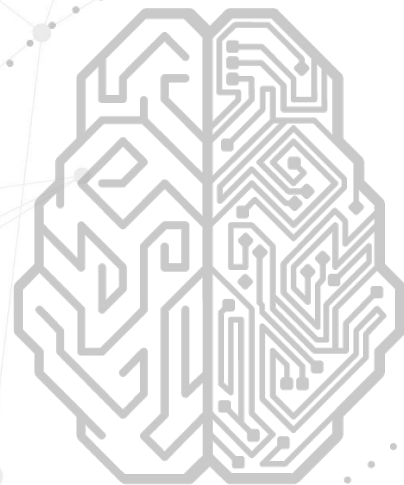


# MERCK

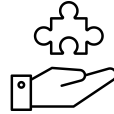
## Breast Cancer Detection



# AGENDA



**Merck  
Case-Study**



**Data  
Understanding**



**Analysis Approach**



**Predictive Model  
Development**



**Recommendations**



# Merck: Case-Study

# Merck: Breast Cancer Detection

*Our past and our future.*

## Business Perspective



Decrease the turnaround time in cancer detection process within patient diagnosis



Reduce the cost of generating the models by leveraging just the oncology data & Increase Insights for cohort treatment

## Business Challenges



**Traditional models** are lacking in accuracy and do affect the overall patient diagnosis



The cost of gathering the patient data from EHR systems like EPIC and Cerner is very high and requires a lot of compliance clearance

## Technical Challenges



**Gail model** used simple statistical architectures and the additional inputs were derived from costly and/or invasive procedures



Lack of centralized analytical platform that can generate the required data for breast cancer detection and provide results instantly

*The Client*



**Merck** is a pharmaceutical and life sciences company and one of the largest pharmaceutical companies in the world. Merck & Co. manufactures a variety of well-known pharmaceutical drugs, vaccines, and animal health products. It made the first smallpox vaccine for commercial use in the U.S. It is also the maker of the painkiller Vioxx and the HPV vaccine Gardasil.

# Merck: All about the company

*Our past and our future.*

## Service Line



Oncology



Vaccines



Infectious diseases



COVID-19



Cardio-metabolic disorders



Discovery & development

## Geo Reach

### North America

# Employees: 12829  
#Revenue: 4,214 million

### Europe

# Employees: 26,715  
#Revenue: 4,735 million

### APAC

# Employees: 12,728  
#Revenue: 5,599 million

### Middle East

# Employees: 1,366  
#Revenue: 591 million

### Latin America

# Employees: 3,433  
#Revenue: 1,012 million

Total Employees: **57K**

Sales: **16K Million**

# Merck: SWOT Analysis

*Strengths and Weaknesses*



## MERCK



### Strengths

- Skilled Staff
- Wide spread base of Buyers
- Resource and Labor Availability
- Setting the Benchmark



### Weakness

- Non-availability of modern technology
- Low profitability and low Margins



### Opportunities

- Government Support
- Growing Space
- Financial Support
- Patents
- Global Market



