Judul : **Klasifikasi Kanker Payudara Menggunakan Random Forest**

Anggota Kelompok : **1. Resti Ramadhani (123220147) (DS-E)**
2. Jeslyn Vicky Hanjaya (123220150) (DS-E)
3. Faiza Nur Rafida (123220159) (DS-D)

1. Business Understanding & Analytic Approach

Kanker payudara adalah salah satu penyakit paling umum yang menyerang wanita di seluruh dunia, sehingga deteksi dini menjadi sangat penting untuk meningkatkan peluang kesembuhan dan menurunkan angka kematian. Tingkat akurasi dalam membedakan sel jinak (*benign*) dan ganas (*malignant*) menjadi tantangan utama karena proses deteksi manual membutuhkan waktu lama dan rentan terhadap kesalahan. Dengan memanfaatkan data historis dari *Breast Cancer Wisconsin* dataset, kami dapat mengembangkan sistem yang lebih cepat dan akurat untuk klasifikasi sel kanker berdasarkan berbagai fitur seluler. Sehingga memahami hubungan antara fitur seluler dan sifat sel memungkinkan pembangunan model klasifikasi yang efisien, mempercepat diagnosis, dan mengurangi kesalahan. Dengan sistem ini, peluang kesembuhan dapat ditingkatkan dan angka kematian dapat dikurangi secara signifikan.

Analytic Approach:
Kami memutuskan untuk membangun model *supervised classification* menggunakan algoritma seperti Random Forest untuk memprediksi apakah sel bersifat jinak atau ganas. Model ini akan dilatih menggunakan Breast Cancer Wisconsin dataset dengan berbagai fitur seluler sebagai input. Pendekatan ini bertujuan untuk menghasilkan sistem yang cepat, akurat, dan dapat diandalkan untuk mendukung diagnosis medis.

Berikut merupakan usulan task dalam project kami:
- Mendefinisikan Business Understanding tentang masalah yang dibahas.
- Menentukan Analytic Approach yang akan digunakan. Pada topik ini, kami menggunakan *Predictive Analytic*.
- Menganalisa dan mengumpulkan data yang dibutuhkan.
- Menentukan jenis model yang digunakan. Pada topik ini, kami menggunakan model *Klasifikasi Data Supervised*.
- Menentukan Algoritma yang digunakan. Pada topik ini, kami menggunakan algoritma *Random Forest*.
- Mendeskripsikan hasil yang telah didapat.

2. Data Requirement & Data Collection

Sumber data : Wolberg, William. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. *https://doi.org/10.24432/C5HP4Z*

Kami mengambil data dari dataset Breast Cancer Wisconsin (Original) yang kami dapatkan dari UCI Machine Learning Repository. Dataset ini berisi informasi terkait karakteristik seluler untuk mengidentifikasi apakah tumor bersifat *malignant* (ganas) atau *benign* (jinak). Dataset ini terdiri dari 699 sampel, tetapi kami hanya menggunakan 680 data untuk analisis. Hal tersebut dikarenakan terdapat *missing value* sebanyak 16 data (699-16 = 683), dikarenakan untuk memudahkan proses *splitting* data *training* dan testing maka data yang diambil sebanyak 680 data saja.

Dataset ini memiliki kolom sebanyak sebelas kolom dan berisi data numeric. Dalam hal ini kami memilih semua kolom dalam dataset kecuali *Sample_code_number*, yaitu: *Clump_thickness, Uniformity_of_cell_size, Uniformity_of_cell_shape, Marginal_adhesion, Single_epithelial_cell_size, Bare_nuclei, Bland_chromatin, Normal_nucleoli,* dan *Mitoses.* Kolom-kolom ini dipilih karena informasi yang dikandungnya dapat digunakan untuk mengklasifikasikan jenis kanker (*Class*), baik jinak maupun ganas, berdasarkan karakteristik seluler yang diamati. Hal ini memberikan wawasan yang penting dalam mendukung diagnosis kanker payudara lebih cepat dan akurat, sehingga pengambilan keputusan klinis dapat dilakukan dengan lebih efektif.

