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

**Proposal Tugas Akhir**

**Kelas TA I**

**Nanda Budi Prayuga**

**1103134395**



**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**2017**

# Lembar Persetujuan

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

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

**Nanda Budi Prayuga**

**NIM: 1103134395**

Proposal ini diajukan sebagai usulan pembuatan tugas akhir pada

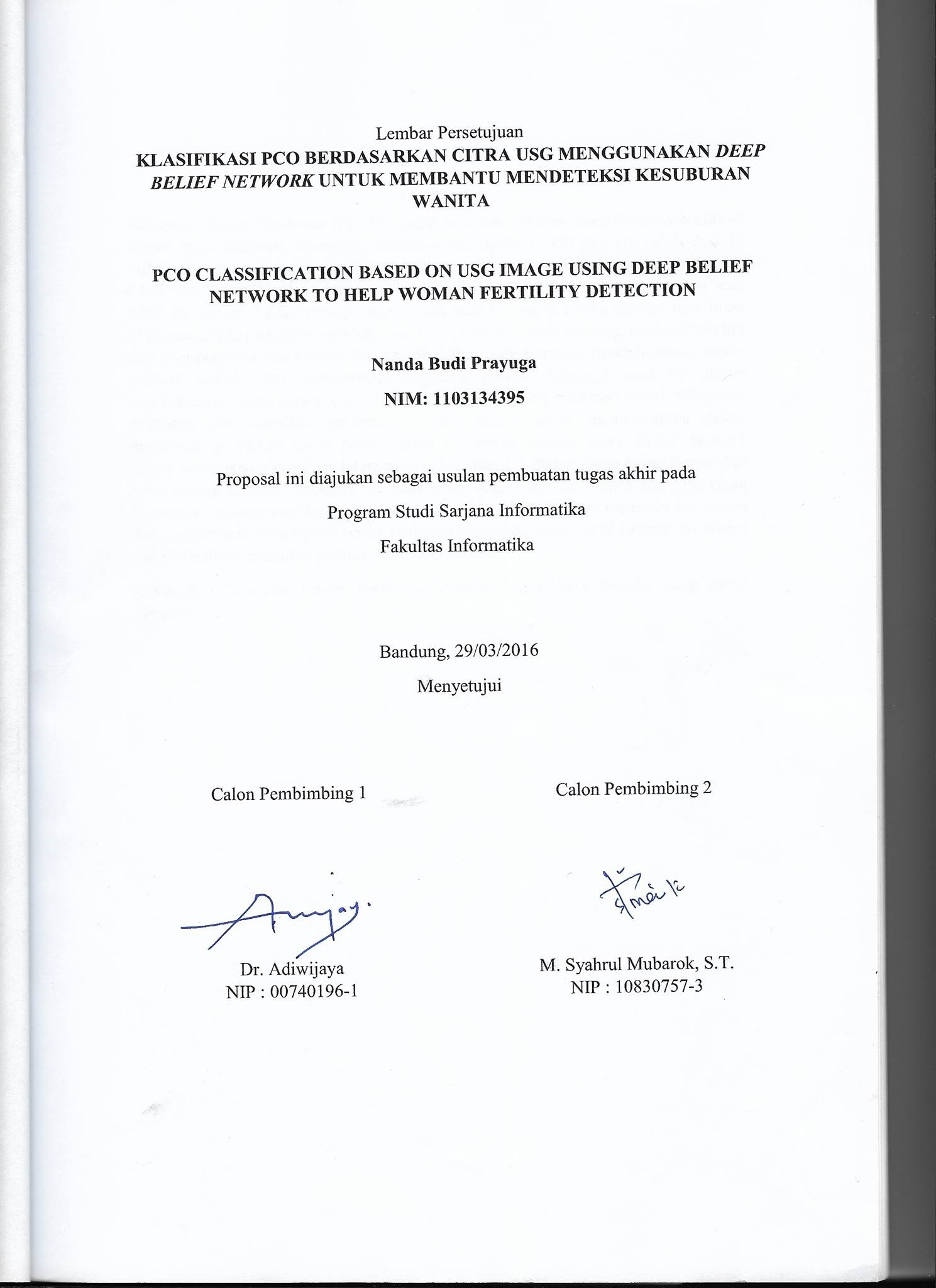
Program Studi Sarjana Informatika

Fakultas Informatika

Bandung, 10/08/2017

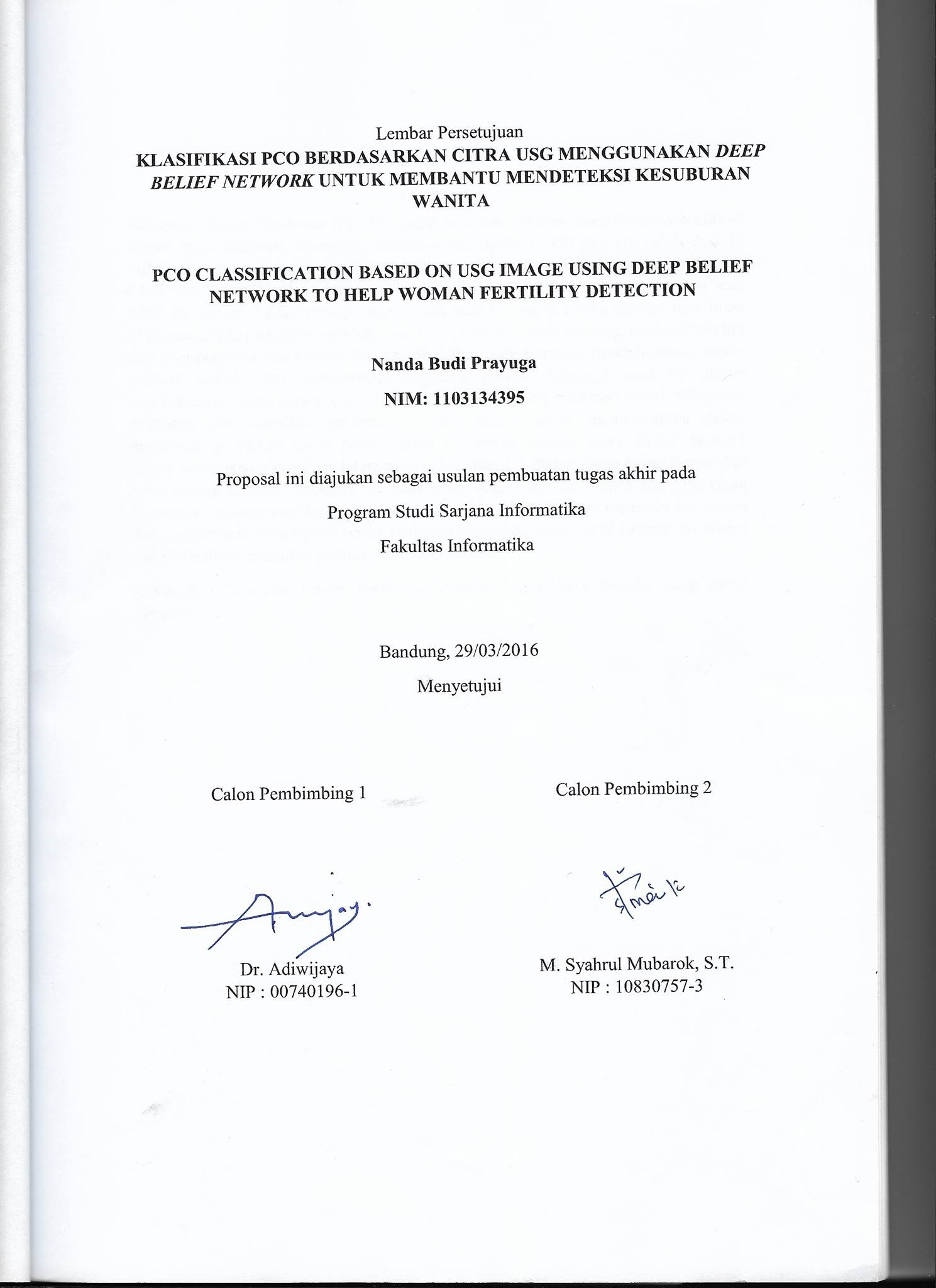
Menyetujui

Calon Pembimbing 1



Prof. Adiwijaya  
NIP : 00740196-1

Calon Pembimbing 2



M. Syahrul Mubarok, S.T.  
NIP : 10830757-3

# **Abstrak**

*Polycystic Ovary Syndrome (PCOS)* adalah kelainan sindrom yang di derita wanita di sistem reproduksinya, seseorang dikatakan menderita *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 *PCOS* dokter harus melakukan scan USG dan secara manual menghitung jumlah *follicle* yang di tandai dengan area hitam di gambar. 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. Sehingga saat ini dokter membutuhkan suatu sistem yang dapat mengekstrak setiap fitur dari citra USG secara otomatis, dan memiliki performansi yang bagus untuk membantunya dalam mendiagnosis PCOS untuk pendeteksian kesuburan wanita. *Gaussian Naïve Bayes* adalah salah satu turunan dari *Bayesian Network* yang sangat berpotensi digunakan dalam pemecahan masalah diatas. *Gaussian Naïve Bayes* di pilih karena keefektifannya dalam melakukan klasifikasi, walaupun dalam teorinya di asumsikan tiap fitur independen (naïve), serta diharapkan sistem yang dihasilkan memiliki performansi yang bagus.

Keyword : *Polycystic Ovary Syndrome*, ovarium, citra USG, *follicle, Bayesian Network , Gaussian Naïve Bayes.*

# **DAFTAR ISI**

[Lembar Persetujuan 2](#_Toc490145350)

[**Abstrak** 3](#_Toc490145351)

[**DAFTAR ISI** 4](#_Toc490145352)

[**DAFTAR GAMBAR** 7](#_Toc490145353)

[**DAFTAR ISTILAH** 8](#_Toc490145354)

[**DAFTAR TABEL** 9](#_Toc490145355)

[**BAB I PENDAHULUAN** 10](#_Toc490145356)

[**1.1** **Latar Belakang** 10](#_Toc490145357)

[**1.2** **Perumusan Masalah** 10](#_Toc490145358)

[**1.3** **Tujuan** 11](#_Toc490145359)

[**1.4** **Hipotesa** 11](#_Toc490145360)

[**1.5** **Rencana Kegiatan** 11](#_Toc490145361)

[**1.6** **Jadwal Kegiatan** 12](#_Toc490145362)

[**BAB II TINJAUAN PUSTAKA** 13](#_Toc490145384)

[**2.1** **Ovarium** 13](#_Toc490145385)

[**2.2** **Gaussian Naïve Bayes** 13](#_Toc490145386)

[**2.3** **Imbalance Data** 15](#_Toc490145387)

[**2.4** **Principal Component Analysis (PCA)** 15](#_Toc490145388)

[**2.5** **Normalization** 16](#_Toc490145389)

[**2.6** **Literatur Terkait** 17](#_Toc490145390)

[**BAB III PEMBAHASAN** 19](#_Toc490145391)