### Threats

- Currency Devaluation
- Inflation
- New Competitors
- Insufficient Power Supply

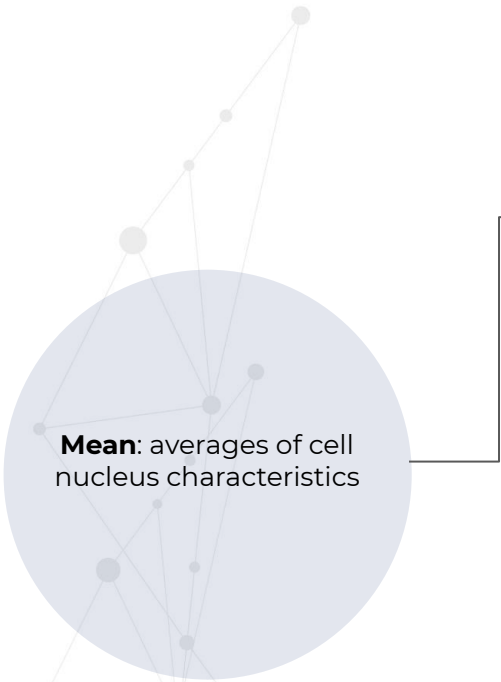


# Data Understanding

# Data Set Description

*Mean values of the cell area*

Variable Name	Description
ID	Id number
Diagnosis	The diagnosis of breast tissues
<b>MEAN</b>	
Radius_mean	Mean of distances from center to points on the perimeter
Texture_mean	Mean Standard deviation of gray-scale values
Perimeter_mean	Mean distance around the cell nucleus
Area_mean	Mean area/size of the cell nucleus
Smoothness_mean	Mean Local variation in radius length
Compactness_mean	Average of calculation: $\text{Perimeter}^2 / \text{area} - 1.0$
Concavity_mean	Mean severity of concave portions of the contour
Concave.points_mean	Mean number of concave portions of the contour
Symmetry_mean	Mean proportionality of cell nucleus
Fractal_dimension_mean	Mean of calculation: coastline approximation - 1



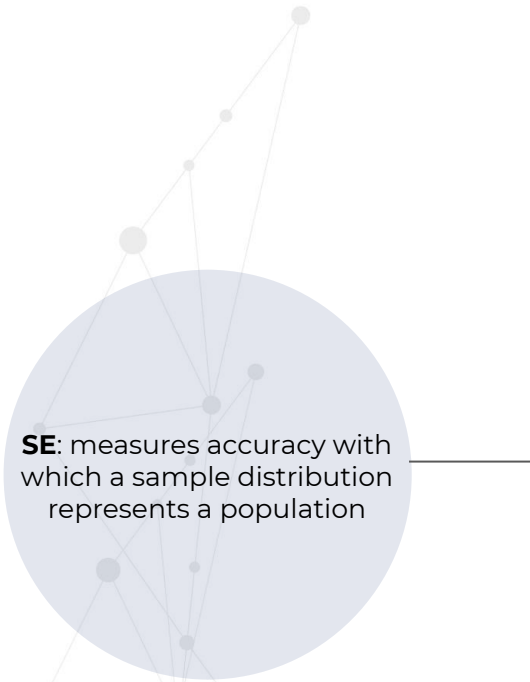
**Mean:** averages of cell nucleus characteristics



# Data Set Description

*Standard Error values of the cell area*

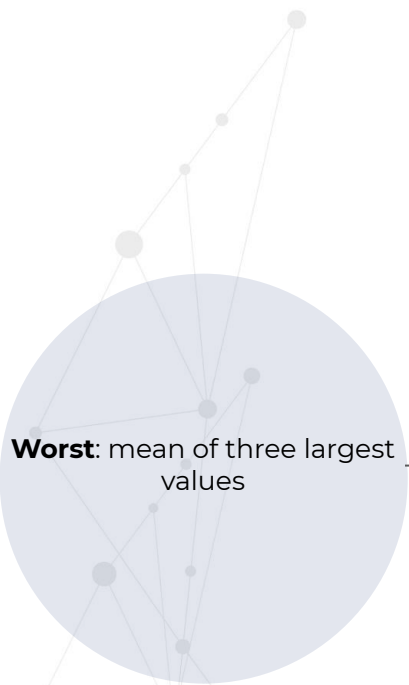
Variable Name	Description
<b>STANDARD ERROR</b>	
Radius_se	Standard error of distances from center to points on the perimeter
Texture_se	Standard error of Standard deviation of gray-scale values
Perimeter_se	Standard error distance around the cell nucleus
Area_se	Standard error area/size of the cell nucleus
Smoothness_se	Standard error Local variation in radius length
Compactness_se	Standard error of calculation: $\text{Perimeter}^2 / \text{area} - 1.0$
Concavity_se	Standard error severity of concave portions of the contour
Concave.points_se	Standard error number of concave portions of the contour
Symmetry_se	Standard error proportionality of cell nucleus
Fractal_dimension_se	Standard error of calculation: coastline approximation - 1