Deskripsi dan Preview Data:

Sample_code_number	Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4
1056784	3	1	1	1	2	1	2	1	1	2
1057013	8	4	5	1	2	?	7	3	1	4
1059552	1	1	1	1	2	1	3	1	1	2
1065726	5	2	3	4	2	7	3	6	1	4
1066373	3	2	1	1	1	1	2	1	1	2

- *Clump_thickness*: Ketebalan kelompok sel yang diukur dari sampel. Nilai tinggi dapat menunjukkan potensi kanker.
- *Uniformity_of_cell_size*: Ukuran sel yang seragam. Ketidakseragaman tinggi dapat mengindikasikan keganasan.
- *Uniformity_of_cell_shape*: Bentuk sel yang seragam. Mirip dengan ukuran, ketidakseragaman tinggi adalah indikator keganasan.
- *Marginal_adhesion*: Kemampuan sel untuk menempel satu sama lain. Sel kanker sering memiliki nilai adhesi yang rendah.
- *Single_epithelial_cell_size*: Ukuran sel epitel individu. Nilai yang lebih besar dari normal dapat menunjukkan kanker.
- *Bare_nuclei*: Jumlah inti sel yang terlihat kosong (tanpa sitoplasma). Inti kosong sering terlihat pada sel kanker.
- *Bland_chromatin*: Kromatin pada inti sel dengan penampilan halus atau tidak terlalu beragam. Ketidaknormalan di sini sering terkait kanker.
- *Normal_nucleoli*: Jumlah nukleolus normal di inti sel. Nukleolus besar atau banyak dapat menunjukkan aktivitas kanker.
- *Mitoses*: Aktivitas mitosis (pembelahan sel). Jumlah mitosis tinggi adalah karakteristik sel kanker.

3. Data Preparation

a. Memberikan nama kolom.

Dataset yang kami gunakan awalnya tidak memiliki nama untuk setiap kolom. Oleh karena itu, pada tahap ini kami menetapkan nama untuk masing-masing kolom dalam dataset.

Sebelum:

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2

Sesudah:

Sample_code_number	Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4

b. Menghilangkan kolom yang tidak dipakai

Menghilangkan kolom "*Sample_code_number*" karena merupakan ID saja dan tidak diperlukan untuk pemrosesan data. Data yang semulanya 11 kolom kini tersisa 10 kolom saja.

Sebelum:

Sample_code_number	Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4

Sesudah:

Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
2	1	1	1	2	1	2	1	1	2
5	3	3	3	2	3	4	4	1	4
1	1	1	1	2	3	3	1	1	2
8	7	5	10	7	9	5	5	4	4
7	4	6	4	6	1	4	3	1	4
4	1	1	1	2	1	2	1	1	2
4	1	1	1	2	1	3	1	1	2
10	7	7	6	4	10	4	1	2	4
6	1	1	1	2	1	3	1	1	2

c. Mengubah tipe data kolom

Karena kolom "*Bare_nuclei*" bertipe data character maka perlu diubah menjadi integer untuk dapat diproses dan kolom "*Class*" dirubah dari integer menjadi factor, karena sifatnya kategorikal.

Sebelum:

```
'data.frame':   699 obs. of  10 variables:
 $ Clump_thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
 $ Uniformity_of_cell_size    : int  1 4 1 8 1 10 1 1 1 2 ...
 $ Uniformity_of_cell_shape   : int  1 4 1 8 1 10 1 2 1 1 ...
 $ Marginal_adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
 $ Single_epithelial_cell_size: int  2 7 2 3 2 7 2 2 2 2 ...
 $ Bare_nuclei                : chr  "1" "10" "2" "4" ...
 $ Bland_chromatin            : int  3 3 3 3 3 9 3 3 1 2 ...
 $ Normal_nucleoli            : int  1 2 1 7 1 7 1 1 1 1 ...
 $ Mitoses                    : int  1 1 1 1 1 1 1 1 5 1 ...
 $ Class                      : int  2 2 2 2 2 4 2 2 2 2 ...
```

