Laporan Final Project DSC UNIPA

“ Breast Cancer Coimbra Classification”

Oleh :

Afiful Fuad (192400025)

PROGRAM STUDI STATISTIKA

FAKULTAS SAINS TEKNOLOGI

UNIVERSITAS PGRI ADI BUANA SURABAYA

TAHUN 2021 / 2022

a. Deskripsi Data

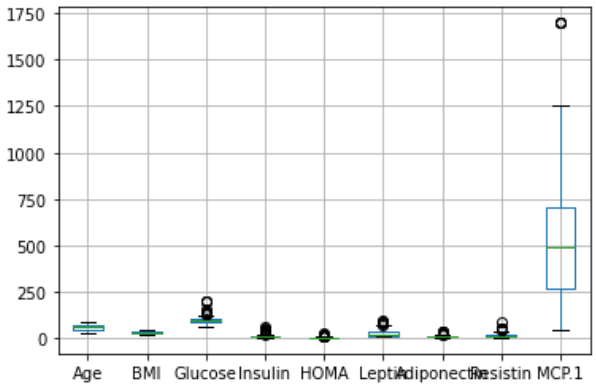
Ada 10 prediktor, semuanya kuantitatif, dan variabel dependen biner, yang menunjukkan ada tidaknya kanker payudara.Prediktor adalah data antropometrik dan parameter yang dapat dikumpulkan dalam analisis darah rutin. Model prediksi berdasarkan prediktor tersebut, jika akurat, berpotensi dapat digunakan sebagai biomarker kanker payudara.

Attribute Information:
Quantitative Attributes:
Age (years)
BMI (kg/m2)
Glucose (mg/dL)
Insulin (µU/mL)
HOMA
Leptin (ng/mL)
Adiponectin (µg/mL)
Resistin (ng/mL)
MCP-1(pg/dL)

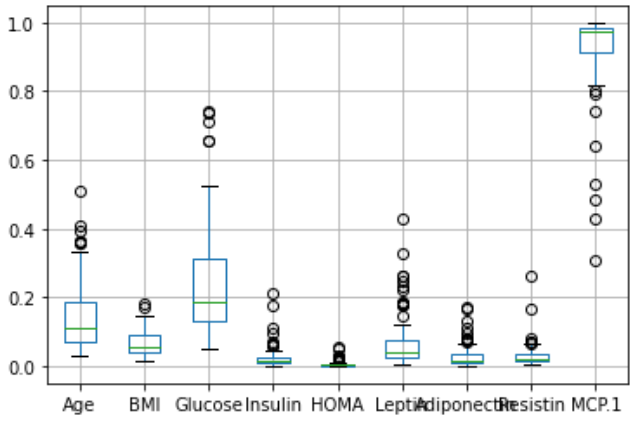
Labels:
1=Healthy controls
2=Patients

b. Data Prepocessing

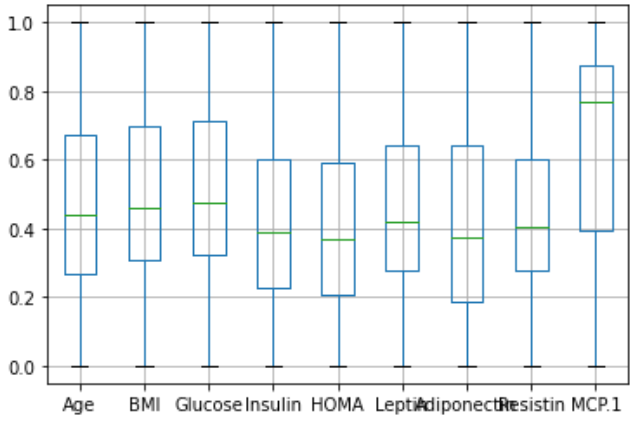
- Pertama, Perbaiki Ketidakseimbangan Kelas menggunakan SMOTE karena data tidak seimbang. Yang mana jumlah kelas 1 (Healthy controls) sebanyak 52, sedangkan kelas 2 (Patients) sebanyak 64, maka dari itu lakukan SMOTE. SMOTE adalah teknik oversampling di mana sampel sintetis dihasilkan untuk kelas minoritas.
- Langkah kedua adalah deteksi outlier,



Dari boxplot diatas, dapat disimpulkan bahwa ada outlier, oleh karena itu lakukan normalisasi data, dan dihasilkan seperti gambar dibawah



Ternyata setelah di normalisasi data masih terdapat outlier, Langkah selanjutnya adalah dengan menggunakan minmaxScaler:



Karena data sudah tidak ada lagi outlier, maka lanjut Langkah selanjutnya

c. Modelling dan Evaluating

- Cross Validation

  Dari running data di Jupiter Notebook didapatkan output:

```
Cross Validation Score: 0.65 (+/- 0.24)
```

- Splitting Results dengan K-Fold Cross Validation vs Stratified K-Fol

```
Splitting Results using K-Fold Cross Validation
Train -  [38 64]   |   Test -  [26]
Train -  [38 64]   |   Test -  [26]
Train -  [64 38]   |   Test -  [ 0 26]
Train -  [64 39]   |   Test -  [ 0 25]
Train -  [52 51]   |   Test -  [12 13]


Splitting Results using Stratified K-Fold Cross Validation
Train -  [51 51]   |   Test -  [13 13]
Train -  [51 51]   |   Test -  [13 13]
Train -  [51 51]   |   Test -  [13 13]
Train -  [51 52]   |   Test -  [13 12]
```

```
Train -  [52 51]   |   Test -  [12 13]
```

- Menampilkan akurasi dari data train dan data testing, didapatkan:

```
Average Accuracy Train : 0.7674852465257949
Average Accuracy Test : 0.6544615384615385
```

d. Kesimpulan

Dari running data mulai dari awal hingga akhir disimpulkan bahwa model klasifikasi dari data *Breast Cancer Coimbra* memiliki rata-rata akurasi pada Training sebesar 76,75%, dan memiliki rata-rata akurasi Testing sebesar 65,47%