Tugas Akhir Mata Kuliah
Pengantar Data Science
Breast Cancer

Disusun oleh :
Josie Esthaliani / 6181801008
Florenthia Kezia Kurniawan / 6181801028

Teknik Informatika
Fakultas Teknologi Informasi dan Science
Universitas Katolik Parahyangan
2020

Daftar Isi

Bab 1 (Pendahuluan)

1.1 Latar Belakang

Kanker payudara merupakan kondisi ketika sel kanker yang menyerang jaringan payudara. Pada masa sekarang ini sekitar 1 dari 8 wanita dapat mengalami kanker payudara. Selain itu kanker payudara sudah menjadi faktor penyebab kematian tertinggi di kalangan wanita. Faktor penyebab dari penyakit ini pun beragam dan beberapa diantaranya sangat dekat dengan kehidupan manusia modern seperti saat ini. Penyakit ini sendiri dapat dikelompokan menjadi 2 yaitu kanker yang ganas dan jinak kedua kondisi tersebut tentu saja dapat memberikan harapan yang berbeda pada setiap penderita.

Eksperimen dengan membuat visualisasi dapat membantu mempermudah untuk melihat berbagai hal yang dapat digali dalam sebuah data. Selain itu klasifikasi dan prediksi juga dapat dilakukan dari data yang ada untuk menemukan hubungan dan prediksi dari data-data yang sudah pernah ada sebagai pola penentuan.

1.2 Tujuan

Kondisi dari seberapa parah dari penyakit kanker payudara dapat mempengaruhi kebutuhan untuk mengobati penyakit maupun penanganannya maka dari itu pada eksperimen kali ini akan dilakukan pengolahan data untuk melihat hal apa saja yang dapat menjadi acuan pengelompokan kondisi kanker payudara dan beberapa hal yang berhubungan dengan kanker payudara lainnya. Eksperimen ini akan membentuk visualisasi dengan menggunakan bantuan dari Tableau dan python.

Selain melihat dari hasil olahan data eksperimen kali ini juga akan membuat sebuah model klasifikasi yang nantinya akan menghasilkan sarana untuk melakukan prediksi kondisi diagnosa kanker. Sehingga akan lebih memudahkan untuk mengetahui apakah ganas atau jinak kanker dari seseorang hanya dengan memasukan beberapa informasi. Eksperimen ini akan melakukan klasifikasi dengan menggunakan algoritma KNN atau K-nearest neighbor untuk melakukan pengelompokan kondisi diagnosa kanker tersebut.

1.3 Penyiapan Data

Dataset yang akan kami gunakan merupakan data Wisconsin Diagnostic Breast Cancer (WDBC) atau data kanker payudara Wisconsin yang berasal dari https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).
Berikut adalah detail dari atribut pada dataset wdbc yang akan kami olah :

Nama atribut	**Tipe data**	**Keterangan**
id	int	Id dari data
diagnosis	object	Diagnosa diamana M = malignant (ganas) dan B = benign (jinak)
mean_radius	float	Rata-rata radius (jarak pusat ke keliling)
mean_texture	float	Rata-rata tekstur
mean_perimeter	float	Rata-rata perimeter / garis keliling
mean_area	float	Rata-rata area
mean_smoothness	float	Rata-rata tingkat kehalusan
mean_compactness	float	Rata-rata (perimeter^2 / luas -1,0)
mean_concavity	float	Rata-rata tingkat kecekungan dari kontur
mean_concave_points	float	Rata-rata jumlah bagian cekung kontur
mean_symmetry	float	Rata-rata simetri
mean_fractal_dimension	float	Rata-rata fractal dimension
se_radius	float	Standar error radius (jarak pusat ke keliling)
se_texture	float	Standar error tekstur
se_parimeter	float	Standar error perimeter / garis keliling

se_area	float	Standar error area
se_smoothness	float	Standar error tingkat kehalusan
se_compactness	float	Standar error (perimeter^2 / luas -1,0)
se_concavity	float	Standar error tingkat kecekungan dari kontur
se_concave_points	float	Standar error jumlah bagian cekung kontur
se_symmetry	float	Standar error simetri
se_fractal_dimension	float	Standar error fractal dimension
worst_radius	float	Worst radius (jarak pusat ke keliling)
worst_texture	float	Worst tekstur
worst_perimeter	float	Worst perimeter / garis keliling
worst_area	float	Worst area
worst_smoothness	float	Worst tingkat kehalusan
worst_compactness	float	Worst (perimeter^2 / luas -1,0)
worst_concavity	float	Worst tingkat kecekungan dari kontur
worst_concave_points	float	Worst jumlah bagian cekung kontur
worst_symmetry	float	Worst simetri
worst_fractal_dimension	float	Worst fractal dimension

Data wdbc berikut ini sudah tidak memiliki missing value tetapi masih memiliki beberapa data outlier, maka dari itu kami melakukan pembersihan data dengan membuat data yang masih memiliki outliers dengan menggunakan rumus z score.

Selain melakukan pembersihan data kami melakukan transformasi pada atribut diagnosis yang bertipe object yang berisi M dan B menjadi int untuk beberapa keperluan

eksperimen kami. Maka dari itu setelah melakukan transform dengan label encoder dihasilkan :

B = 0

M = 1

Semua kode untuk melakukan pembersihan, transformasi, dan eksperimen berikut ini dapat dilihat pada TA_7.py. Pada eksperimen kali ini kami menggunakan tools Tableau dan python.

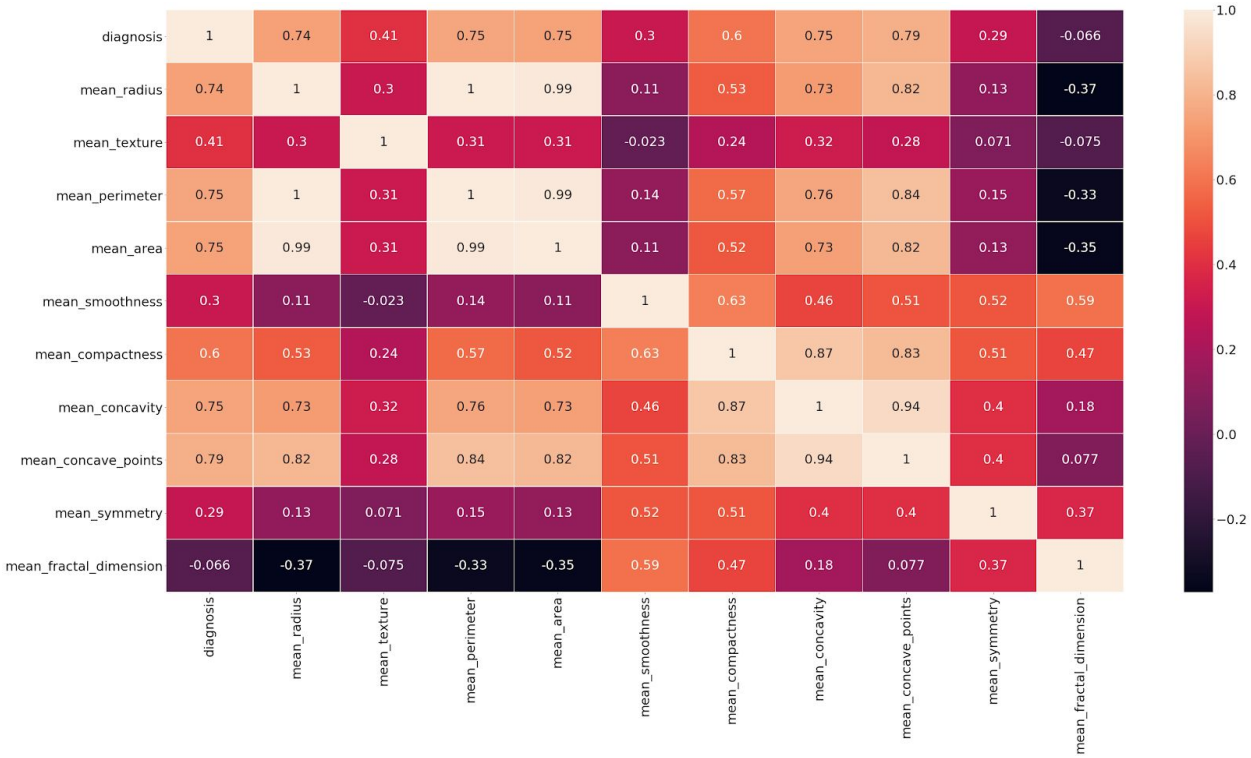
Bab 2 (Isi)