Sesudah:

```
'data.frame':   699 obs. of  10 variables:
 $ Clump_thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
 $ Uniformity_of_cell_size    : int  1 4 1 8 1 10 1 1 1 2 ...
 $ Uniformity_of_cell_shape   : int  1 4 1 8 1 10 1 2 1 1 ...
 $ Marginal_adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
 $ Single_epithelial_cell_size: int  2 7 2 3 2 7 2 2 2 2 ...
 $ Bare_nuclei                : int  1 10 2 4 1 10 10 1 1 1 ...
 $ Bland_chromatin            : int  3 3 3 3 3 9 3 3 1 2 ...
 $ Normal_nucleoli            : int  1 2 1 7 1 7 1 1 1 1 ...
 $ Mitoses                    : int  1 1 1 1 1 1 1 1 5 1 ...
 $ Class                      : Factor w/ 2 levels "ganas","jinak": 2 2 2 2 2 1 2 2 2 2 ...
```

d. Menghilangkan Nilai kosong NA

Dikarenakan terdapat nilai bukan angka/NA dalam data maka perlu pembersihan data. Di sini kami menghapus baris yang mengandung komponen bukan angka. Dan karena hal ini data yang semula 699 baris tersisa 683 baris, data yang mengandung NA ada 16 baris.

Sebelum:

1	1	2	1	3	NA	1	1	1	jinak
9	9	10	3	6	10	7	10	6	ganas
10	7	7	4	5	10	5	7	2	ganas
4	1	1	1	2	1	3	2	1	jinak
3	1	1	1	2	1	3	1	1	jinak
1	1	1	2	1	3	1	1	7	jinak
5	1	1	1	2	NA	3	1	1	jinak
4	1	1	1	2	2	3	2	1	jinak

140 to 167 of 699 entries, 10 total columns

Sesudah:

Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	Class
5	1	1	1	2	1	3	1	1	jinak
5	4	4	5	7	10	3	2	1	jinak
3	1	1	1	2	2	3	1	1	jinak
6	8	8	1	3	4	3	7	1	jinak
4	1	1	3	2	1	3	1	1	jinak
8	10	10	8	7	10	9	7	1	ganas
1	1	1	1	2	10	3	1	1	jinak
2	1	2	1	2	1	3	1	1	jinak
2	1	1	1	2	1	1	1	5	jinak
4	2	1	1	2	1	2	1	1	jinak
1	1	1	1	1	1	3	1	1	jinak
2	1	1	1	2	1	2	1	1	jinak
5	3	3	3	2	3	4	4	1	ganas
1	1	1	1	2	3	3	1	1	jinak
8	7	5	10	7	9	5	5	4	ganas
7	4	6	4	6	1	4	3	1	ganas
4	1	1	1	2	1	2	1	1	jinak
4	1	1	1	2	1	3	1	1	jinak
10	7	7	6	4	10	4	1	2	ganas
6	1	1	1	2	1	3	1	1	jinak
7	3	2	10	5	10	5	4	4	ganas
10	5	5	3	6	7	7	10	1	ganas
3	1	1	1	2	1	2	1	1	jinak
1	1	1	1	2	1	3	1	1	jinak
5	2	3	4	2	7	3	6	1	ganas
3	2	1	1	1	1	2	1	1	jinak
5	1	1	1	2	1	2	1	1	jinak

to 27 of 683 entries, 10 total columns

e. Mengambil baris yang akan digunakan, dalam hal ini kami mengambil 680 baris data teratas dari dataset untuk digunakan dalam analisis ini.

4. Modelling & Evaluation

Jenis model yang akan kami pilih yaitu klasifikasi (*Supervised Model*) untuk memprediksi apakah seorang pasien termasuk dalam klasifikasi kanker payudara dengan diagnosis sel menjadi jinak (*benign*) atau ganas (*malignant*).

Algoritma yang kami gunakan adalah *Random Forest* karena sesuai dengan dataset yang tersedia yaitu adanya banyak fitur untuk mengklasifikasikan antara sel jinak dan ganas. Algoritma ini juga memberikan informasi tentang fitur yang paling berpengaruh (*feature importance*) dari semua fitur yang terdapat dalam dataset ini. Selain itu, *Random Forest* cenderung memberikan hasil yang akurat saat jumlah data tidak terlalu besar, seperti dalam dataset *Breast Cancer Wisconsin*.

Modelling dilakukan pada dataset *Breast Cancer Wisconsin* dengan rasio 80:20 untuk data *training* dan data *testing*. Sehingga dari dataset yang berjumlah 680 data, jumlah data *training* sebanyak 543 data dan data *testing* yang digunakan untuk modeling sebanyak 137 data. Kemudian data *testing* tersebut akan dievaluasi dan menghasilkan jumlah prediksi yang kemudian dilakukan perhitungan untuk mengetahui nilai Akurasi, *Precision*, *Recall*, dan *F1-Score*.

