Early-Stage Diabetes Risk Prediction

1. Latar Belakang

Dilansir dari Kompas.com, berdasarkan data International Diabetes Federation (IDF), Indonesia berstatus waspada diabetes karena menempati urutan ke-7 dari 10 negara dengan jumlah pasien diabetes tertinggi. Pravalensi pasien pengidap diabetes di Indonesia mencapai 6.2%, yang artinya ada lebih dari 10,8 juta orang menderita diabetes per tahun 2020.

Ketua Umum Perkumpulan Endokrinologi Indonesia (Perkeni), Prof. Dr. dr. Ketut Suastika, SpPD-KEMD mengatakan bahwa angka ini diperkirakan meningkat menjadi 16,7 juta pasien per tahun 2045. Dengan data tahun ini, 1 dari 25 penduduk Indonesia atau 10% dari penduduk Indonesia mengalami diabetes. Tingkat kesadaran masyarakat Indonesia yang rendah juga menjadi salah satu unsur penyebab diabetes terus menerus merenggut kehidupan masyarakat luas ini tanpa disadari.

Fakta seputar penyakit diabetes yaitu erdapat 425 juta pasien diabetes per tahun 2017 di dunia. Angka ini diperkirakan akan meningkat sebesar 45% atau setara dengan 629 juta pasien per tahun 2045. Komplikasi pada jantung dan ginjal menjadi penyebab utama kematian pasien diabetes di dunia. 75% pasien diabetes pada tahun 2017 berusia 20-64 tahun.

2. Identifikasi Masalah

Berdasarkan latar belakang yang telah dijelaskan, permasalahan-permasalahan yang ingin diketahui yaitu sebagai berikut:

1. Metode manakah yang mampu melakukan klasifikasi dari diabetes yang diderita pasien secara tepat dan berapa besar tingkat akurasi klasifikasi model tersebut?

3. Bahan dan Metode

Data yang akan digunakan yaitu sebanyak 520 responden dan 17 variabel. Sumber data diperoleh dari University of California Irvine’s (UCI) Machine Learning Repository. Umur responden yang diperoleh dari 16 tahun – 90 tahun. Menggunakan model klasifikasi sebagai berikut:

1. Logistic Regression
2. Support Vector Machines Linear
3. SVM RBF
4. KNN
5. Naive Bayes Gaussian
6. Decision Trees
7. Random Forest

Dibawah ini adalah atribut yang digunakan yaitu sebagai berikut:

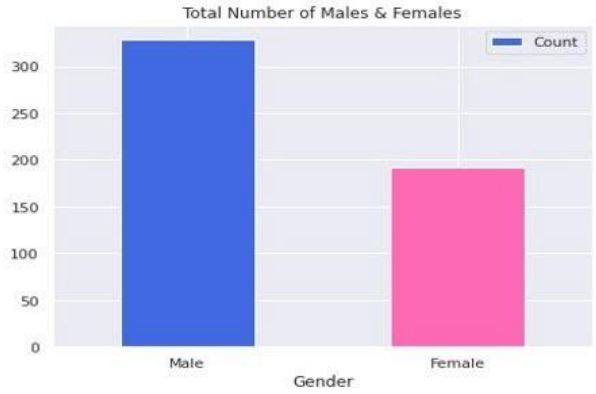
Tabel 3.1 Deskripsi Variabel

Name of Variable	Description of variable	Type of Variable
Age	The age of the person	Numeric
Gender	The gender of the person	Male/Female
Polyuria	The occurrence of polyuria in a person.	Yes/No
Polydipsia	The occurrence of polydipsia in a person.	Yes/No
sudden weight loss	If the person experiences sudden weight loss	Yes/No
weakness	If the person experiences genital thrush	Yes/No
Polyphagia	The occurrence of polyphagia in a person.	Yes/No
Genital thrush	If the person experiences genital thrush	Yes/No
visual blurring	If the person experiences visual blurring	Yes/No
Itching	If the person experiences itching	Yes/No
Irritability	If the person experiences irritability	Yes/No
delayed healing	If the person experiences delayed healing	Yes/No
partial paresis	If the person experiences partial paresis	Yes/No
muscle stiffness	If the person experiences muscle stiffness	Yes/No
Alopecia	The occurrence of alopecia in a person.	Yes/No
Obesity	If the person is obese	Yes/No
class	Whether the person is diabetic or not	Positive/Negative