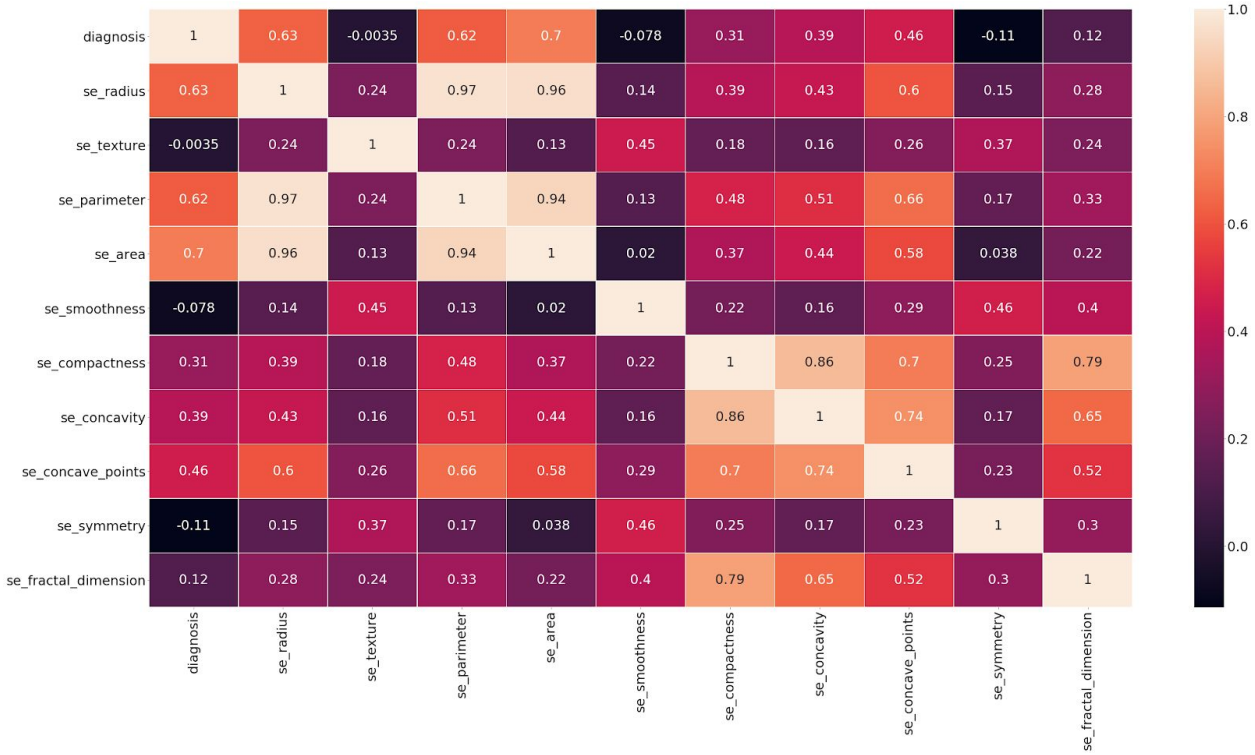
2.1 Penggalian Insights

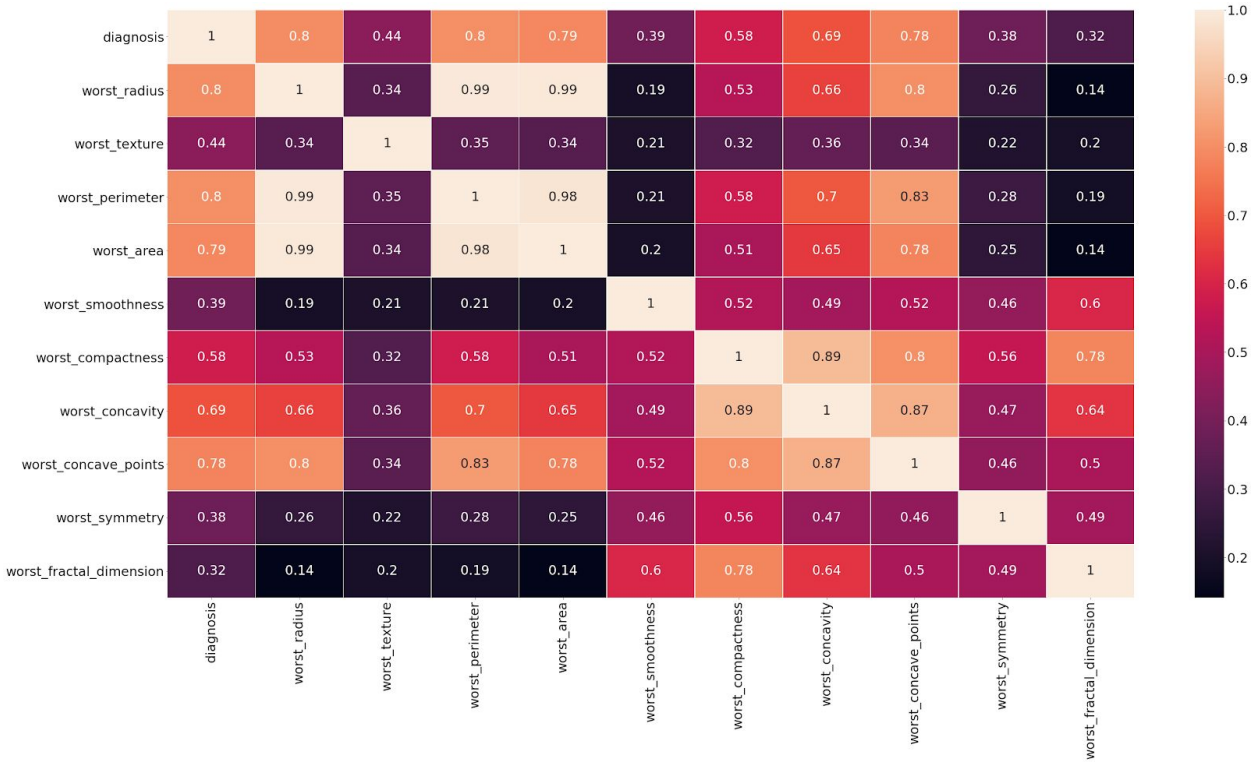
Dalam rangka mengetahui hal-hal penyebab atau hal-hal yang mempengaruhi kanker pada seseorang, kami membuat beberapa insight yang akan memberikan informasi mengenai kanker, antara lain sebagai berikut :