Dibawah ini merupakan hasil pemodelan dengan menggunakan *Random Forest*:

```
Call:
 randomForest(formula = Class ~ Clump_thickness + Uniformity_of_cell_size +      Uniformity_of_cell_shape
+ Marginal_adhesion + Single_epithelial_cell_size +      Bare_nuclei + Bland_chromatin + Normal_nucleoli
+ Mitoses,      data = breast_cancer_train, ntree = 100, mtry = 3, importance = TRUE,      seed = 420)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 3

        OOB estimate of  error rate: 3.5%
Confusion matrix:
      ganas jinak class.error
ganas   181     7  0.03723404
jinak    12   343  0.03380282
```

Dari pemodelan tersebut didapat bahwa *Out-of-Bag* (OOB) Error Rate: 3.5%. Error ini cenderung termasuk angka yang kecil. Hal ini dapat disimpulkan bahwa hanya model salah memprediksi 3.5% dari data pelatihan.

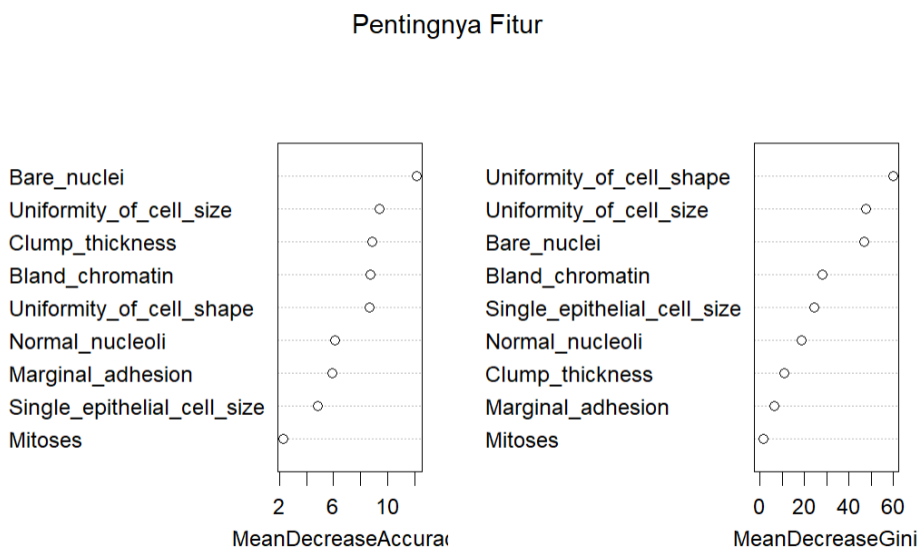
```
Call:
 randomForest(formula = Class ~ Clump_thickness + Uniformity_of_cell_size +      Uniformity_of_cell_shape +
Marginal_adhesion + Single_epithelial_cell_size +      Bare_nuclei + Bland_chromatin + Normal_nucleoli + Mitoses,
data = breast_cancer_train, ntree = 50, mtry = 3, importance = TRUE,      seed = 420)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 3

        OOB estimate of  error rate: 3.68%
Confusion matrix:
      ganas jinak class.error
ganas   180     8  0.04255319
jinak    12   343  0.03380282
```

