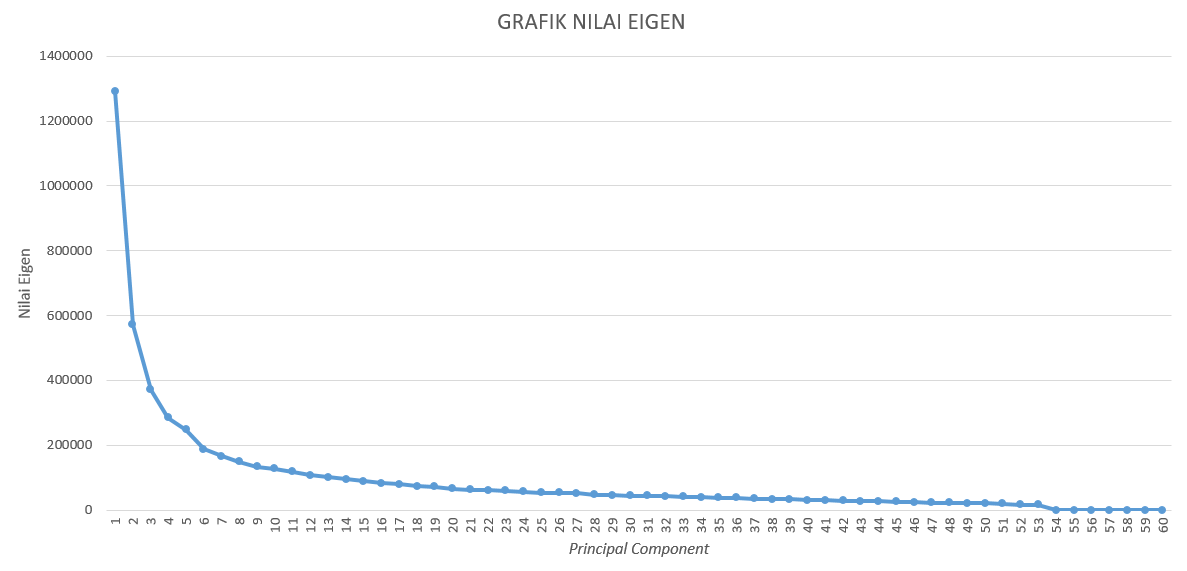
|  |  |  |  |
| --- | --- | --- | --- |
|  | Jumlah data 40-40 | Jumlah data 14-40 | Jumlah data 14-14 |
| F1 Score | 84.76% | 38.75% | 28.16% |
| Recall | 76.45% | 49.23% | 23.59% |
| Precision | 99.00% | 38.24% | 46.38% |

Tabel 4. 2 Recall, Precision dan F1 Score: Observasi Jumlah Distrubusi Data Setiap Kelas

Berdasarkan grafik pada Gambar 4.1. dan Tabel 4.2 dapat dilihat bahwa jumlah data tiap kelas mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, dan *f1 score* yang berbeda. Menggunakan metode *Random Oversampling* untuk memperbanyak data dan *K-Fold Cross Validation* dengan k=8 serta dilakukan percobaan sebanyak 50 kali, parameter uji dengan jumlah data kelas negative PCOS 40 dan jumlah data kelas positif PCOS 40 memberikan hasil terbaik dengan nilai *recall* sebesar 76.45%, *precision* 99.00% dan *f1 score* 84.76%. Jumlah distribusi data di setiap kelas mempengaruhi performansi karena metode klasifikasi Naïve Bayes adalah metode klasifikasi yang sensitif dengan jumlah data set tiap kelasnya yang harus seimbang, dengan jumlah data set yang seimbang di setiap kelasnya menjadikan nilai prior antar kelas akan sama besarnya. Selain itu dengan jumlah data training yang semakin banyak, maka sistem yang dibangun bisa melakukan proses *learning* terhadap data yang lebih beragam. Hal tersebut juga mempengaruhi nilai *mean* dan varian yang nantinya akan digunakan sebagai parameter *learning* pada tahap klasifikasi.