2.1.1. Apa saja yang paling berpengaruh dari sel nukleus terhadap diagnosa kanker? Apa yang paling bisa mengelompokan kanker seseorang termasuk kanker ganas atau kanker jinak?

Untuk menjawab insight diatas kami menggunakan visualisasi heatmap berdasarkan data yang sudah disiapkan.







Melalui heatmap diatas dapat dilihat bahwa atribut **mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter,**

worst_area, dan worst_concave_points memiliki nilai yang paling mendekati 1 jika dibandingkan dengan atribut lainnya. Karena jika nilai semakin mendekati 1 atau -1 maka dikatakan bahwa atribut tersebut berpengaruh terhadap atribut target. Maka dapat diketahui bahwa atribut **mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter, worst_area, dan worst_concave_points** merupakan atribut yang paling berpengaruh terhadap atribut target. Untuk lebih meyakinkan, maka kami melakukan pencarian nilai R^2 dan MAE melalui **regresi**.

Berikut ini merupakan tabel yang berisi nilai R^2 dan MAE untuk setiap atribut :

Atribut	**R^2**	**MAE**
mean_radius dan diagnosa	0.3488320973572109	0.2556002006753207
mean_texture dan diagnosa	0.1066200261458542	0.33739659795570076
mean_perimeter dan diagnosa	0.37706025768596607	0.25009360011214093
mean_area dan diagnosa	0.3596330896287683	0.2427917181032283
mean_smoothness dan diagnosa	0.057561718517956106	0.37788117749497047
mean_compactness dan diagnosa	0.2825453289731782	0.28101136059141657
mean_concavity dan diagnosa	0.5649253806530173	0.20716481855562296
mean_concave_points dan diagnosa	0.5948944855419112	0.19514072287384873
mean_symmetry dan diagnosa	-0.02627119497232333	0.3881026820306393

mean_fractal_dimension dan diagnosa	-0.09650658391465905	0.42166047216225017
se_radius dan diagnosa	0.24546508603497408	0.2894967424841747
se_texture dan diagnosa	-0.09564215300614576	0.42389133722859274
se_parimeter dan diagnosa	0.2417402890713003	0.28918728560431517
se_area dan diagnosa	0.3541064157995286	0.2640872706253759
se_smoothness dan diagnosa	-0.11590520628119028	0.4244815033457778
se_compactness dan diagnosa	0.0015653819613381525	0.3806446630727261
se_concavity dan diagnosa	0.0331032701776659	0.3522981642200149
se_concave_points dan diagnosa	0.14643518099757025	0.32857115135648257
se_symmetry dan diagnosa	-0.08142954783009104	0.41811262092716056
se_fractal_dimension dan diagnosa	-0.06622347764375114	0.41599502401636573
worst_radius dan diagnosa	0.4784129969309605	0.2305914928349405
worst_texture dan diagnosa	0.2097497982600912	0.3123155936918188
worst_perimeter dan diagnosa	0.49890825939571415	0.22809626951337908
worst_area dan diagnosa	0.4710525359742621	0.2220126090761169
worst_smoothness dan	0.18819727890853155	0.3430082053382407

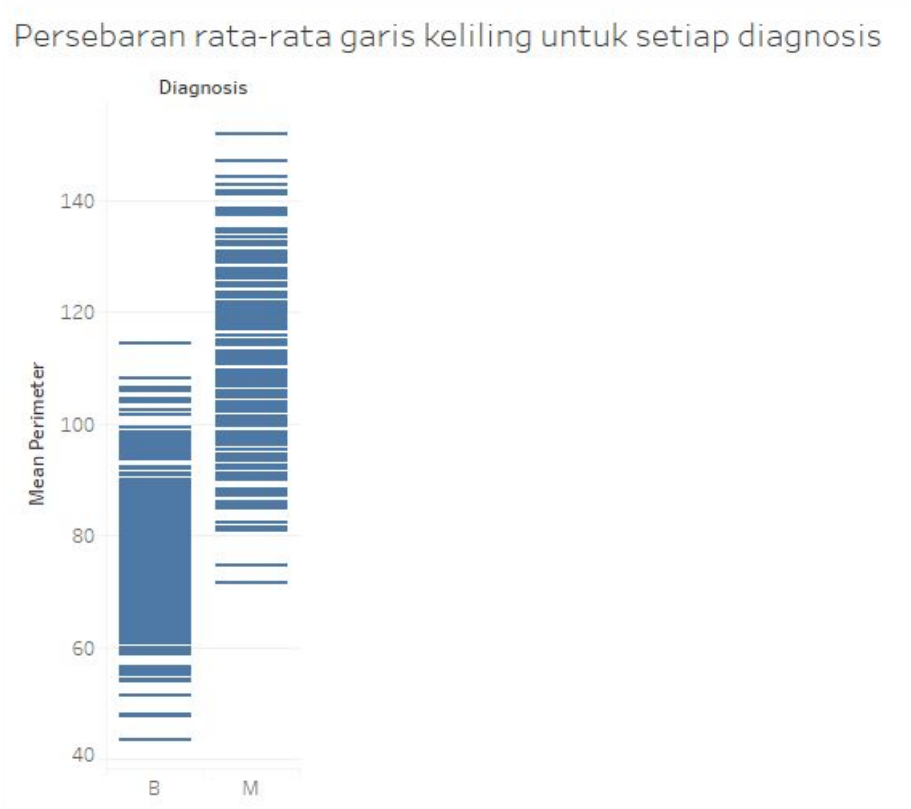
diagnosa		
worst_compactness dan diagnosa	0.265887116862928	0.28079704784348397
worst_concavity dan diagnosa	0.41252391811933764	0.2426622260343118
worst_concave_points dan diagnosa	0.5536543989270011	0.22294410902879566
worst_symmetry dan diagnosa	0.11169728974996218	0.35799628623741053
worst_fractal_dimension dan diagnosa	0.03793828016192757	0.3775609488073335