Sumber: University of California Irvine's (UCI) Machine Learning Repository

4. Hasil dan Pembahasan

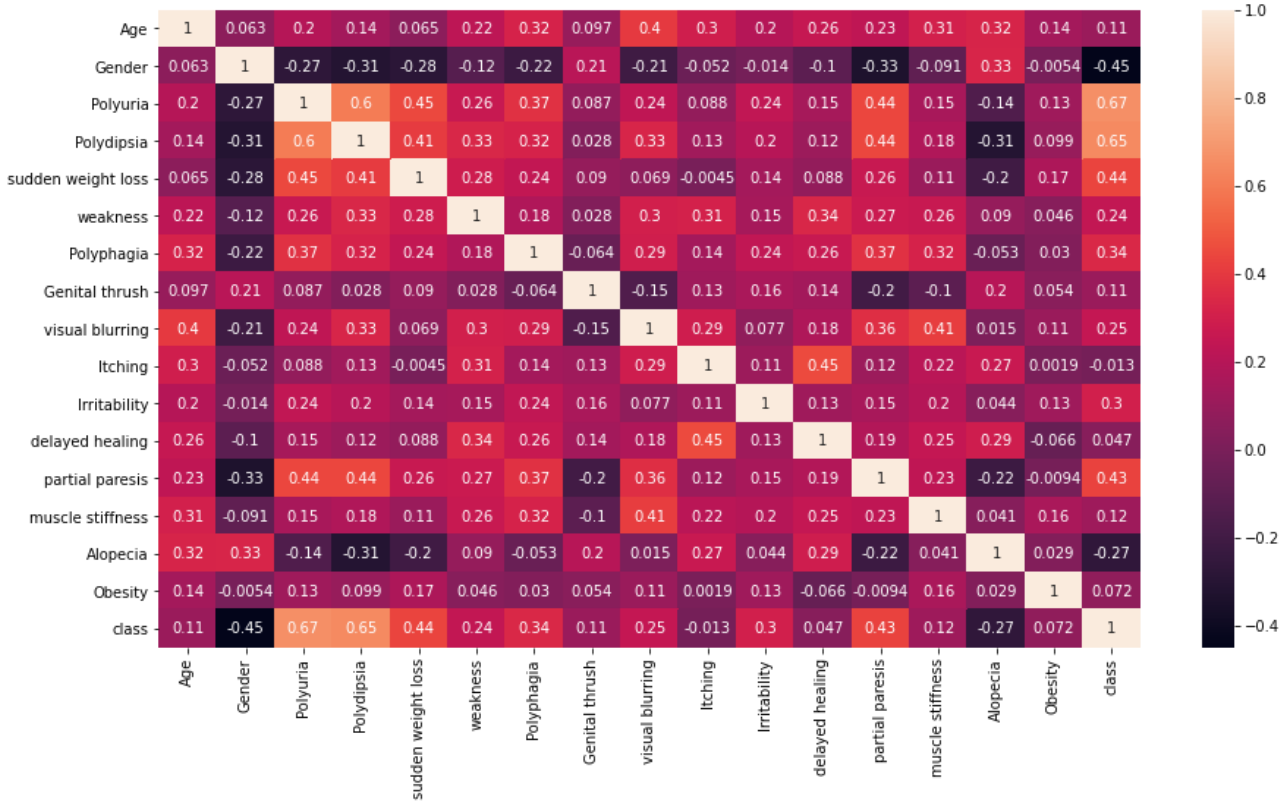
Tabel 4.1 Total Responden Berdasarkan Jenis Kelamin



Sumber: data diolah peneliti

Berdasarkan Tabel 4.1 dapat diketahui total responden berdasarkan jenis kelamin yaitu pada laki-laki lebih banyak daripada perempuan yaitu sebanyak 328 atau sebesar 63,08% pada responden laki-laki dan 192 atau sebesar 36,92% pada responden perempuan.

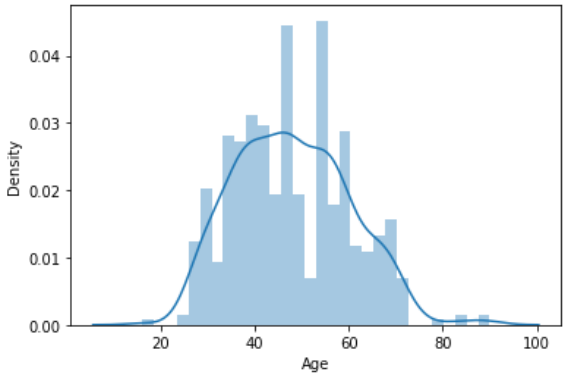
Tabel 4.2 Deskripsi Data Korelasi Responden Diabetes