[**3.1** **Data Set** 19](#_Toc490145392)

[**3.2** **Gambaran Umum Sistem** 19](#_Toc490145393)

[**3.3** **Perancangan Sistem Klasifikasi** 22](#_Toc490145394)

[3.3.1 Perancangan *Balancing* Data 22](#_Toc490145395)

[3.3.2 Perancangan Normalisasi Data 23](#_Toc490145396)

[*3.3.3* Perancangan *Principal Component Analysis (PCA)* 23](#_Toc490145397)

[3.3.4 Perancangan *Gaussian Naïve Bayes* 24](#_Toc490145398)

[3.3.5 Proses Pengujian 25](#_Toc490145399)

[**3.4** **Spesifikasi Kebutuhan Sistem** 26](#_Toc490145400)

[**BAB 4 PENGUJIAN DAN ANALISIS** 27](#_Toc490145401)

[**4.1** **Tujuan Pengujian** 27](#_Toc490145406)

[**4.2** **Pengujian Sistem** 27](#_Toc490145407)

[4.2.1 Skenario Pengujian Sistem 27](#_Toc490145408)

[**4.3** **Analisis** 28](#_Toc490145409)

[4.3.1 Pengujian Skenario 1 28](#_Toc490145410)

[4.3.2 Pengujian Skenario 2 29](#_Toc490145411)

[4.3.3 Pengujian Skenario 3 31](#_Toc490145412)

[**BAB 5 KESIMPULAN DAN SARAN** 33](#_Toc490145413)

[**5.1** **Kesimpulan** 33](#_Toc490145419)

[**5.2** **Saran** 33](#_Toc490145420)

[**DAFTAR PUSTAKA** 34](#_Toc490145421)

# **DAFTAR GAMBAR**

[Gambar 2. 1. Ilustrasi model Gaussian 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 Gaussian 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 ISTILAH**

*Prior   
Likelihood   
Posterior*  
Fitur  
*Imbalanced*  
Dimensi  
*Naïve*  
Numpy   
Recall  
Precision  
F1 Score  
USG  
*Follicle*