Melalui hasil R^2 dan MAE di atas terlihat bahwa atribut **mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter, worst_area, dan worst_concave_points** memiliki R^2 yang paling mendekati 1 jika dibandingkan atribut lain dan MAE yang kecil. Diketahui bahwa jika R^2 semakin mendekati 1 atau -1 dan MAE semakin kecil itu menunjukkan bahwa atribut prediktor tersebut berpengaruh atau memiliki relasi yang kuat dengan atribut target. **Melalui visualisasi heatmap dan hasil regresi dapat disimpulkan bahwa rata-rata garis keliling inti sel, rata-rata luas inti sel, rata-rata kecekungan, rata-rata jumlah bagian cekung dari kontur nukleus (inti sel), rata-rata dari tiga nilai tertinggi radius inti sel, rata-rata dari tiga nilai tertinggi garis keliling inti sel, rata-rata dari tiga nilai tertinggi luas inti sel, rata-rata dari tiga nilai tertinggi jumlah bagian cekung dari kontur inti sel berpengaruh terhadap diagnosa kanker sehingga dapat menentukan apakah seseorang didiagnosa kanker berat atau tidak**

2.1.2. Bagaimana pengaruh persebaran setiap atribut terhadap diagnosanya?

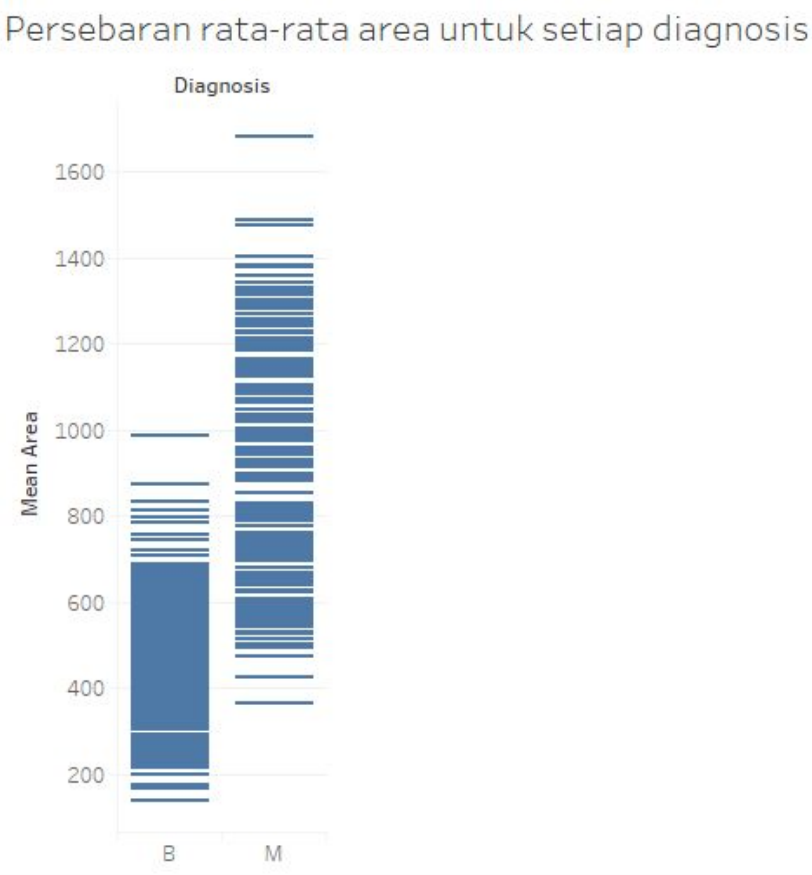
Berdasarkan heatmap dan hasil dari perhitungan R^2 dan MAE berikut adalah persebaran untuk setiap atribut dengan target.

A. Persebaran rata-rata garis keliling / perimeter untuk setiap diagnosa.



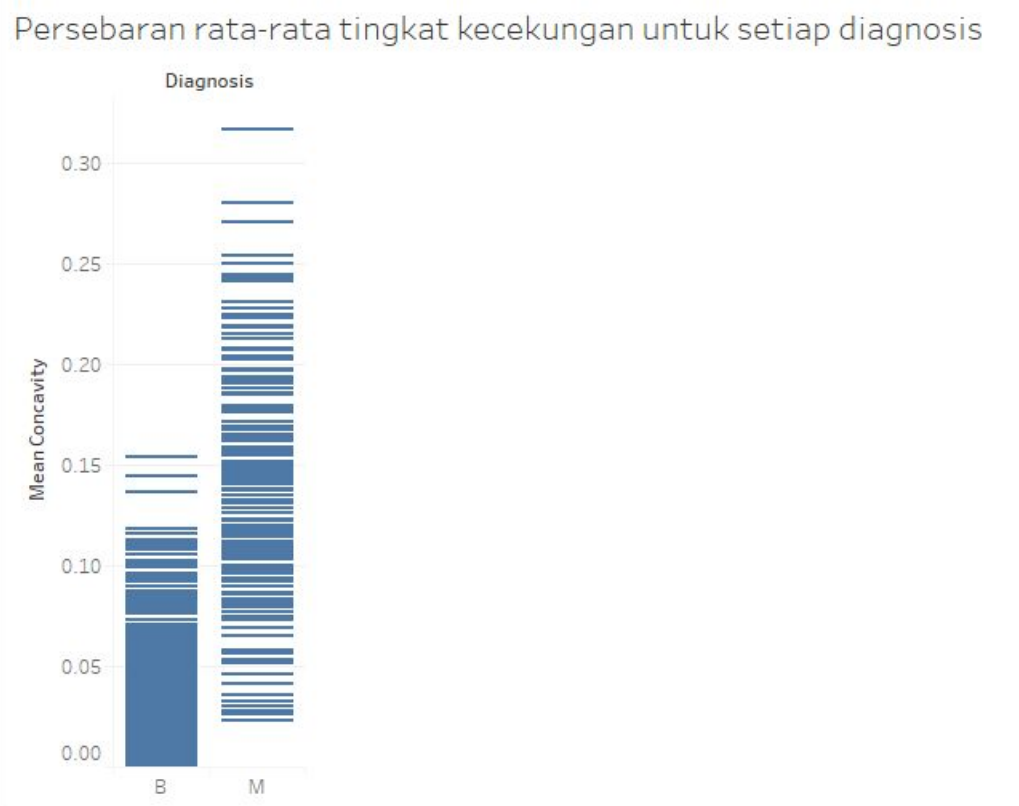
Melalui visualisasi diatas dapat terlihat persebaran pada data rata-rata perimeter/ garis keliling untuk setiap diagnosa dimana untuk diagnosa B atau jinak angka rata-rata perimeternya lebih tersebar di sekitar angka 60-100 dengan rata-rata tertinggi hampir mencapai 120, bahkan data terendah untuk rata-rata perimeternya berada pada bagian jinak ini. Sedangkan untuk kanker ganas rata-rata perimeternya tersebar di antara rentang 70 - 160. Pada kondisi M atau ganas ini ada rata-rata tertinggi dari seluruh data. Dari hal-hal tersebut dapat disimpulkan bahwa rata-rata perimeter yang tinggi lebih memiliki kecenderungan akan di diagnosa pada kondisi kanker payudara yang ganas.