**SE:** measures accuracy with which a sample distribution represents a population

# Data Set Description

*Worst values of the cell area*

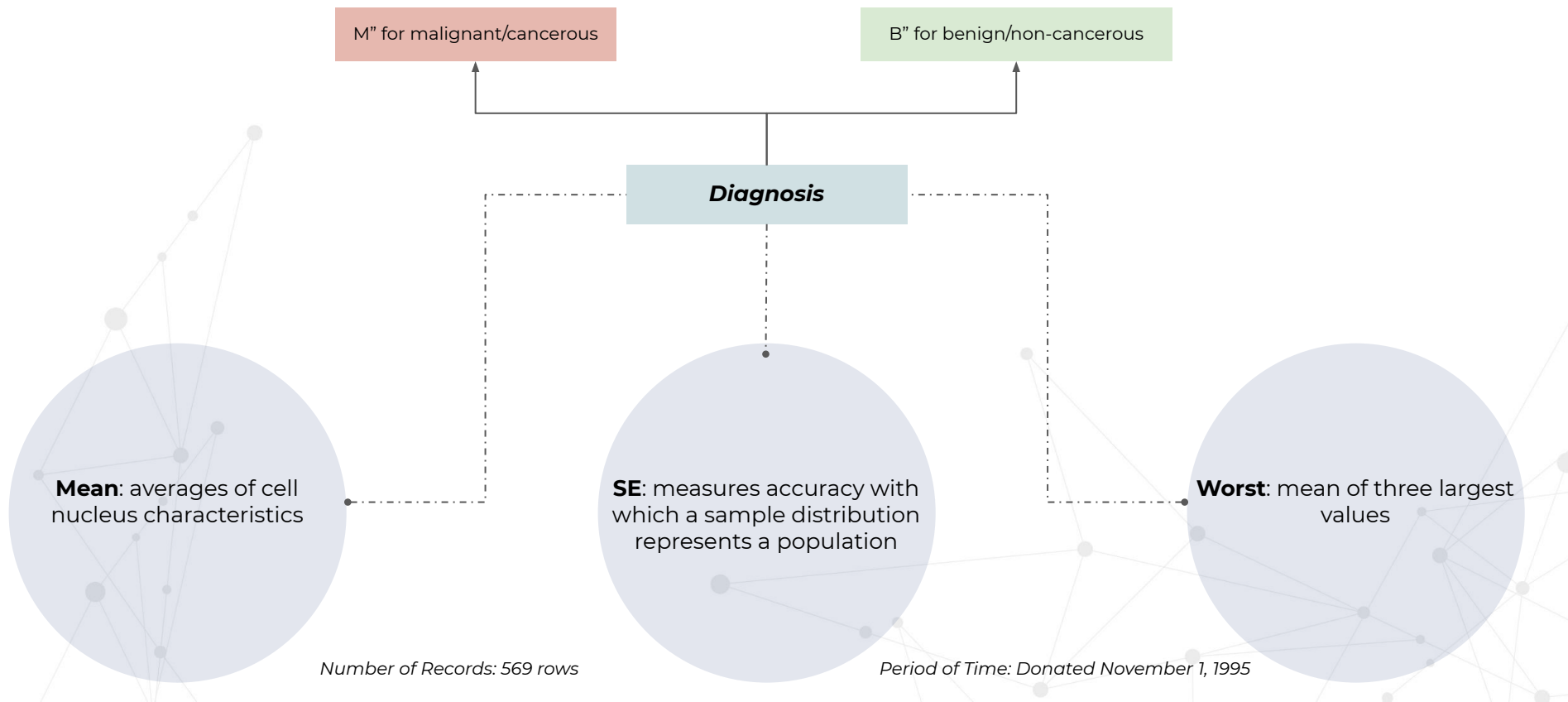


Variable Name	Description
<b>WORST</b>	
Radius_worst	Worst of distances from center to points on the perimeter
Texture_worst	Worst Standard deviation of gray-scale values
Perimeter_worst	Worst distance around the cell nucleus
Area_worst	Worst area/size of the cell nucleus
Smoothness_worst	Worst Local variation in radius length
Compactness_worst	Worst of calculation: $\text{Perimeter}^2 / \text{area} - 1.0$
Concavity_worst	Worst severity of concave portions of the contour
Concave.points_worst	Worst number of concave portions of the contour
Symmetry_worst	Worst proportionality of cell nucleus
Fractal_dimension_worst	Worst of calculation: coastline approximation - 1

# Target Variable

Predicting the class of the oncology data

[Data Source](#)