Sumber: data diolah penulis

Dari heatmap 4.2 di atas, dapat diketahui ternyata korelasi yang paling tinggi adalah antara class dengan polyuria. Sementara itu, beberapa atribut lain yang juga berkorelasi dengan cukup kuat adalah polydipsia dengan class, polyuria dengan polydipsia, sudden weight loss dengan polyuria, sudden weight loss dengan class, delayed healing dengan itching, partial paresis dengan polyuria dan partial paresis dengan polydipsia. Di sisi lain, obesity tidak terlihat berkorelasi kuat dengan class.

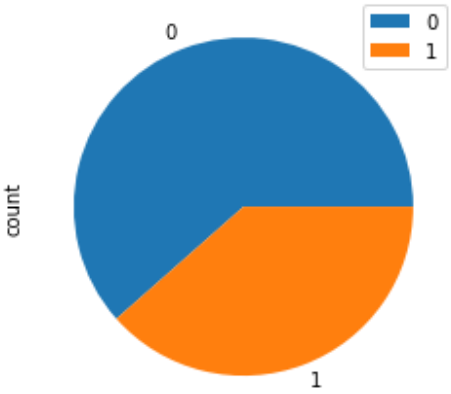
Tabel 4.3 Data Responden Berdasarkan Umur



Sumber: data diolah peneliti

Berdasarkan Tabel 4.3 dapat diketahui data responden berdasarkan umur dimulai dari 16 tahun – 90 tahun. Apabila dilihat dari grafik terjadi lonajakan dimana rata-rata responden tersebut berada di umur 39 tahun – 57 tahun dan paling banyak berada di umur 48 tahun.

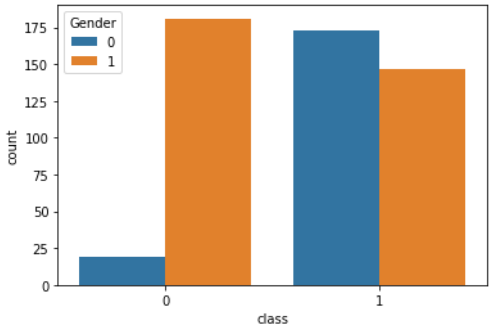
Tabel 4.4 Data Responden Berdasarkan Class



Sumber: data diolah penulis

Berdasarkan Tabel 4.4 dapat diketahui data responden berdasarkan class yaitu pada responden positif diabetes lebih banyak daripada responden negative diabetes yaitu sebanyak 320 atau sebesar 61,54% untuk responden positif diabetes dan 200 atau sebesar 38,46% untuk responden negative diabetes.

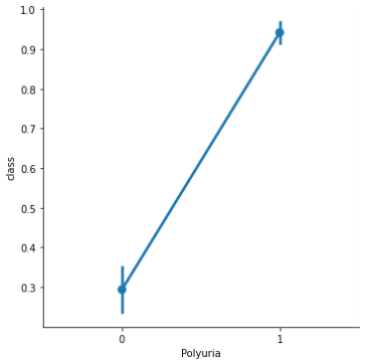
Tabel 4.5 Data Responden Berdasarkan Jenis Kelamin



Sumber: data diolah penulis

Berdasarkan Tabel 4.5 dapat diketahui data responden berdasarkan jenis kelamin yaitu pada responden negative diabetes pada laki-laki lebih banyak daripada perempuan sedangkan responden positif diabetes pada perempuan lebih banyak daripada laki-laki.

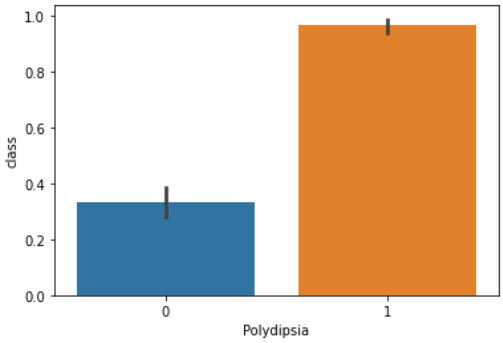
Tabel 4.6 Data Responden Berdasarkan Gejala Polyuria



Sumber: data diolah peneliti

Berdasarkan Tabel 4.6 dapat diketahui data responden berdasarkan gejala polyuria yaitu ketika seseorang dinyatakan positif diabetes maka muncul gejala polyuria atau kondisi ketika tubuh menghasilkan urin secara berlebihan.