B. Persebaran rata-rata area untuk setiap diagnosa.



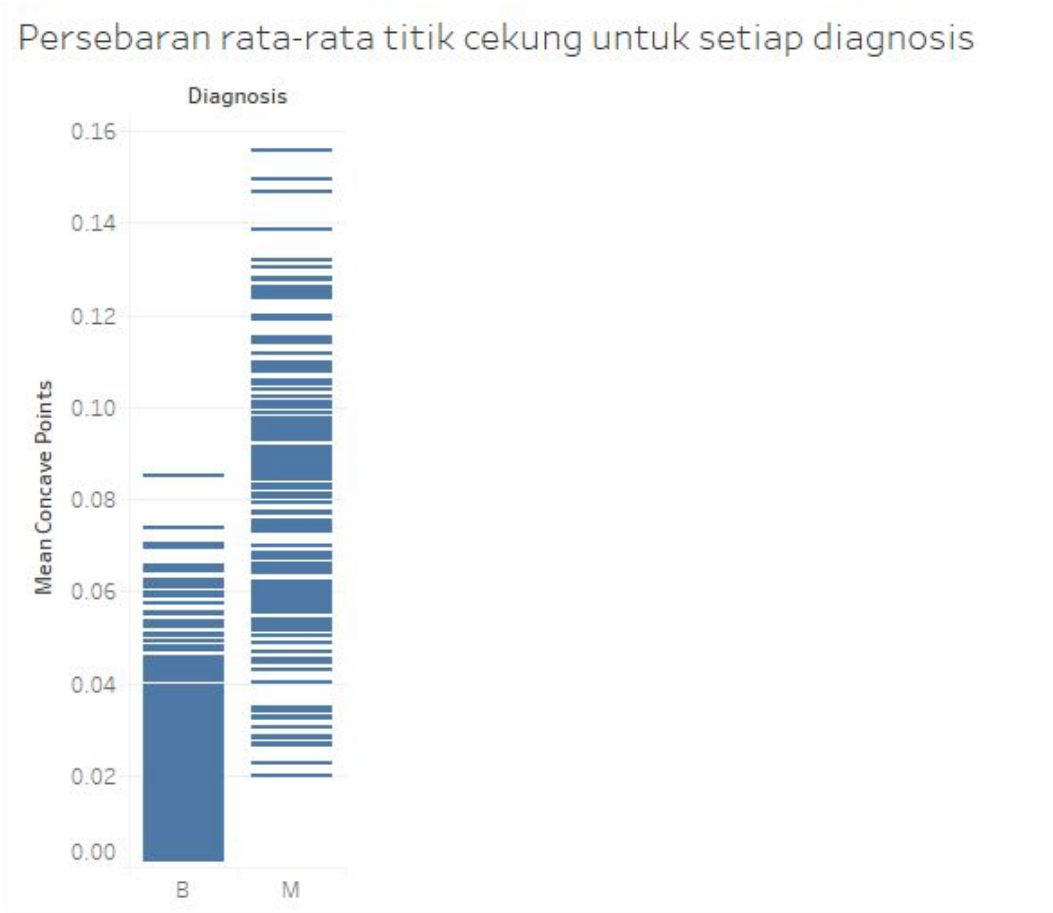
Visualisasi diatas dapat menunjukan hubungan dari rata-rata area dengan diagnosa kanker payudara yang ganas maupun jinak. Dari yang terlihat pada persebarannya untuk kanker yang jinak berada di bawah angka 1000 dan paling banyak tersebar di rentang 200-600. Sedangkan pada diagnosa ganas rata-rata area tersebar mulai dari range 300-1700 dan kondisi ini memiliki nilai rata-rata area tertinggi. Dari hal tersebut dapat disimpulkan bahwa diagnosa kanker yang jinak lebih cenderung memiliki rata-rata area dengan nilai yang rendah dibandingkan diagnosa ganas.

C. Persebaran rata-rata tingkat kecekungan / concavity untuk setiap diagnosa.



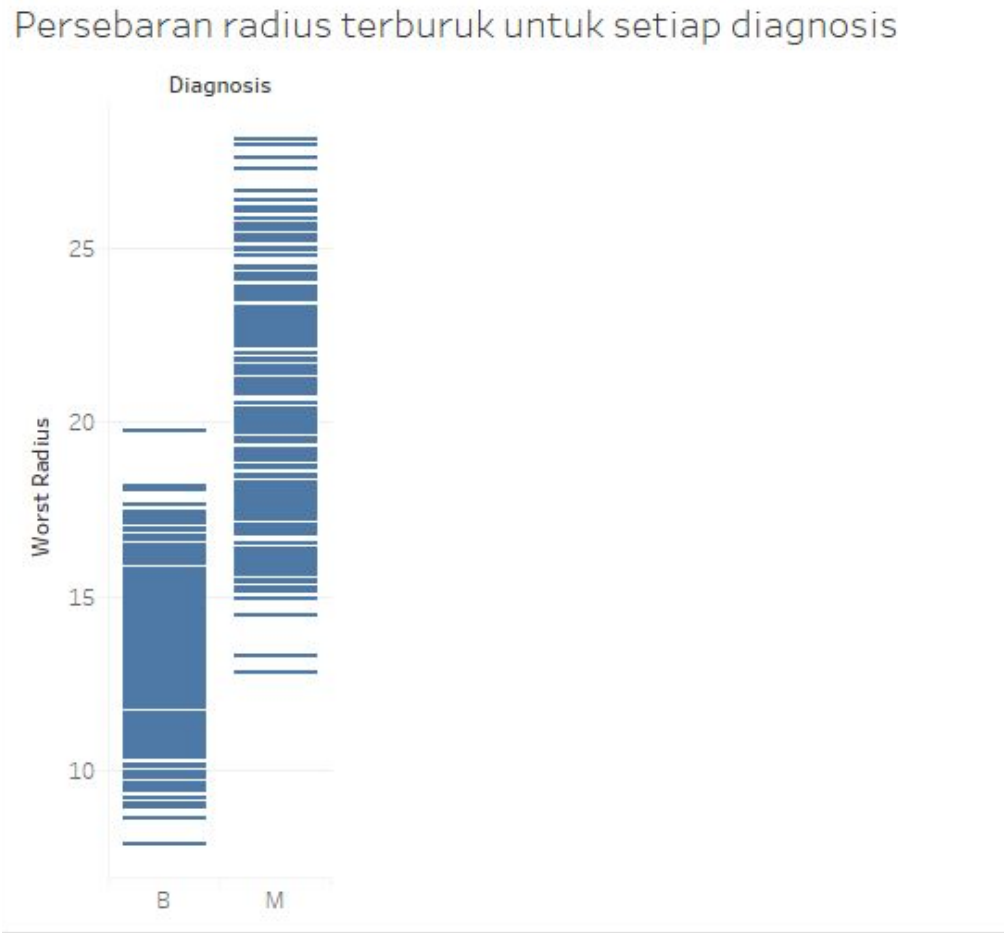
Berdasarkan visualisasi diatas dapat terlihat persebaran rata-rata tingkat kecekungan untuk diagnosa kanker payudara ganas maupun jinak. Dari persebarannya dapat terlihat bahwa rata-rata tingkat kecekungan untuk kanker jinak menyebar dari 0 sebagai nilai terendah dan nilai tertinggi berada disekitar 0.15. Sedangkan untuk diagnosa kanker ganas mulai tersebar dengan nilai terendah sekitar 0.02 sampai nilai tertingginya sekitar 0.31. Dari hal tersebut dapat disimpulkan bahwa rata-rata tingkat kecekungan sel yang tinggi lebih berpeluang terdiagnosa kanker ganas.