# Analysis Approach

## Solution Approach



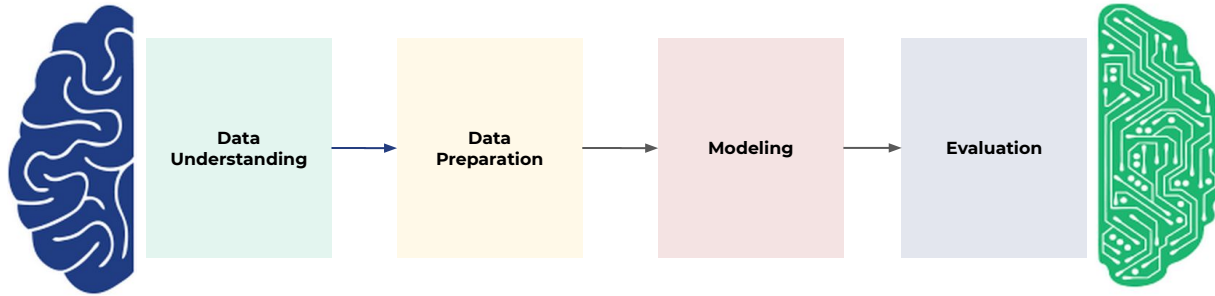
Explore the dataset and summarize the relations between the variables. Create descriptive statistics, box plots, scatter plots, and a correlation matrix for all variables.



Use **logistic regression** to predict the diagnosis of breast tissue based on all or selected cell nucleus characteristics.

Build a **classification model** (single and pruned decision trees, random forest) and extract the rules to be used for the prediction of breast tissue diagnosis.

## Solution Design

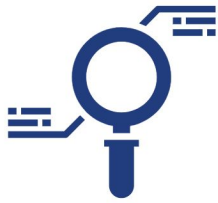


## Solution Aim : Recommendation

- Build a highly efficient model with an accuracy better than the traditional models used in Breast Cancer Risk Prediction Tool (BCRAT)
- Develop data mining models that used highly accessible personal health data to predict breast cancer risk
- Implement non-invasive and cost-effective risk stratification tools to increase early breast cancer detection and prevention

## Deliverables

- Research and gather the data set
- Clean the data and build a predictive model
- Generate final report of the analysis and provide recommendations