### **Pengujian Skenario 2**

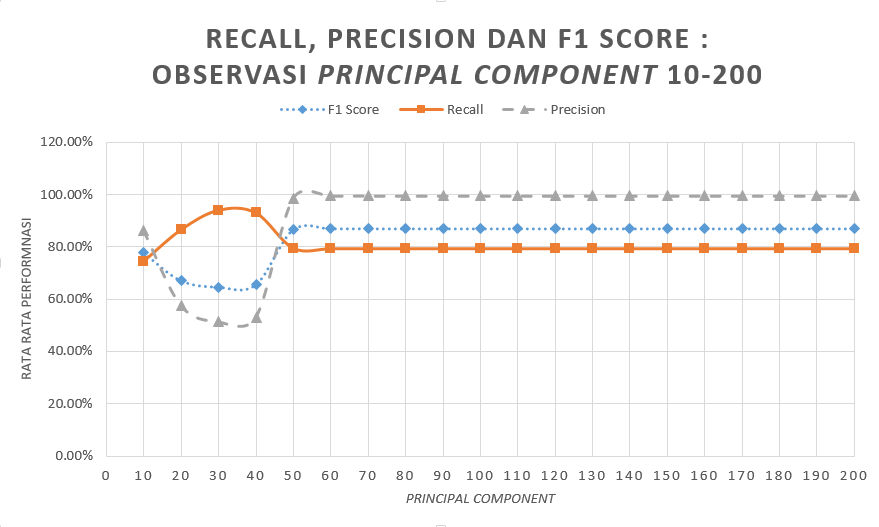
Pada skenario ini diukur pengaruh penggunaan PCA dan jumlah *principal components* yang digunakan terhadap performansi sistem dengan menghitung *recall*, *precision*, dan *f1 score* terhadap data set. Hasil pengujian pada skenario ini dapat dilihat di bawah ini.



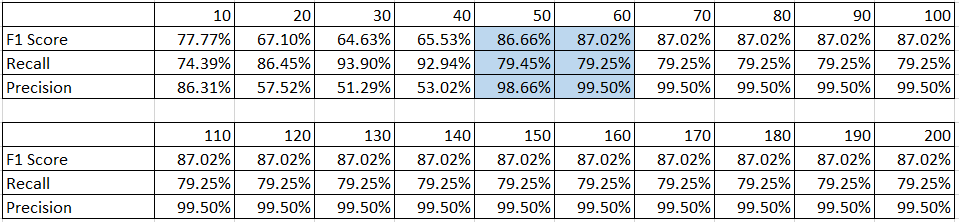
Gambar 4. 2. Grafik Nilai Eigen

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Principal Component | Nilai Eigen | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 1292486 | 571133.3 | 371341.1 | 285117.9 | 248015.5 | 188456.2 | 166803.8 | 148993.4 | 134404.7 | 127934.7 |
| 10 | 117583.6 | 108042.6 | 101069.4 | 94869.78 | 88263.39 | 82613.72 | 79780.61 | 73653.06 | 71713.38 | 65640.81 |
| 20 | 62294.51 | 60983.49 | 59898.92 | 56964.99 | 53569.14 | 52900.55 | 52075.3 | 47981.86 | 46551.1 | 45021.79 |
| 30 | 44047.68 | 43139.1 | 41391.7 | 39264.25 | 38659.77 | 37485.41 | 35557.12 | 34089.95 | 33440.56 | 30962.52 |
| 40 | 30558.55 | 28655.88 | 27216.87 | 26545.72 | 26020.15 | 24135.28 | 22944.26 | 22304.59 | 21086.09 | 20469.13 |
| 50 | 20083.73 | 17128.67 | 15749.55 | 0.190546 | 0.142884 | 0.01706 | 0.010676 | 0.008359 | 0.007935 | 0.005308 |