# **DAFTAR TABEL**

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

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

# **BAB I PENDAHULUAN**

## **Latar Belakang**

*Polycystic Ovary Syndrome (PCOS)* adalah kelainan sindrom yang di derita wanita di sistem reproduksinya, seseorang dikatakan menderita *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 *PCOS* akan mengalami masalah pada kesuburannya serta mempunyai hormon androgen dan insulin yang tinggi. Efeknya orang tersebut kan 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 *PCOS* belum di tentukan, maka penderita *PCOS* ditangani berdasarkan gejala penyakitnya, bukan menangani penyebab *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 eksplisit, sehingga sistem tidak bisa secara mandiri menentukan deteksi penderita PCOS berdasarkan gambar USGnya.

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

## **Perumusan Masalah**

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

1. Bagaimana merancang sistem yang dapat mengekstrak setiap fitur dari citra USG secara otomatis?
2. Bagaimana mengimplementasikan sistem yang dirancang sehingga dapat mengklasifikasikan *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 mengekstrak setiap fitur dari citra USG secara otomatis.
2. Mengimplementasikan sistem yang dirancang sehingga dapat mengklasifikasikan *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 eksplisit, sehingga sistem tidak bisa secara mandiri menentukan deteksi penderita PCOS berdasarkan gambar USGnya.

*Gaussian Naïve Bayes* sangat berpotensi digunakan dalam pemecahan masalah deteksi *PCOS*. *Gaussian Naïve Bayes* di pilih karena keefektifannya dalam melakukan klasifikasi, walaupun dalam teorinya di asumsikan tiap fitur independen (naïve). Selain itu *Gaussian Naïve Bayes* mudah di optimasi dan digabungkan dengan metode-metode lainnya. Hasil akhirnya diharapkan dengan menggunakan metode *Gaussian Naïve Bayes* di dapatkan sistem yang powerful, efisien, dan memiliki tingkat akurasi yang tinggi untuk membantu dokter dalam mendiagnosis *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 PCOS. Data set yang digunakan berjumlah 100 buah, dimana nanti akan dibagi menjadi 3, yaitu : 50 buah untuk data training, 25 buah untuk data validasi, dan 25 buah untuk data testing.

* **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 di jelaskan gambaran tentang metode yang digunakan dalam menyelesaikan masalah yang di definiskan 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 tingkat keberhasilan dari model sistem, sehingga hasil akhirnya dapat di tarik kesimpulan.

* **Pembuatan Laporan Tugas Akhir**

Tahap terakhir yatiu pembuatan laporan tugas akhir, dimana hasil implementasi dan analisi di dokumentasikan 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 | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 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**

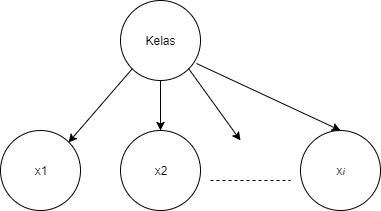
## **Ovarium**

Untuk mengetahui ukuran ovarium normal dapat di lihat saat masa menstruasi, melalui sebuah tes telah dilakukan di Creighton University School Of Medicine di dapat hasil dimana ciri-ciri ovarium normal yaitu : ukuran awal folikel 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 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 PCOS adalah volume dari ovarium. Ovarium normal memiliki volume kurang dari 10 cm3 (7.94 ± 2.34 cm3), sedangkan ovarium penderita PCOS memiliki volume melebihi 10 cm3 (14.04 ± 7.36 cm3)[8, 9].

## **Gaussian Naïve Bayes**

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



Gambar 2. 1. Ilustrasi model Gaussian Naive Bayes

Persamaan yang di gunakan pada model Gaussian Naïve Bayes bisa di lihat di bawah ini

(1)

atau jika di jabarkan 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 di lihat di persamaan (1), setiap data akan di cari posterior probabilitinya yang kemudian di tentukan 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 di kalikan 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 *Gaussian Naïve Bayes* sangat berpotensi digunakan dalam pemecahan masalah deteksi *PCOS*. *Gaussian Naïve Bayes* di pilih karena keefektifannya dalam melakukan klasifikasi, walaupun dalam teorinya di asumsikan tiap fitur independen (naïve). Selain itu *Gaussian Naïve Bayes* mudah di optimasi dan digabungkan dengan metode-metode lainnya. Hasil akhirnya diharapkan dengan menggunakan metode *Gaussian Naïve Bayes* di dapatkan sistem yang powerful, efisien, dan memiliki tingkat akurasi yang tinggi untuk membantu dokter dalam mendiagnosis *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 di sebut *major class,* sedangkan data set kelas yang jumlahnya paling sedikit di *sebut minor class*[10]. Permasalahan yang biasanya di hadapi 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. 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.*

## **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 :

value baru dari *x*

minimum value dari *x*

maximum value dari *x*

maximum value baru dari *x*

minimum 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 di lihat di bawah ini.