# Data Explorations

# Summary Statistics

Summarizing each variable

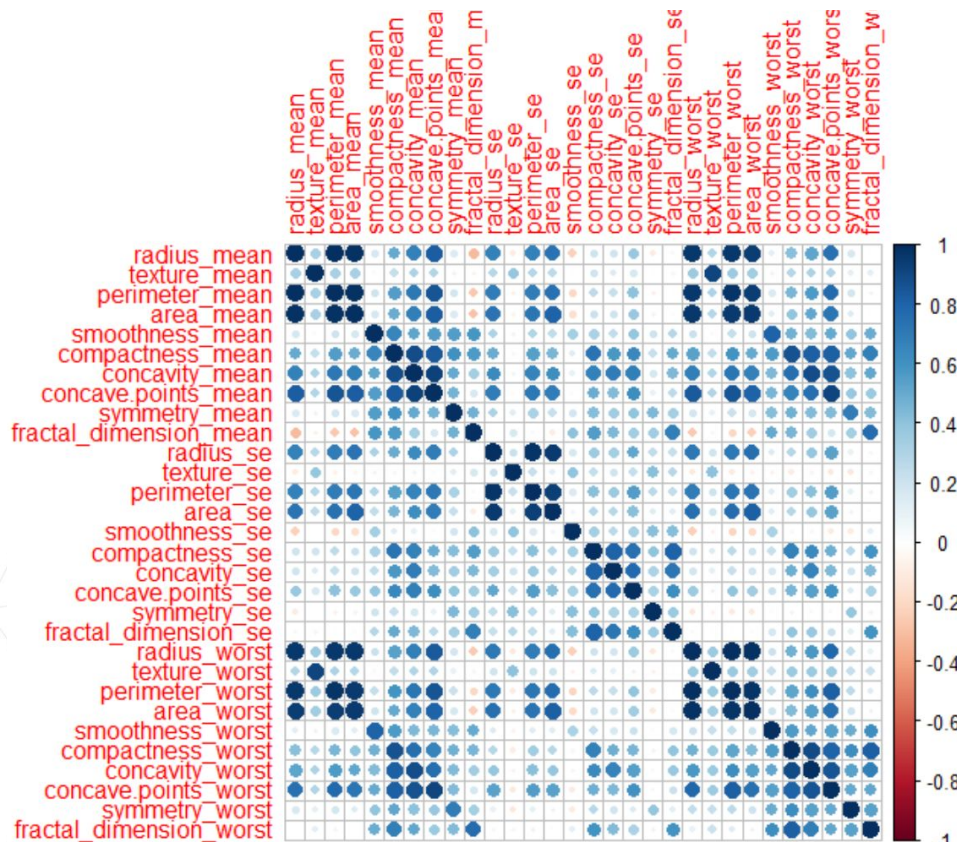
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
diagnosis	569						
... B	357	62.7%					
... M	212	37.3%					
radius_mean	569	14.127	3.524	6.981	11.7	15.78	28.11
texture_mean	569	19.29	4.301	9.71	16.17	21.8	39.28
perimeter_mean	569	91.969	24.299	43.79	75.17	104.1	188.5
area_mean	569	654.889	351.914	143.5	420.3	782.7	2501
smoothness_mean	569	0.096	0.014	0.053	0.086	0.105	0.163
compactness_mean	569	0.104	0.053	0.019	0.065	0.13	0.345
concavity_mean	569	0.089	0.08	0	0.03	0.131	0.427
concave.points_mean	569	0.049	0.039	0	0.02	0.074	0.201
symmetry_mean	569	0.181	0.027	0.106	0.162	0.196	0.304
fractal_dimension_mean	569	0.063	0.007	0.05	0.058	0.066	0.097
radius_se	569	0.405	0.277	0.112	0.232	0.479	2.873
texture_se	569	1.217	0.552	0.36	0.834	1.474	4.885
perimeter_se	569	2.866	2.022	0.757	1.606	3.357	21.98
area_se	569	40.337	45.491	6.802	17.85	45.19	542.2
smoothness_se	569	0.007	0.003	0.002	0.005	0.008	0.031
compactness_se	569	0.025	0.018	0.002	0.013	0.032	0.135
concavity_se	569	0.032	0.03	0	0.015	0.042	0.396
concave.points_se	569	0.012	0.006	0	0.008	0.015	0.053
symmetry_se	569	0.021	0.008	0.008	0.015	0.023	0.079
fractal_dimension_se	569	0.004	0.003	0.001	0.002	0.005	0.03
radius_worst	569	16.269	4.833	7.93	13.01	18.79	36.04
texture_worst	569	25.677	6.146	12.02	21.08	29.72	49.54
perimeter_worst	569	107.261	33.603	50.41	84.11	125.4	251.2
area_worst	569	880.583	569.357	185.2	515.3	1084	4254
smoothness_worst	569	0.132	0.023	0.071	0.117	0.146	0.223
compactness_worst	569	0.254	0.157	0.027	0.147	0.339	1.058
concavity_worst	569	0.272	0.209	0	0.114	0.383	1.252
concave.points_worst	569	0.115	0.066	0	0.065	0.161	0.291
symmetry_worst	569	0.29	0.062	0.156	0.25	0.318	0.664
fractal_dimension_worst	569	0.084	0.018	0.055	0.071	0.092	0.208

## Key insights

- The overall data don't require any normalization cause there is not much difference in the scale of each variable
- There is a need to convert the diagnosis variable into factors for classification models and into numeric for regression model
- As of now, there are 31 variables and not all variables are required for model development

# Correlation Matrix

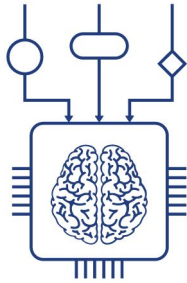
Feature Selection for Model development



## Key insights

- Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features
- Radius\_mean, Perimeter\_mean, Area\_mean, Radius\_worst, Perimeter\_worst & Area\_worst are highly correlated. Radius\_mean can be considered for the model development and rest of the variables can be discarded
- Radius\_se, perimeter\_se & area\_se are highly correlated and only one column instead of three can be considered
- As of now, there are 24 variables for developing predictive models