D. Persebaran rata-rata jumlah titik cekung / concave points untuk setiap diagnosa.



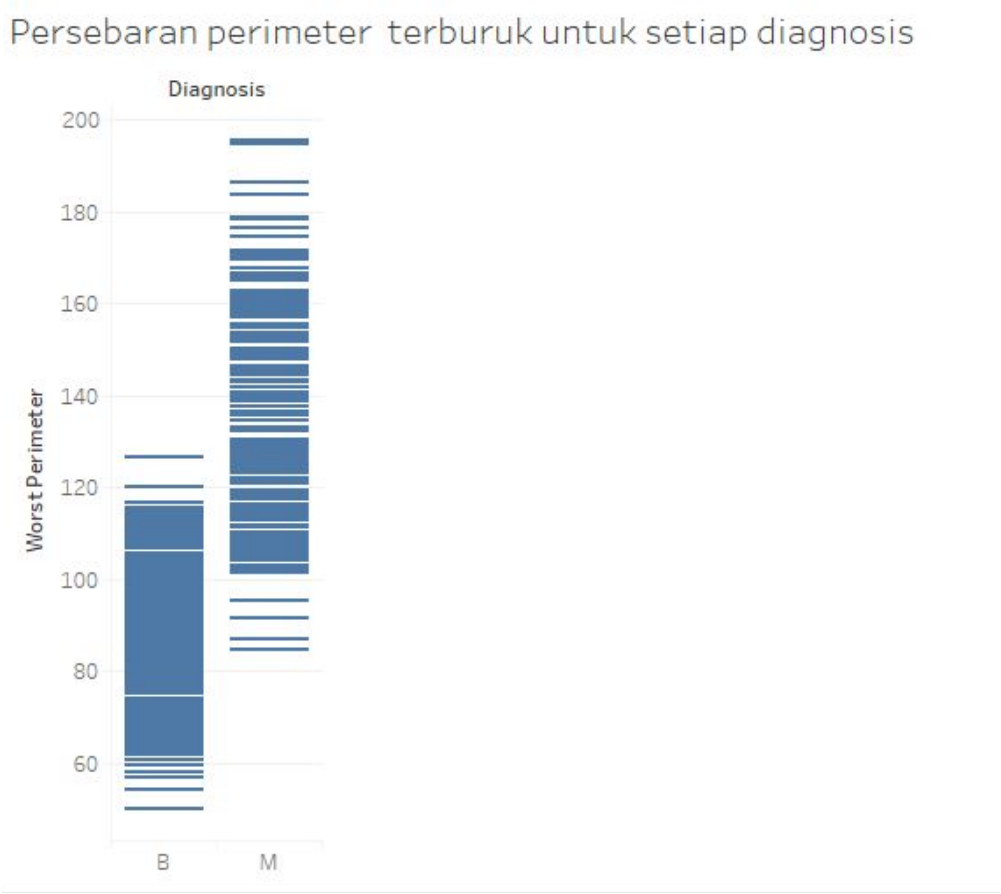
Visualisasi diatas dapat menunjukan hubungan dari rata-rata titik cekung dengan diagnosa kanker payudara yang ganas maupun jinak. Dari yang terlihat pada persebarannya untuk kanker yang jinak berada di bawah angka 0.09 dan paling banyak tersebar di rentang 0 - 0.06. Sedangkan pada diagnosa ganas rata-rata titik cekung tersebar mulai dari range 0.02 - 0.16. Nilai terendah dari seluruh data berada pada diagnosa jinak sedangkan nilai tertingginya berada pada diagnosa ganas. Sudah sangat jelas dari hal tersebut dapat disimpulkan bahwa diagnosa kanker yang jinak lebih cenderung memiliki rata-rata jumlah titik cekung dengan nilai yang rendah dibandingkan diagnosa ganas.

E. Persebaran radius terburuk untuk setiap diagnosa.



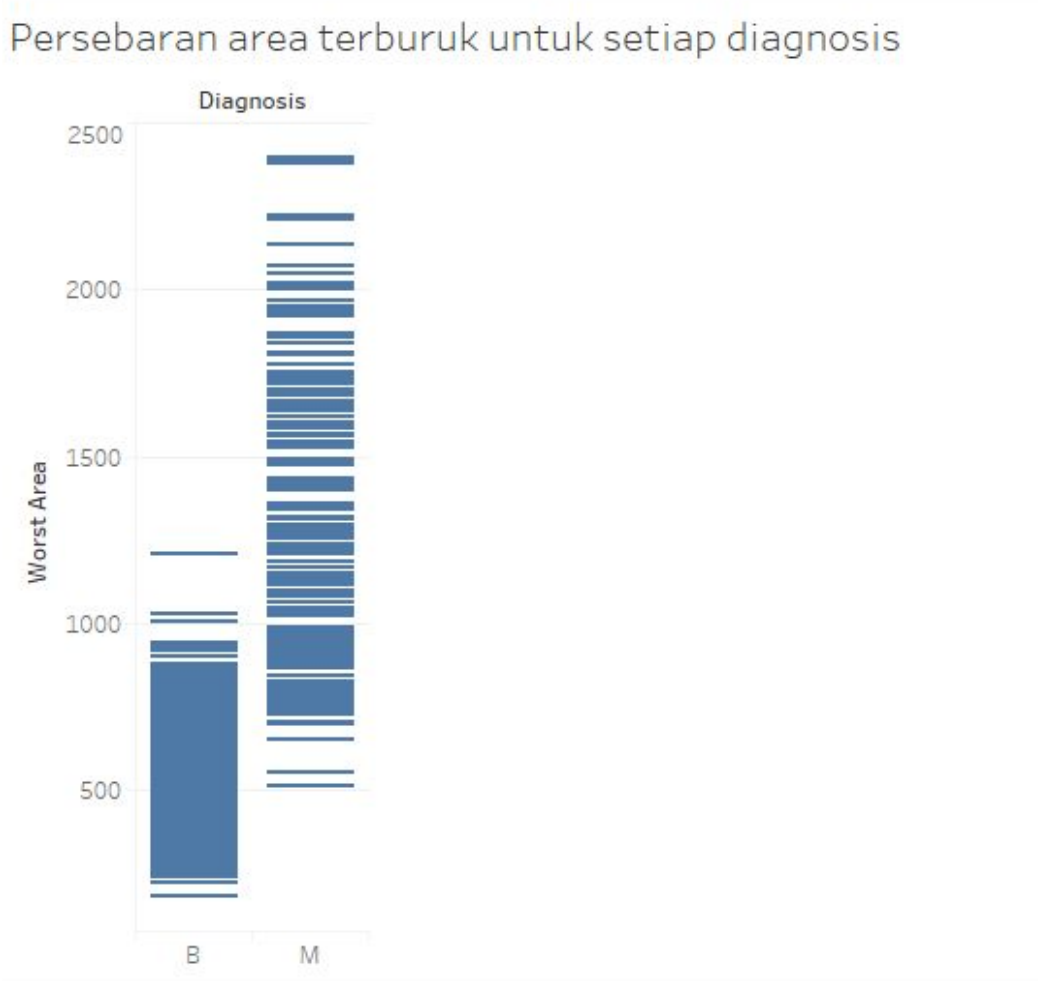
Berdasarkan visualisasi diatas dapat terlihat persebaran radius terburuk untuk diagnosa kanker payudara ganas maupun jinak. Dari persebarannya dapat terlihat bahwa radius terburuk untuk kanker jinak menyebar dari sekitar 5 sebagai nilai terendah dan nilai tertinggi berada disekitar 20. Sedangkan untuk diagnosa kanker ganas mulai tersebar dengan nilai terendah sekitar 12 sampai nilai tertingginya sekitar 28. Dari hal tersebut dapat disimpulkan bahwa radius terburuk sel yang tinggi lebih berpeluang terdiagnosa kanker ganas dibandingkan dengan pemilik radius terburuk yang rendah.