Penjelasan dari rumus di atas :

value baru dari *x*

rata-rata value dari *x*

standar devisiasi dari *x*

## **Literatur Terkait**

Berdasarkan penelitian yang di publikasikan National Institutes of Health (NIH) seseorang dikatakan menderita *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 *PCOS* dokter harus melakukan scan USG dan secara manual menghitung jumlah *follicle* yang di tandai dengan area hitam di gambar. Metode yang digunakan untuk mendeteksi *PCOS* bisa menggunakan NIH 1990, Rotterdam 2003, atau AE-PCOS Society 2006. Sedangkan untuk menangani permasalahan *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 *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 *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 di improvisasi 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 *PSO image clustering* dengan *contrast enhancement preprocessing* akan menghasilkan jarak intracluster dan *quantization error* yang lebih besar daripada metode *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**

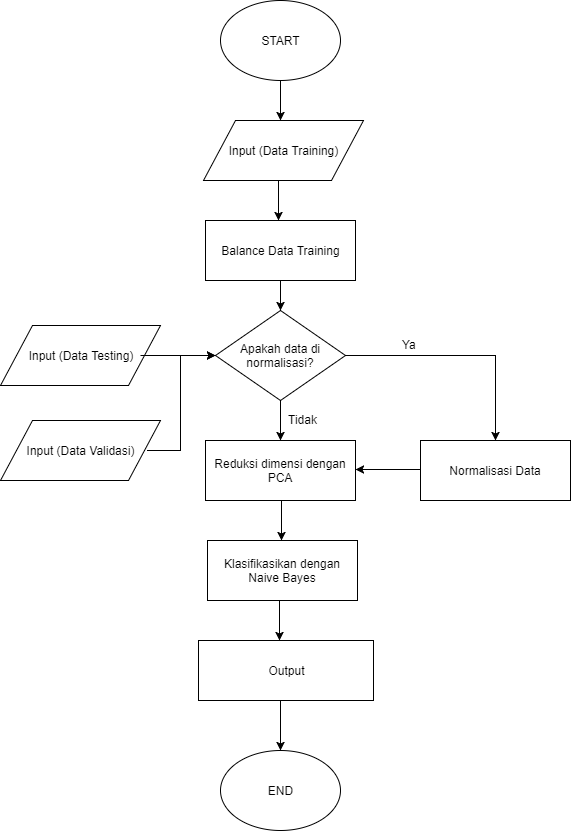
**Macintosh HD:Users:hendro:Downloads:Data Set.png**

Gambar 3. 1. Data Set

Data set adalah data yang akan di gunakan untuk proses pembuatan model dari sistem. Data set berupa ada gambar USG ovarium yang di bagi menjadi 2 kelas label. Gambar ovarium positif terkena PCOS dan negative PCOS. Setelah itu data set di pecah menjadi 3 bagian, yaitu : 25% data testing, 50% data training, dan 25% data validasi, lalu dari data tersebut akan digunakan untuk membangun model sistem.