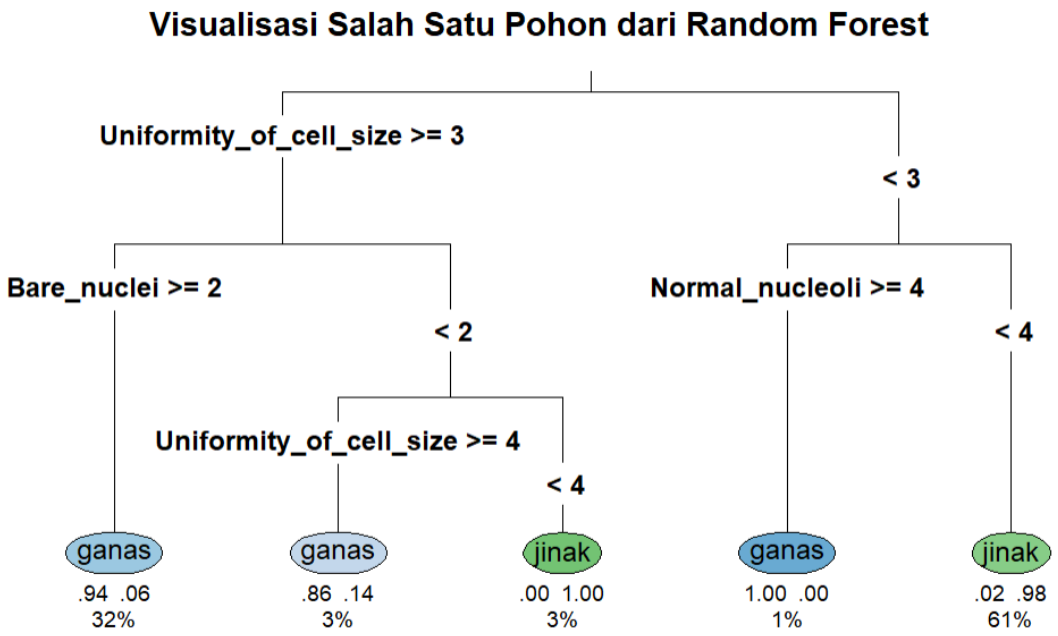
Jika Number of trees yang digunakan sebanyak lebih sedikit dari nilai 100, maka OOB yang dihasilkan sebesar 3.68%. Sehingga disimpulkan bahwa jumlah pohon 100 lebih baik dibandingkan jika jumlah pohon 50.

Tingkat kepentingan (*feature importance*) yang dihasilkan oleh pemodelan *Random Forest*. Variabel "Bare_nuclei" menjadi variabel yang paling berdampak terhadap performa model secara keseluruhan. Pembagian data (*split*) dilakukan pada setiap pohon untuk memisahkan data menjadi grup yang lebih homogen.

Fitur "Uniformity_of_cell_shape" berkontribusi paling besar dalam memprediksi apakah data termasuk “jinak” atau “ganas”. Dengan ini membantu mengidentifikasi fitur yang paling relevan dalam proses prediksi.



Salah satu visualisasi pohon dari 100 pohon dalam pemodelan *Random Forest* dengan 3 fitur:



Berikut adalah output yang menunjukkan hasil prediksi model pada data testing, dengan label klasifikasi "jinak" dan "ganas". Setiap angka di bagian atas mewakili indeks data dalam data testing, sedangkan label di bawahnya menunjukkan hasil prediksi model untuk setiap data.

```
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
jinak jinak jinak ganas ganas jinak ganas ganas ganas ganas jinak jinak jinak jinak jinak jinak
   17    18    19    20    21    22    23    24    25    26    27    28    29    30    31    32
ganas jinak jinak jinak jinak jinak ganas ganas jinak jinak jinak jinak jinak jinak ganas jinak
   33    34    35    36    37    38    39    40    41    42    43    44    45    46    47    48
jinak ganas jinak jinak ganas jinak jinak jinak ganas ganas ganas ganas ganas jinak jinak jinak
   49    50    51    52    53    54    55    56    57    58    59    60    61    62    63    64
ganas ganas jinak ganas ganas ganas ganas ganas ganas jinak ganas jinak ganas jinak jinak ganas
   65    66    67    68    69    70    71    72    73    74    75    76    77    78    79    80
jinak jinak jinak ganas ganas jinak jinak ganas jinak jinak jinak jinak ganas jinak jinak jinak
   81    82    83    84    85    86    87    88    89    90    91    92    93    94    95    96
ganas jinak ganas jinak jinak jinak jinak ganas jinak ganas ganas jinak jinak jinak jinak ganas
   97    98    99   100   101   102   103   104   105   106   107   108   109   110   111   112
ganas ganas jinak jinak jinak jinak ganas jinak jinak jinak jinak jinak jinak jinak jinak ganas
  113   114   115   116   117   118   119   120   121   122   123   124   125   126   127   128
jinak ganas jinak jinak ganas ganas jinak jinak ganas jinak jinak jinak jinak ganas jinak jinak
  129   130   131   132   133   134   135   136   137
jinak jinak jinak jinak jinak jinak jinak ganas jinak
Levels: ganas jinak
```