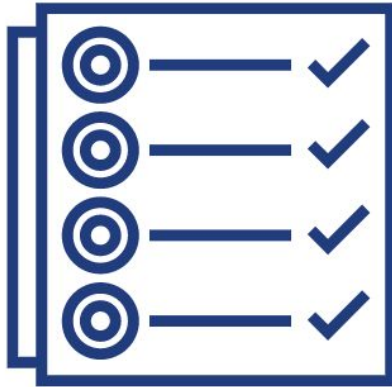




# Predictive Model Development

# Model Goals & Objectives

*Setting up the goals for model development*



**Target Variable:** Diagnosis (characterized by “M” for malignant/cancerous and “B” for benign/non-cancerous)

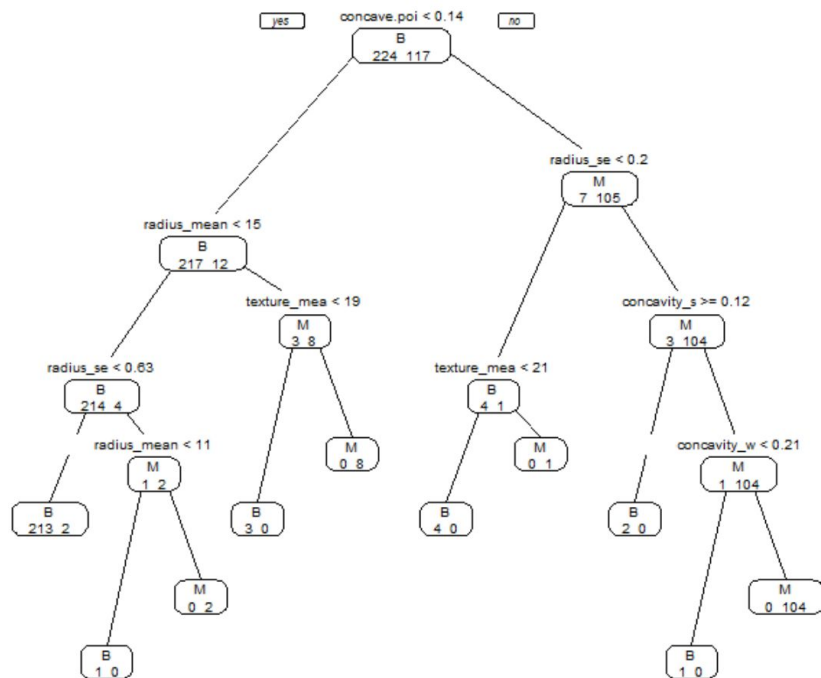
Use logistic regression to predict the diagnosis of breast tissue based on all or selected cell nucleus characteristics.

Build a classification model (single and pruned decision trees, random forest) and extract the rules to be used for prediction of breast tissue diagnosis.

Determine the cell nucleus characteristics that are highly associated with malignant cells.

# Regular Decision Tree

Predicting malignant and benign



## Confusion Matrix and Statistics

Prediction \ Reference	Reference	
	B	M
B	119	12
M	14	83

Accuracy : 0.886

95% CI : (0.8374, 0.9241)

No Information Rate : 0.5833

P-value [Acc > NIR] : <0.0000000000000002

Kappa : 0.7661

Mcnemar's Test P-value : 0.8445

Sensitivity : 0.8947

Specificity : 0.8737

Pos Pred Value : 0.9084

Neg Pred Value : 0.8557

Prevalence : 0.5833

Detection Rate : 0.5219

Detection Prevalence : 0.5746

Balanced Accuracy : 0.8842

'Positive' Class : B

**Accuracy: 88.6%**

# Deeper Decision Tree

Predicting malignant and benign

## Confusion Matrix and Statistics

Prediction \ Reference	B	M
	B	M
B	119	11
M	14	84

Accuracy : 0.8904

95% CI : (0.8424, 0.9278)