## **Gambaran Umum Sistem**

Gambar 3.2 menggambarkan proses secara umum sistem yang akan dibuat

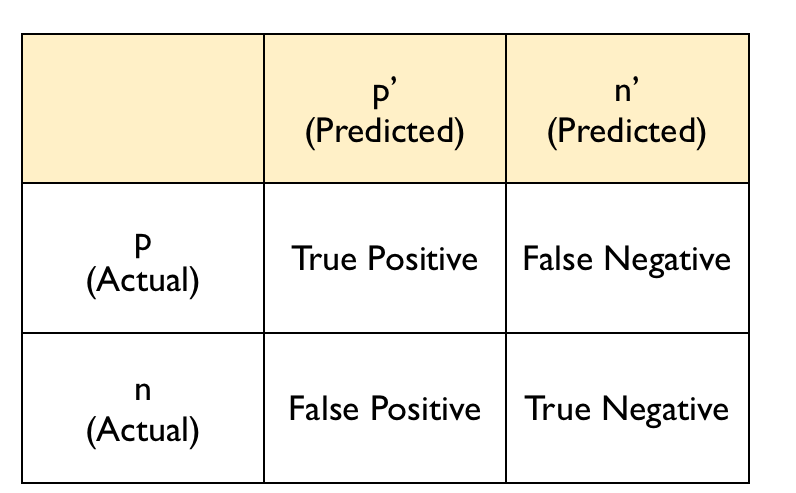
****

Gambar 3. 2 Gambaran Umum Sistem

Pada dasarnya sistem yang akan dibangun di bagi menjadi empat bagian utama, yaitu : *balancing* data, normalisasi data, reduksi dimensi, dan klasifikasi. Sebelum melalui ketiga proses tersebut data set di bagi menjadi tiga bagian yaitu data training, data testing dan data validasi seperti di jelaskan di bab 3.1. Gaussian Naïve Bayes adalah metode klasifikasi deterministik yang sensitif dengan jumlah data set tiap kelasnya yang harus seimbang. Salah satu masalah yang di hadapi dalam pembuatan sistem ini adalah jumlah data set yang sedikit (80 gambar) dan jumlah gambar tiap kelas yang tidak seimbang. Oleh karena itu sebelum di lakukan proses PCA dan klasifikasi dengan Gaussian Naïve Bayes harus dilakukan *balancing* data (jumlah data training dalam setiap kelas di seimbangkan).

Tahap selanjutnya setelah di lakukan *balancing* data adalah tahap reduksi dimensi menggunakan PCA. Reduksi dimensi menggunakan PCA di lakukan untuk mendapatkan fitur-fitur penting dari gambar dan juga mempercepat proses komputasi. Data yang sudah di reduksi dimensinya kemudian akan masuk ke tahap klasifikasi menggunakan Gaussian Naïve Bayes untuk menentukan label dari setiap gambar yang di inputkan.

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



Tabel 3. 1. Confusion Matrix

Sedangkan rumus perhitungannya sendiri yaitu :

Pada tugas akhir ini akan di observasi pengaruh penggunaan metode balancing data dan PCA terhadap sistem yang dibangun. Akan ada empat scenario yang kan di uji coba pada sistem ini, yaitu :

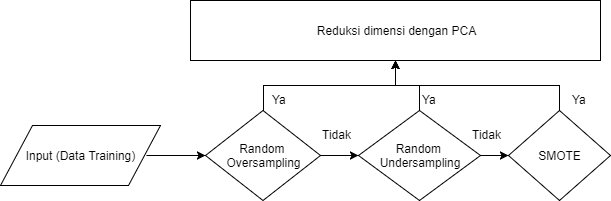
* 1. Skenario pertama data training akan melalui proses balancing data dan PCA.
  2. Skenario kedua data training hanya akan melalui proses PCA.
  3. Skenario ketiga data training akan melalui proses balancing data tetapi tidak melalui proses PCA
  4. Skenario keempat data training tidak akan melalui proses balancing data dan juga PCA.

## **Perancangan Sistem Klasifikasi**

Perancangan sistem klasifikasi yang akan dibangun di bagi menjadi empat bagian utama, yaitu : *balancing* data, normalisasi data, reduksi dimensi, dan klasifikasi.

### Perancangan *Balancing* Data

Perancangan sistem *balancing* datadalam sistem yang akan dibangun dapat di lihat di alur diagram pada Gambar 3.3 di bawah ini*.*



Gambar 3. 3. Alur Diagram Balancing Data

Sebelum dilakukan proses reduksi dimensi dengan PCA, input data training akan di *balancing* dengan salah satu dari tiga 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

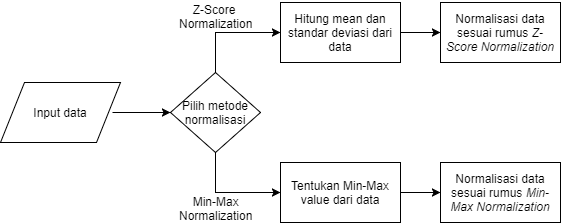
Kebalikan dari random oversampling, random undersampling adalah suatu mekanisme untuk mengurangi data set *mayor class* dengan cara memilih data yang tersedia dari *mayor class* secara acak, lalu hilangkan data yang sudah di pilih dari data set *mayor class.*

1. Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) adalah suatu mekanisme untuk menambah data set *minor class* dengan cara memilih data yang tersedia dari *minor class* berdasarkan similaritas antara data set di *minor class* menggunakan metode KNN. Data yang sudah di pilih lalu salin dan tambahkan ke data set *minor class.*

### Perancangan Normalisasi Data

Perancangan sistem normalisasidatadalam sistem yang akan dibangun dapat di lihat di alur diagram pada Gambar 3.4 di bawah ini*.*



Gambar 3. 4. Alur Diagram Normalisasi Data

Dilihat dari alur diagram normalisasi di atas (Gambar 3.4) ada dua metode yang nantinya akan di ujicoba 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 di lihat di alur diagram pada Gambar 3.5 di bawah ini.