Berikut adalah prediksi probabilitas data dalam *data testing* yang menunjukkan apakah data tersebut cenderung termasuk Class “jinak” atau “ganas”:

	ganas	jinak
1	0.00	1.00
2	0.00	1.00
3	0.42	0.58
4	0.87	0.13
5	0.98	0.02
6	0.00	1.00
7	0.90	0.10
8	0.96	0.04
9	0.87	0.13
10	0.79	0.21
11	0.09	0.91
12	0.02	0.98
13	0.00	1.00
14	0.09	0.91
15	0.00	1.00
16	0.00	1.00
17	0.99	0.01
18	0.00	1.00
19	0.00	1.00
20	0.00	1.00
21	0.00	1.00
22	0.00	1.00
23	0.89	0.11
24	0.86	0.14
25	0.17	0.83
26	0.33	0.67
27	0.02	0.98
28	0.00	1.00
29	0.00	1.00
30	0.00	1.00
31	0.94	0.06
32	0.41	0.59
33	0.00	1.00
34	0.56	0.44
35	0.00	1.00
36	0.00	1.00
37	0.94	0.06
38	0.00	1.00
39	0.00	1.00
40	0.00	1.00
41	0.98	0.02
42	0.99	0.01
43	0.96	0.04
44	0.90	0.10
45	0.92	0.08
46	0.12	0.88
47	0.00	1.00
48	0.13	0.87
49	0.92	0.08
50	0.83	0.17
51	0.00	1.00
52	0.85	0.15
53	0.96	0.04
54	0.94	0.06
55	1.00	0.00
56	1.00	0.00
57	0.91	0.09
58	0.00	1.00
59	1.00	0.00
60	0.00	1.00
61	0.97	0.03
62	0.00	1.00
63	0.00	1.00
64	0.96	0.04
65	0.00	1.00
66	0.00	1.00
67	0.00	1.00
68	1.00	0.00
69	1.00	0.00
70	0.47	0.53
71	0.00	1.00
72	1.00	0.00
73	0.00	1.00
74	0.00	1.00
75	0.00	1.00
76	0.05	0.95
77	0.97	0.03
78	0.00	1.00
79	0.00	1.00
80	0.04	0.96
81	1.00	0.00
82	0.43	0.57
83	0.97	0.03
84	0.00	1.00
85	0.00	1.00
86	0.00	1.00
87	0.05	0.95
88	1.00	0.00
89	0.00	1.00
90	1.00	0.00
91	1.00	0.00
92	0.00	1.00
93	0.00	1.00
94	0.00	1.00
95	0.05	0.95
96	1.00	0.00
97	0.99	0.01
98	0.82	0.18
99	0.00	1.00
100	0.00	1.00
101	0.00	1.00
102	0.00	1.00
103	0.98	0.02
104	0.00	1.00
105	0.00	1.00
106	0.00	1.00
107	0.00	1.00
108	0.00	1.00
109	0.00	1.00
110	0.00	1.00
111	0.05	0.95
112	0.99	0.01
113	0.04	0.96
114	0.86	0.14
115	0.00	1.00
116	0.00	1.00
117	0.98	0.02
118	0.96	0.04
119	0.00	1.00
120	0.00	1.00
121	0.57	0.43
122	0.00	1.00
123	0.01	0.99
124	0.00	1.00
125	0.00	1.00
126	1.00	0.00
127	0.00	1.00
128	0.00	1.00
129	0.00	1.00
130	0.00	1.00
131	0.00	1.00
132	0.00	1.00
133	0.00	1.00
134	0.00	1.00
135	0.00	1.00
136	0.65	0.35
137	0.00	1.00

Hasil Permodelan:

Dataset Breast Cancer

Jumlah Jinak
444

Jumlah Ganas
236

Pilih Jenis Class
ganas

Show 10 entries

Search:

	Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	Class
1	8	10	10	8	7	10	9	7	1	ganas
2	5	3	3	3	2	3	4	4	1	ganas
3	8	7	5	10	7	9	5	5	4	ganas
4	7	4	6	4	6	1	4	3	1	ganas
5	10	7	7	6	4	10	4	1	2	ganas
6	7	3	2	10	5	10	5	4	4	ganas
7	10	5	5	3	6	7	7	10	1	ganas
8	5	2	3	4	2	7	3	6	1	ganas
9	10	7	7	3	8	5	7	4	3	ganas
10	10	10	10	8	6	1	8	9	1	ganas

Showing 1 to 10 of 236 entries

Previous 1 2 3 4 5 ... 24 Next

Data Splitting

Data Training
543

Data Testing
137

Metrik Evaluasi

Akurasi
0.985

Precision
0.979

Recall
0.979

F1-Score
0.979

Uji Coba Klasifikasi

Clump Thickness
7

Uniformity of Cell Size
1

Uniformity of Cell Shape
1

Marginal Adhesion
1

Single Epithelial Cell Size
7

Bare Nuclei
8

Bland Chromatin
1

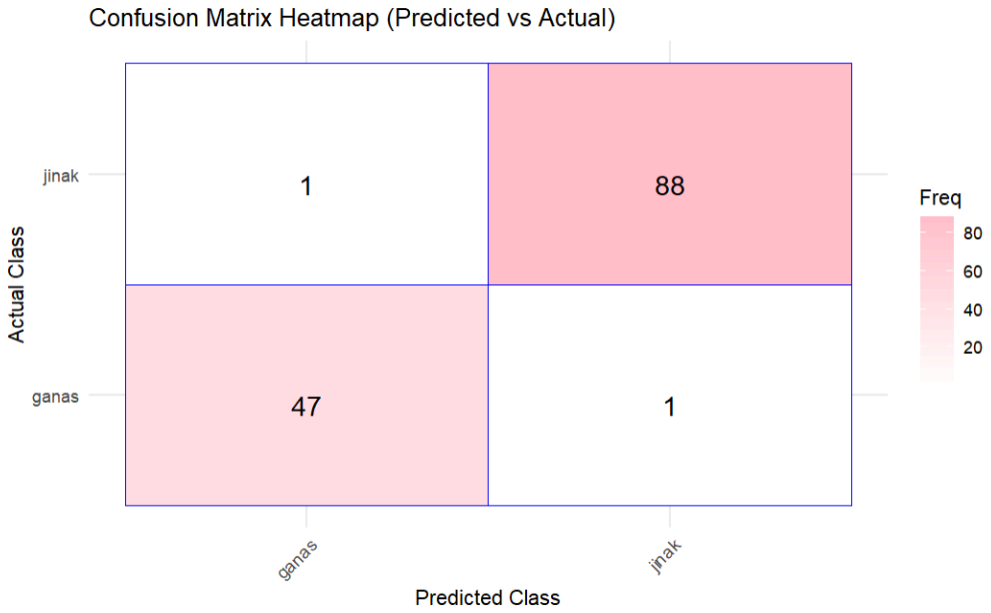
Normal Nucleoli
6

Mitoses
1

Klasifikasi Kelas

ganas

Berikut adalah *Confusion Matrix* yang dihasilkan:



Distribusi hasil evaluasi model (*Accuracy*, *Precision, Recall, F1-Score*):

```
Precision: 0.979
Recall: 0.979
F1-Score: 0.979
Accuracy: 0.985
```

Dalam kasus klasifikasi kanker payudara ini, *accuracy* digunakan untuk mengevaluasi performa secara keseluruhan. *Precision* digunakan untuk memastikan prediksi kanker ganas akurat. *Recall* memaksimalkan deteksi kasus ganas, dan *F1-score* menyeimbangkan *precision* serta *recall*, terutama pada dataset yang tidak seimbang, sehingga model dapat memberikan hasil yang tepat. Dari evaluasi tersebut dapat diketahui bahwa model *Random Forest* yang kami gunakan sudah mampu mengklasifikasikan kanker payudara menjadi kelas "ganas" atau "jinak" dengan sangat baik.

5. **Kesimpulan**

Melalui analisis ini, kami bertujuan untuk meningkatkan efektivitas dan akurasi dalam deteksi dini kanker payudara. Dengan memahami karakteristik seluler dari kanker payudara, tujuan kami adalah membangun model klasifikasi yang dapat membedakan antara sel jinak (benign) dan sel ganas (malignant) secara cepat dan akurat. Hal ini diharapkan dapat mempercepat proses diagnosis, mengurangi potensi kesalahan akibat deteksi manual, serta meningkatkan peluang kesembuhan pasien dan menurunkan angka kematian.