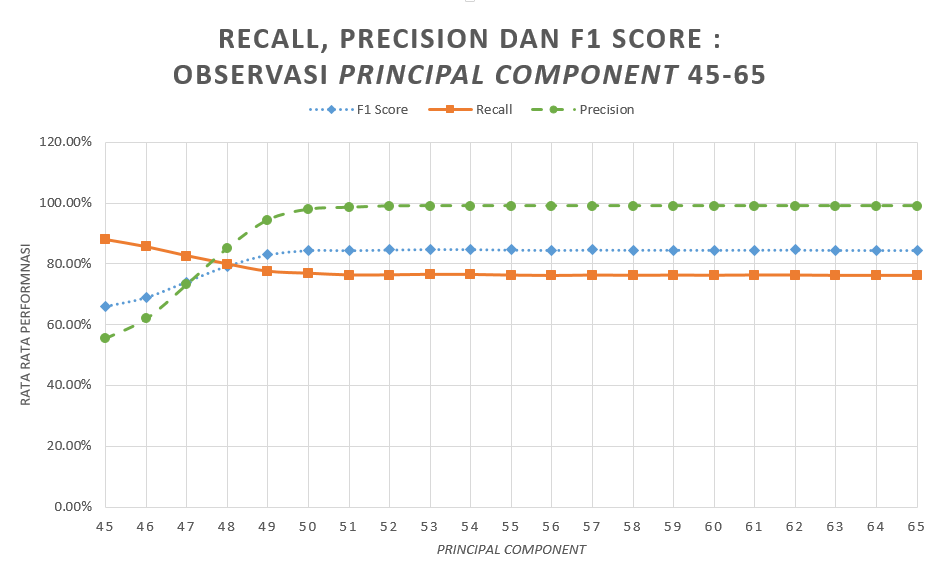
Tabel 4. 3 Nilai Eigen 1-60



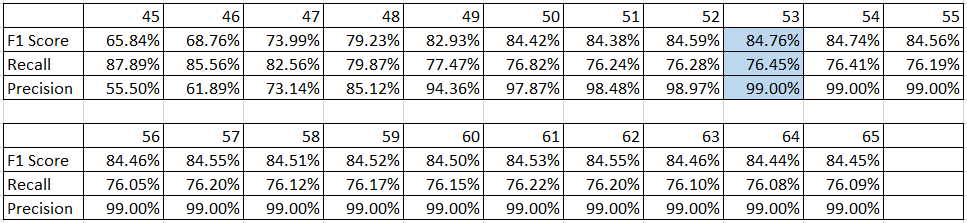
Gambar 4. 3. Grafik Recall, Precision dan F1 Score: Observasi Principal Component 10-200



Tabel 4. 4 Recall, Precision dan F1 Score:   
Observasi Principal Component 10-200



Gambar 4. 4. Grafik Recall, Precision dan F1 Score: Observasi Principal Component 45-65



Tabel 4. 5 Grafik Recall, Precision dan F1 Score: Observasi Principal Component 45-65

Didalam teori PCA dikatakan bahwa semakian besar nilai eigen suatu komponen maka semakin besar juga informasi yang terkandung didalamnya. Ada beberapa cara yang bisa dilakukan untuk mencari jumlah *principal components* paling optimal diantaranya yaitu: *Scree test* [12], dan *cross validation* [13]. Menurut Cattell [12] *scree test* adalah teknik grafis yang digunakan untuk menggambarkan nilai eigen dari komponen PCA sehingga dari grafis yang ditampilkan dapat diketahui jumlah *principal component* paling optimal. Seperti yang terlihat di Gambar 4.2 dan Tabel 4.3, dari observasi yang dilakukan dengan menggunakan *scree test* dapat dilihat bahwa jumlah *principal component* paling optimal adalah 53 komponen (nilai eigen komponen ke 54 sangat kecil = 0.190546). Setelah diketahui jumlah *principal component* paling optimal menggunakan *scree test,* langkah selanjutnya kita lakukan validasi dengan *cross validation*. Penjelasan proses yang dilakukan saat validasi menggunakan *cross validation* dijabarkan sebagai berikut:

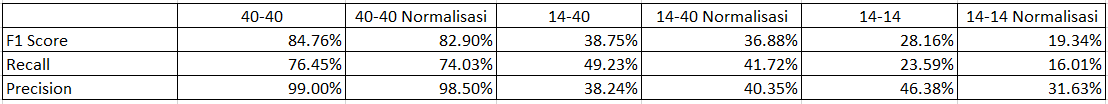
1. Dilakukan observasi dengan nilai *principal component* dari 10, 20, 30, .., 200 dengan percobaan sebanyak 50 kali. Pada Gambar 4.3 dan Tabel 4.4 dapat dilihat bahwa hasil *F1 Score* naik saat menggunakan *principal component* sebanyak 50, dan nilai *F1 Score* stabil disaat nilai *principal component* 60. Dari hasil percobaan di atas dapat ditarik kesimpulan bahwa penggunaan *principal component* paling optimal ada direntang 50-60 komponen.
2. Setelah dipercobaan sebelumnya diketahui bahwa *principal component* paling optimal ada di rentang 50-60 komponen, maka di percobaan kedua ini akan di lakukan observasi dengan nilai *principal component* dari 45,46,47, .., 65 dengan percobaan sebanyak 50 kali. Pada Gambar 4.4 dan Tabel 4.5 dapat dilihat bahwa hasil *F1 Score* terbaik muncul saat menggunakan *principal component* sebanyak 53.

Berdasarkan observasi, penggunaan PCA dan jumlah *principal components* yang digunakan terbukti mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, dan *f1 score* yang berbeda. Penggunaan PCA akan mereduksi dimensi data yang akan diproses dengan sesedikit mungkin menghilangkan informasi yang terkandung di dalamnya, sehingga sangatlah penting untuk mengetahui jumlah *principal componen*t paling optimal yang bisa digunakan. Parameter uji dengan jumlah *principal components* = 53 memberikan hasil terbaik dengan nilai *recall* sebesar 76.45%, *precision* 99.00% dan *f1 score* 84.76%.

### **Pengujian Skenario 3**

Pada skenario ini diukur performansi sistem dengan dengan menghitung *recall*, *precision*, dan *f1 score* terhadap data set. Hasil pengujian pada skenario ini dapat dilihat pada Gambar 4.5.