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 di lakukan 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 di lakukan dengan menggunakan rumus yang sudah di jelaskan di bab 2

1. Hitung Nilai Eigen dan Vektoe Eigen

Hasil dari perhitungan matriks covariance di langkah sebelumnya kemudian di olah lagi untuk mendapatkan nilai eigen dan vektor eigen nya.

1. Tranformasikan ke Domain Eigen

Setelah nilai eigen dan vektor eigen di dapat, 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 di dapat yaitu dimensi dari data set yang sudah tereduksi dan *feature vector.*

### Perancangan *Gaussian Naïve Bayes*

Perancangan sistem *Gaussian Naïve Bayes* dalam sistem yang akan dibangun dapat di lihat di alur diagram pada Gambar 3.6 di bawah ini.

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

Gambar 3. 6. Alur Diagram Gaussian 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 di lakukan dalam adalah menghitung prior dari tiap kelas dengan cara jumlah data di tiap kelas dibagi denganjumlah data.

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

Dalam proses ini akan di cari *mean* dan *variance* tiap fitur data training yang kemudian akan di pakai di proses berikutnya

1. Hitung *Posterior Probability* Data terhadap Kelas

Proses perhitungan *posterior probability* data terhadap kelas yang dilakukan sudah di jabarkan 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 di dapat, 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 di lihat di bawah ini.

1. Skenario I – Pengujian berdasarkan jumlah data training tiap kelasnya terhadap sistem yang dibangun
   1. Tujuan : Mendapatkan jumlah data training paling optimal untuk tiap kelasnya
   2. Pengukuran : F1 Score Measure
   3. Variabel Peubah : Jumlah data tiap kelas
   4. Variabel Tetap : Nilai PCA
   5. Cara Pengukuran : Menghitung F1 Score berdasarkan jumlah data tiap kelas yang telah di tentukan 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 di gunakan
6. Variabel Tetap : Jumlah data tiap kelas
7. Cara Pengukuran : Menghitung F1 Score berdasarkan jumlah *principal component* yang telah di tentukan 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 data tiap kelas, Nilai PCA
13. Cara Pengukuran : Menghitung F1 Score berdasarkan penggunaan normalisasi yang telah di tentukan 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 : Macbook
  + Model : Macbook Air 11 inch 2013
  + Prosesor : 1.3 GHz Intel Core i5
  + RAM : 4 GB 1600 MHz DDR3
  + Storage : SSD 128 GB
  + GPU : Intel HD Graphics 5000 1536 MB

**Spesifikasi Perangkat Lunak**

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

* + OS : OS X Yosemite 10.10
  + Aplikasi : Microsoft Word 2013, Python 3, Numpy.

# **BAB 4 PENGUJIAN DAN ANALISIS**

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



## **Tujuan Pengujian**

Tujuan pengujian pada sistem yang akan 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 akurasi yang tinggi, sehingga membantu dokter dalam mendiagnosis PCOS untuk pendeteksian kesuburan wanita.

## **Pengujian Sistem**

Pengujian terhadap sistem yang dibangun menggunakan skenario-skenario yang sudah di jelaskan di bab 3. Performansi sistem dapat diukur dengan dengan menghitung recall, precision, akurasi dan f1 measure untuk setiap skenario terhadap datasets yang ada. Hasil performansi dari setiap skenario yang diukur akan dijelaskan di bab ini. Parameter-parameter yang akan di observasi di sistem yang dibangun yaitu :

* Jumlah data training tiap kelasnya pada tahap *balancing data*
* Penggunaan PCA dan jumlah *principal components* yang di gunakan pada tahap dimensi reduksi
* Penggunaan normalisasi terhadap data set pada tahap *normalization*

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

* Mengetahui performansi sistem dan menganalisis pengaruh jumlah data training tiap kelasnya terhadap sistem yang dibangun .
* Mengetahui performansi sistem dan menganalisis pengaruh penggunaan PCA dan jumlah *principal components* yang di gunakan terhadap sistem yang dibangun .
* Mengetahui performansi sistem dan menganalisis pengaruh penggunaan normalisasi pada data set terhadap sistem yang dibangun .