F. Persebaran perimeter terburuk untuk setiap diagnosa.



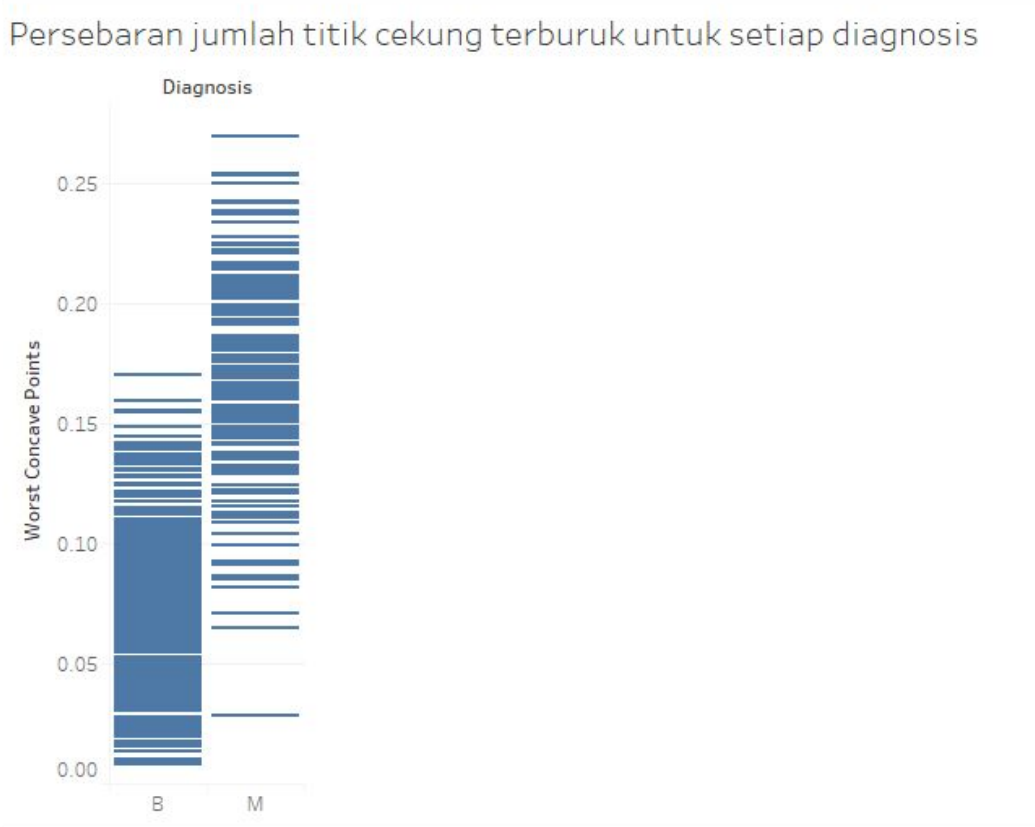
Berdasarkan visualisasi diatas dapat terlihat persebaran perimeter terburuk untuk diagnosa kanker payudara ganas maupun jinak. Dari persebarannya dapat terlihat bahwa perimeter terburuk untuk kanker jinak mulai menyebar dari sekitar 50 sebagai nilai terendah dan nilai tertinggi berada di sekitar 130. Persebarannya paling banyak tersebar di rentang 60-120. Sedangkan untuk diagnosa kanker ganas mulai tersebar dengan nilai terendah sekitar 90 sampai nilai tertingginya sekitar 190 menuju 200 dan paling banyak tersebar di rentang 100-160. Dari hal tersebut dapat disimpulkan bahwa perimeter terburuk sel yang tinggi lebih memungkinkan akan terdiagnosa kanker ganas dibandingkan dengan pemilik perimeter terburuk yang rendah.

G. Persebaran terburuk untuk setiap diagnosa.



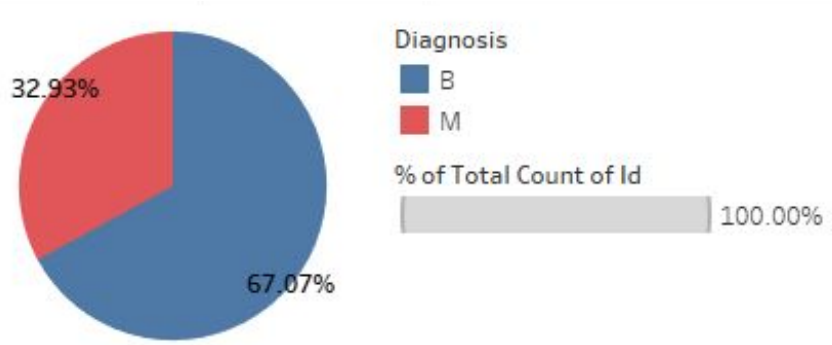
Berdasarkan visualisasi diatas dapat terlihat persebaran area terburuk untuk diagnosa kanker payudara ganas dan jinak. Dari persebarannya dapat terlihat bahwa area terburuk untuk kanker jinak mulai menyebar dari sekitar 180 sebagai nilai terendah dan nilai tertinggi berada di sekitar 1200 selain itu persebaran paling banyak berada pada rentang 200-900. Sedangkan untuk diagnosa kanker payudara ganas mulai tersebar dengan nilai terendah sekitar 500 sampai nilai tertingginya hampir menyentuh angka 2500 dan paling banyak tersebar di rentang 700 - 1900. Dari hal tersebut dapat disimpulkan bahwa area terburuk sel yang rendah memiliki peluang untuk diberikan diagnosa kanker payudara jinak dibandingkan sel dengan area terburuk yang tinggi.