Gambar 4. 5. Recall, Precision dan F1 Score:   
Observasi Normalisasi



Tabel 4. 6 Recall, Precision dan F1 Score: Observasi Normalisasi

Berdasarkan Gambar 4.5. dapat dilihat bahwa proses normalisasi mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, dan *f1 score* yang berbeda. Parameter uji dengan menggunakan proses normalisasi memberikan hasil terbaik dengan nilai *recall* sebesar 80%, *precision* 82% dan *f1 score* 85%. Penggunaan proses normalisasi mempengaruhi performansi karena …..

# **BAB 5 KESIMPULAN DAN SARAN**

Pada bab ini akan di jelaskan kesimpulan dari hasil observasi dan analisis yang telah di lakukan berserta saran untuk penelitian ke depan.



## **Kesimpulan**

Berdasarkan observasi dan analisis dari percobaan yang telah di lakukan, maka dapat di tarik kesimpulan sebagai berikut :

1. Jika menggunakan metode klasifikasi Naïve Bayes, jumlah data yang seimbang di tiap kelasnya mempengaruhi sistem yang dibangun . Dari skenario yang sudah diuji, menggunakan metode SMOTE untuk memperbanyak data, parameter uji dengan jumlah data kelas normal 32 dan jumlah data kelas *Polycystic Ovary Syndrome* (PCOS) 32 memberikan hasil terbaik dengan nilai *recall* sebesar 80%, *precision* 82% dan *f1 score* 85%.
2. Penggunaan PCA terbukti mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, dan *f1 score* yang berbeda. Parameter uji dengan jumlah *principal components* = 100 memberikan hasil terbaik dengan nilai *recall* sebesar 80%, *precision* 82% dan *f1 score* 85%.
3. Proses normalisasi terbukti mempengaruhi performansi sistem yang dibuat, hal ini dapat di lihat dari nilai *recall*, *precision*, dan *f1 score* yang berbeda. Parameter uji dengan menggunakan proses normalisasi memberikan hasil terbaik dengan nilai *recall* sebesar 80%, *precision* 82% dan *f1 score* 85%.

## **Saran**

Adapun saran dari peneliti yang sekiranya dapat di lakukan pada penelitian selanjutnya, yaitu:

1. Data set yang digunakan harus lebih banyak dan variatif, selain itu juga jumlah data antar kelasnya harus seimbang.
2. Mencoba metode SMOTE (Synthetic Minority Over-sampling Technique) untuk *balancing* data.
3. Mencoba metode klasifikasi *Bayesian Network* atau *Neural Network* untuk data set *Polycystic Ovary Syndrome* (PCOS).

# **DAFTAR PUSTAKA**

[1] Purnama, B., Hasyim, A., Septiani, M. D., Wisesty, U. N., & Astuti, W. (2015). Follicle Detection on the USG Images to Support Determination of Polycystic Ovary Syndrome. In *Journal of Physics: Conference Series* (Vol. 622, No. 1, p. 012027). IOP Publishing.

[2] Guttmacher, A. E. (2012). Final Report The National Institutes of Health Polycystic Ovary Syndrome . NATIONAL INSTITUTES OF HEALTH Evidence-based Methodology Workshop on, 1-40.

[3] Setiawati, E., & Tjokorda, A. B. W. (2015, May). Particle Swarm Optimization on follicles segmentation to support PCOS detection. In Information and Communication Technology (ICoICT), 2015 3rd International Conference on (pp. 369-374). IEEE.

[4] Creighton University School of Medicine. (2005). Ultrasound of Uterus and Ovary,1-16. <http://www.toledoxray.com/sec/Guides/Ultrasound/Ultrasound%20of%20Uterus%20and%20Ovary.pdf>

[5] Järvelä, I. Y., Mason, H. D., Sladkevicius, P., Kelly, S., Ojha, K., Campbell, S., & Nargund, G. (2002). Characterization of normal and polycystic ovaries using three-dimensional power Doppler ultrasonography. Journal of assisted reproduction and genetics, 19(12), 582-590.

[6] Balen, A. H., Laven, J. S., Tan, S. L., & Dewailly, D. (2003). Ultrasound assessment of the polycystic ovary: international consensus definitions. Human reproduction update, 9(6), 505-514.

[7] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

[8] Junaedi, H., Budianto, H., Maryati, I., & Melani, Y. DATA TRANSFORMATION PADA DATA MINING.

[9] Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

[10] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), 1-6.

[11] Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In Encyclopedia of database systems (pp. 532-538). Springer US.

[12] Cattell, R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245-276.

[13] Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. Biology direct, 2(1), 2.

# **LAMPIRAN**

## **Contoh Dataset**

## **Tabel Hasil Pengujian**