### Skenario Pengujian Sistem

Skenario pengujian sistem yang akan di lakukan pada tugas akhir ini dapat di lihat pada table 4.1 di bawah ini.

|  |  |  |  |
| --- | --- | --- | --- |
| Skenario ke- | Tujuan | Parameter Uji | Parameter Konstan |
| 1 | Mengetahui performansi sistem dan menganalisis pengaruh jumlah data training tiap kelasnya terhadap sistem yang dibangun | Jumlah data tiap kelas : 16-8, 8-8, 16-16 (SMOT), 32-32 (SMOT) | Nilai PCA : 100 |
| 2 | Mengetahui performansi sistem dan menganalisis pengaruh penggunaan PCA dan jumlah *principal components* yang di gunakan di tahap dimensi reduksi terhadap sistem yang dibangun . | Jumlah *principal components* yang di gunakan:  Tanpa PCA, 50, 60, 70, .., 200 | Hasil skenario 1 |
| 3 | Mengetahui performansi sistem dan menganalisis pengaruh penggunaan normalisasi pada data set yang di lakukan di tahap *normalization* terhadap sistem yang dibangun . | Menggunakan normalisasi atau tidak | Hasil skenario 2 |

Tabel 4. 1. Skenario Pengujian

## **Analisis**

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

### Pengujian Skenario 1

Pada skenario ini akan diukur performansi sistem dengan dengan menghitung *recall*, *precision*, akurasi dan *f1 score* terhadap data set. Hasil pengujian pada skenario ini dapat di lihat pada gambar 4.1 di bawah ini.

Gambar 4. 1. Recall, Precision dan F1 Score

Berdasarkan Gambar 4.1. dapat dilihat bahwa jumlah data tiap kelas mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, akurasi dan *f1 score* yang berbeda. Menggunakan metode SMOTE untuk memperbanyak data, parameter uji dengan jumlah data kelas normal 32 dan jumlah data kelas PCOS 32 memberikan hasil terbaik dengan nilai *recall* sebesar 80%, *precision* 82% dan *f1 score* 85%. Jumlah data tiap kelas mempengaruhi performansi karena metode klasifikasi Gaussian Naïve Bayes adalah metode klasifikasi deterministik yang sensitif dengan jumlah data set tiap kelasnya yang harus seimbang. Selain itu dengan jumlah data training yang semakin banyak maka sistem yang dibangun bisa melakukan proses *learning* terhadap data yang lebih beragam.

### Pengujian Skenario 2

Pada skenario ini akan diukur performansi sistem dengan dengan menghitung *recall*, *precision*, akurasi dan *f1 score* terhadap data set. Hasil pengujian pada skenario ini dapat di lihat pada gambar 4.2 di bawah ini.

Gambar 4. 2. Recall

Gambar 4. 3. Precision

Gambar 4. 4. F1 Score

Berdasarkan Gambar 4.2, 4.3 dan 4.4 dapat di lihat penggunaan PCA dan jumlah *principal components* yang digunakan terbukti mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, akurasi 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%. Hal ini dapat terjadi di karenakan penggunaan PCA akan mengurangi dimensi data yang akan di proses dan menghilangkan fitur-fitur yang tidak berpengaruh terhadap karakteristik data.

### Pengujian Skenario 3

Pada skenario ini akan di ukur performansi sistem dengan dengan menghitung *recall*, *precision*, akurasi dan *f1 score* terhadap data set. Hasil pengujian pada skenario ini dapat dilihat pada gambar 4.5 di bawah ini.

Gambar 4. 5. Recall, Precision dan F1 Score

Berdasarkan Gambar 4.5. dapat dilihat bahwa proses normalisasi mempengaruhi performansi sistem yang dibuat, hal ini dapat dilihat dari nilai *recall*, *precision*, akurasi 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 Gaussian 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 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*, akurasi 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*, akurasi 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. Jika jumlah data set sedikit atau tidak seimbang metode SMOTE (Synthetic Minority Over-sampling Technique) bisa di coba untuk meningkatkan performansi dari sistem yang dibuat.
3. Mencoba metode klasifikasi *Bayesian Network* atau *Neural Network* untuk data set 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.