Hasil analisis menunjukkan bahwa algoritma *Random Forest* sebagai model klasifikasi memiliki tingkat akurasi yang tinggi. Model ini memiliki sensitivitas yang baik untuk mendeteksi sel ganas dan sel jinak. Selain itu, model ini juga mampu mengidentifikasi fitur-fitur penting seperti *Uniformity_of_cell_size* dan *Bare_nuclei*, yang berperan signifikan dalam proses prediksi kanker payudara. Kebermanfaatannya bagi institusi medis sangat

signifikan, terutama dalam mempercepat proses diagnosis, mengurangi beban kerja dokter, dan meningkatkan efisiensi operasional rumah sakit atau laboratorium diagnostik.

Namun, meskipun hasil analisis cukup baik, terdapat beberapa tantangan yang perlu diperhatikan. Salah satunya adalah kemungkinan kesulitan dalam membedakan antara beberapa jenis kanker payudara yang memiliki karakteristik seluler yang sangat mirip, yang dapat menyebabkan penurunan akurasi pada kasus-kasus tertentu. Selain itu, hasil analisis sangat bergantung pada kualitas dataset yang digunakan dalam pelatihan. Jika dataset tersebut kurang representatif atau mengandung data yang tidak seimbang, hal ini dapat mempengaruhi hasil prediksi model.

Selain itu, implementasi model ini menghadirkan beberapa tantangan teknis, salah satunya adalah meningkatnya kompleksitas komputasi seiring dengan bertambahnya jumlah pohon yang digunakan dalam algoritma. Sebagai contoh, penggunaan 100 pohon menghasilkan akurasi yang lebih tinggi dibandingkan 50 pohon, tetapi juga membutuhkan waktu pemrosesan yang lebih lama dan dapat menjadi kendala pada perangkat dengan kapasitas komputasi terbatas. Model juga dapat rentan terhadap *overfitting* jika tidak diatur dengan baik dan mungkin sulit diinterpretasikan oleh pengguna non-teknis.

Untuk mendukung implementasi model Random Forest dalam deteksi dini kanker payudara, kami merekomendasikan beberapa strategi. Pertama, integrasi model ini ke dalam sistem diagnostik berbasis perangkat lunak sangat disarankan untuk memperoleh hasil yang cepat dan akurat. Kedua, pelatihan bagi tenaga medis sangat diperlukan agar mereka dapat memahami dan memanfaatkan hasil keluaran model secara efektif dalam pengambilan keputusan klinis. Ketiga, dataset yang digunakan untuk melatih model harus diperbarui secara berkala untuk memastikan model tetap relevan dengan pola data terkini. Berdasarkan hasil analisis, penggunaan 100 pohon dinilai optimal untuk mencapai keseimbangan antara akurasi yang tinggi dan waktu komputasi yang wajar.

Dengan implementasi yang baik, sistem diagnostik berbasis Random Forest memiliki potensi untuk mempercepat proses diagnosis dini, yang pada akhirnya dapat meningkatkan peluang keselamatan pasien dan kualitas hidup mereka. Keberhasilan implementasi model ini dapat meningkatkan efisiensi dan akurasi diagnosis, serta mendukung peningkatan kualitas pelayanan kesehatan secara keseluruhan. Namun, perhatian terhadap pemeliharaan kualitas data dan evaluasi berkala terhadap kinerja model juga sangat penting untuk memastikan keberlanjutan dan relevansi model dalam menghadapi variasi data yang lebih luas.

6. **Daftar Pustaka**

[1] Pahlevi, O., Amrin, & Handrianto, Y. (2023). Implementasi Algoritma Klasifikasi Random Forest untuk Penilaian Kelayakan Kredit. *Jurnal Infortech*, 5(1), 71. http://ejournal.bsi.ac.id/ejurnal/index.php/infortech

*definisi bisnis yang dimaksud di bagian ini bukan wajib bidang bisnis, namun berarti berbagai bidang (sesuai yang telah dijelaskan/ dicoba implementasinya pada mini project)