H. Persebaran titik cekung terburuk untuk setiap diagnosa.



Berdasarkan visualisasi diatas dapat terlihat persebaran titik cekung terburuk untuk diagnosa kanker payudara ganas maupun jinak. Dari persebarannya dapat terlihat bahwa nilai titik cekung terburuk untuk kanker jinak mulai menyebar dari 0 yang pastinya merupakan nilai terendah dan nilai tertinggi berada di sekitar 0.17 Untuk persebaran paling banyak berada pada rentang 0.02 - 0.11. Sedangkan untuk diagnosa kanker payudara ganas mulai tersebar dengan nilai terendah sekitar 0.02 sampai nilai tertingginya sekitar 0.27 dan paling banyak tersebar di rentang 0,12 - 0,21. Dari hal tersebut dapat disimpulkan bahwa nilai titik cekung terburuk sel yang rendah memiliki peluang untuk diberikan diagnosa kanker payudara jinak dibandingkan sel dengan titik cekung terburuk sel yang tinggi.

2.1.3. Bagaimana persentase banyaknya pasien kanker payudara untuk setiap diagnosa?



Kami menggunakan pie chart sebagai visualisasi untuk melihat persentase dari banyaknya pasien penderita kanker payudara untuk setiap diagnosa. Seperti yang sudah tertulis pada keterangan di samping pie chart tersebut bahwa merah merupakan M atau ganas sedangkan biru merupakan B atau jinak. Ternyata persentase dari penderita dengan diagnosa jinak lebih banyak dengan persentase 67.07% sedangkan M atau ganas hanya memiliki 32.93%.

2.2 Klasifikasi

Pada bagian sebelumnya kami telah mendapat beberapa informasi berdasarkan insight yang kami buat. Pada bagian ini kami akan menggunakan algoritma kNN untuk klasifikasi sebagai salah satu bagian dari predictive analysis. Kami akan mencoba untuk memprediksi diagnosa yang dimiliki seseorang penderita kanker (apakah orang itu tergolong memiliki kanker benign / jinak atau malignant / ganas). Pada klasifikasi ini kami menggunakan data wdbc.csv.

2.2.1. Algoritma kNN

Setelah menyiapkan dan membersihkan data, hal yang pertama kami lakukan adalah melakukan pemilihan features (atribut prediktor) dengan mendapatkan setiap score menggunakan chi dari atribut tersebut, antara lain sebagai berikut :

Atribut	Scores
mean_radius	181.212
mean_texture	68.1844
mean_perimeter	1351.32
mean_area	36313.3
mean_smoothness	0.0789501
mean_compactness	3.1882
mean_concavity	14.3726
mean_concave_points	7.55728
mean_symmetry	0.129698
mean_fractal_dimension	0.00107858
se_radius	22.4268
se_texture	0.00114086

se_perimeter	148.413
se_area	5308.39
se_smoothness	0.00265583
se_compactness	0.342067
se_concavity	0.871955
se_concave_points	0.221308
se_symmetry	0.0123916
se_fractal_dimension	0.00561437
worst_radius	336.88
worst_texture	123.94
worst_perimeter	2446.32
worst_area	76151.4
worst_smoothness	0.252174
worst_compactness	11.4988
worst_concavity	28.4052
worst_concave_points	9.54746
worst_symmetry	0.630638
worst_fractal_dimension	0.126083

Melalui hasil diatas, akan dipilih pasangan atribut / features berdasarkan scores terbaik. Pada eksperimen kali ini kami akan memilih features sebagai berikut :

No.	**Features (pasangan atribut)**	**Akurasi model**

1	mean_perimeter, mean_area, se_area, worst_radius, worst_perimeter, worst_area	0.9194630872483222
2	mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter, worst_area, dan worst_concave_points	0.9731543624161074

Karena akurasi model terbesar dimiliki oleh features nomor 2 **(mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter, worst_area, dan worst_concave_points), yaitu dengan akurasi sebesar 97.3%** dengan menggunakan kNN, maka features tersebutlah yang akan dijadikan sebagai atribut prediktor untuk memprediksi nilai diagnosa dari data baru.

Selanjutnya kami membuat data baru yang terdiri dari 5 baris dan memiliki atribut seperti pada features terpilih, lalu kami akan memprediksi termasuk diagnosa (benign atau malignant) apakah kanker yang dialami orang pada data tersebut :

mean_perimeter	mean_area	mean_concavity	mean_concave_points	worst_radius	worst_perimeter	worst_area	worst_concave_points	class
132.4	1123.0	0.2065	0.1118	0.0409	0.01284	20.96	151.7	M
94.74	684.5	0.4727	1.24	0.1464	0.2914	0.1609	0.08216	M
60.34	273.9	0.02956	0.02076	10.23	65.13	314.9	0.06227	B
133.9	1210.0	0.2089	0.1130	0.0408	0.01390	21.19	150.9	M
61.18	274.1	0.0301	0.02921	10.90	63.12	315.2	0.0589	B

Setelah membuat data baru, kami dapat memprediksi diagnosa kanker dari seseorang. diagnosa B melambangkan benign atau jinak, diagnosa M melambangkan malignant atau ganas). Dengan demikian hasil prediksi ini akurat sebesar 97.3%.

Bab 3 (Kesimpulan)

3.1 Kesimpulan

Eksperimen terhadap data wisconsin diagnostic breast cancer ini menghasilkan beberapa visualisasi dimana ditemukan bahwa berdasarkan heatmap terhadap target diagnosis hal yang mempengaruhi diagnosis adalah mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter, worst_area, dan worst_concave_points dengan relasi positif. Berdasarkan hasil dari eksperimen tersebut juga kami mencoba memperlihatkan persebaran dari data berpengaruh tersebut. Setelah dilihat dari seluruh visualisasi persebaran tersebut semuanya memiliki kesimpulan yang sesuai dengan relasi positif, dimana semakin besar nilai atribut berpengaruh maka semakin besar juga peluang berada pada diagnosa ganas dan juga sebaliknya.

Eksperimen dilanjutkan dengan klasifikasi menggunakan algoritma kNN dimana setiap atribut dilihat scorenya lalu dibentuk featuresnya berdasarkan nilai score terbaik. Setiap features yang dibuat kemudian dicek akurasinya. Pada eskperimen kali ini akurasi tertinggi dimiliki oleh pasangan atribut mean_perimeter, mean_area, mean_concavity, mean_concave_points, worst_radius, worst_perimeter, worst_area, dan worst_concave_points dengan akurasi sebesar 97.3%. Atribut ini akan digunakan untuk membuat insight. Melalui predictive analysis dengan kNN ini, kami dapat memprediksi diagnosis kanker seseorang berdasarkan data baru yang kami buat