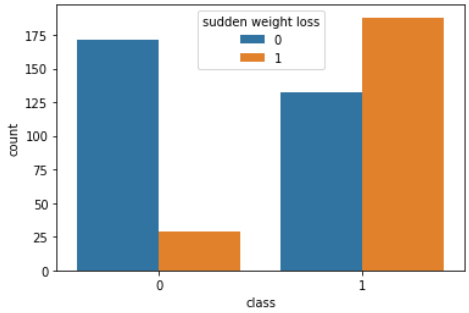
Tabel 4.7 Data Responden Berdasarkan Gejala Polydipsia



Sumber: data diolah peneliti

Berdasarkan Tabel 4.7 dapat diketahui data responden berdasarkan gejala polydipsia yaitu ketika seseorang dinyatakan positif diabetes maka muncul gejala polydipsia atau rasa haus yang berlebihan.

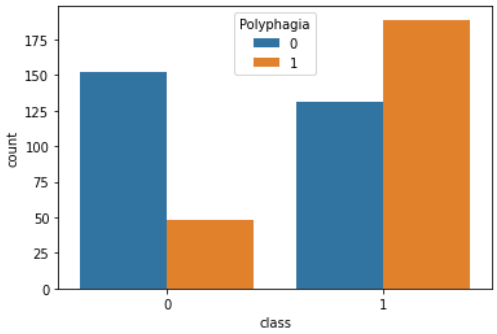
Tabel 4.8 Data Responden Class Berdasarkan Penurunan Berat Badan



Sumber: data diolah penulis

Berdasarkan Tabel 4.8 dapat diketahui data responden berdasarkan penurunan berat badan secara drastis yaitu pada responden negatif diabetes tidak terjadi penurunan berat badan secara drastis sedangkan pada responden positif diabetes terjadi penurunan berat badan secara drastis.

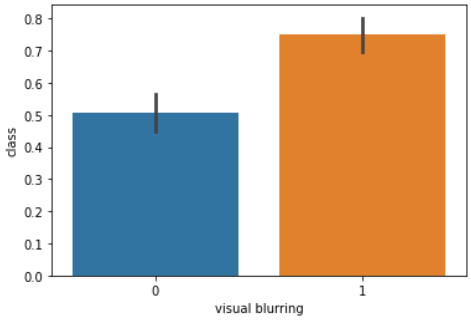
Tabel 4.9 Data Responden Class Berdasarkan Gejala Polyphagia (Rasa Lapar Berlebihan)



Sumber: data diolah penulis

Berdasarkan Tabel 4.9 dapat diketahui data responden berdasarkan polyphagia atau rasa lapar berlebihan yaitu pada responden negatif diabetes tidak terjadi rasa lapar yang berlebihan sedangkan pada responden positif diabetes terjadi polyphagia atau bisa disebut rasa lapar yang berlebihan.

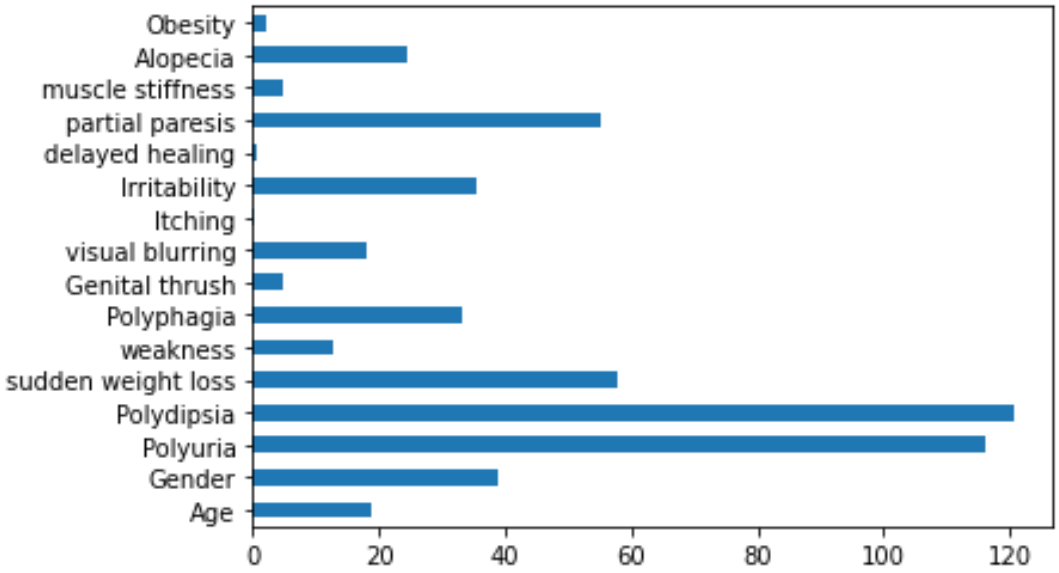
Tabel 4.10 Data Responden Berdasarkan Visual Blurring



Sumber: data diolah peneliti

Berdasarkan Tabel 4.10 dapat diketahui apabila penderita dinyatakan positif diabetes mengalami penglihatan yang tidak jelas.

Tabel 4.11 Data Responden Berdasarkan Penderita Diabetes



Sumber: data diolah penulis

Berdasarkan Tabel 4.11 dapat diketahui apabila penderita diabetes mengalami gejala awal yaitu diantaranya polydipsia atau rasa haus yang berlebihan, polyuria atau kondisi ketika tubuh menghasilkan urin secara berlebihan, dan kehilangan berat badan secara drastis

Metode Model Klasifikasi

1. Logistic Regression

Dalam kasus ini, didapat:

Accuracy : `0.8942307692307693`

```
[[34  5]
 [ 6 59]]
```

Prediksi positif benar sebanyak 34, positif salah sebanyak 5, negatif salah sebanyak 6 dan negatif benar sebanyak 59. Dari angka tersebut kita juga dapat menghitung presisi, akurasi dan penarikan kesimpulan. Pada metode logistic regression didapat akurasi sebesar 0.8942307692307693.

2. Support Vector Machines Linear

Accuracy: `0.9038461538461539`

```
[[34  4]
 [ 6 60]]
```

Prediksi positif benar sebanyak 34, positif salah sebanyak 4, negatif salah sebanyak 6 dan negatif benar sebanyak 60. Dari angka tersebut kita juga dapat menghitung presisi, akurasi dan penarikan kesimpulan. Pada metode SVM Linear didapat akurasi sebesar 0.9038461538461539.

3. SVM RBF

Accuracy: 0.9807692307692307
```
[[39  1]
 [ 1 63]]
```

Prediksi positif benar sebanyak 39, positif salah sebanyak 1, negatif salah sebanyak 1 dan negatif benar sebanyak 63. Dari angka tersebut kita juga dapat menghitung presisi, akurasi dan penarikan kesimpulan. Pada metode SVM RBF didapat akurasi sebesar 0.9807692307692307.

4. KNN

```
[98.08, 98.08, 98.08, 98.08, 97.12]
```

Pada metode KNN didapat akurasi sebesar 98,08.

5. Naïve Bayes Gaussian

Accuracy:0.8557692307692307
```
[[32  7]
 [ 8 57]]
```

Prediksi positif benar sebanyak 32, positif salah sebanyak 7, negatif salah sebanyak 8 dan negatif benar sebanyak 57. Dari angka tersebut kita juga dapat menghitung presisi, akurasi dan penarikan kesimpulan. Pada metode Naïve Bayes Gaussian didapat akurasi sebesar 0.8557692307692307.

6. Decision Trees

Accuracy: 0.9711538461538461
```
[[39  2]
 [ 1 62]]
```

Prediksi positif benar sebanyak 39, positif salah sebanyak 2, negatif salah sebanyak 1 dan negatif benar sebanyak 62. Dari angka tersebut kita juga dapat menghitung presisi, akurasi dan penarikan kesimpulan. Pada metode Naïve Bayes Gaussian didapat akurasi sebesar 0.9711538461538461.

7. Random Forest

Accuracy: 0.9807692307692307
```
[[39  1]
 [ 1 63]]
```

Prediksi positif benar sebanyak 39, positif salah sebanyak 1, negatif salah sebanyak 1 dan negatif benar sebanyak 63. Dari angka tersebut kita juga dapat menghitung presisi, akurasi dan penarikan kesimpulan. Pada metode Naïve Bayes Gaussian didapat akurasi sebesar 0.9807692307692307.

5. Kesimpulan

```
Logistic regression: 0.8942307692307693
svmlinear: 0.9038461538461539
svmrbf: 0.9807692307692307
```

```
knn: [98.08]
naive bayes: 0.8557692307692307
Decision tress: 0.9711538461538461
Random forest: 0.9807692307692307
```

Berdasarkan data diatas didapat model yang tepat yaitu SVM, KNN dan Random Forest sebanyak 98%.