No Information Rate : 0.5833

P-Value [Acc > NIR] : <0.0000000000000002

Kappa : 0.7754

McNemar's Test P-Value : 0.6892

Sensitivity : 0.8947

Specificity : 0.8842

Pos Pred value : 0.9154

Neg Pred value : 0.8571

Prevalence : 0.5833

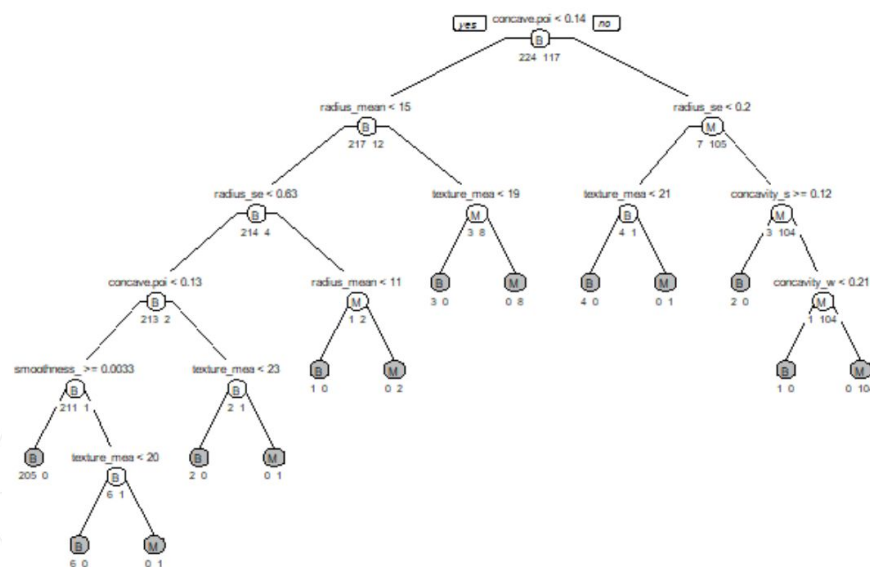
Detection Rate : 0.5219

Detection Prevalence : 0.5702

Balanced Accuracy : 0.8895

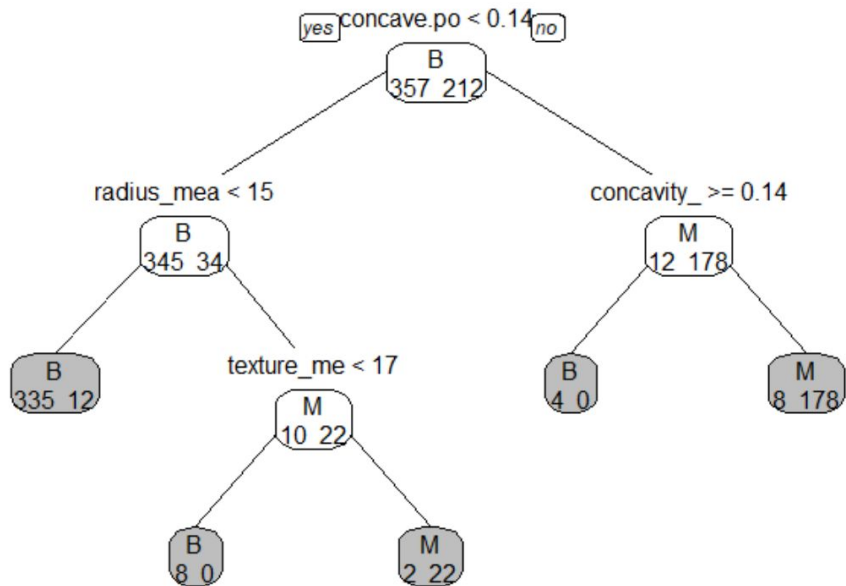
'Positive' class : B

**Accuracy: 89%**



# Pruned Decision Tree

*Predicting malignant and benign*



## Confusion Matrix and Statistics

Prediction \ Reference	Reference	
	B	M
B	127	6
M	6	89

Accuracy : 0.9474

95% CI : (0.9099, 0.9725)

No Information Rate : 0.5833

P-Value [Acc > NIR] : <0.0000000000000002

Kappa : 0.8917

Mcnemar's Test P-Value : 1

Sensitivity : 0.9549

Specificity : 0.9368

Pos Pred Value : 0.9549

Neg Pred Value : 0.9368

Prevalence : 0.5833

Detection Rate : 0.5570

Detection Prevalence : 0.5833

Balanced Accuracy : 0.9459

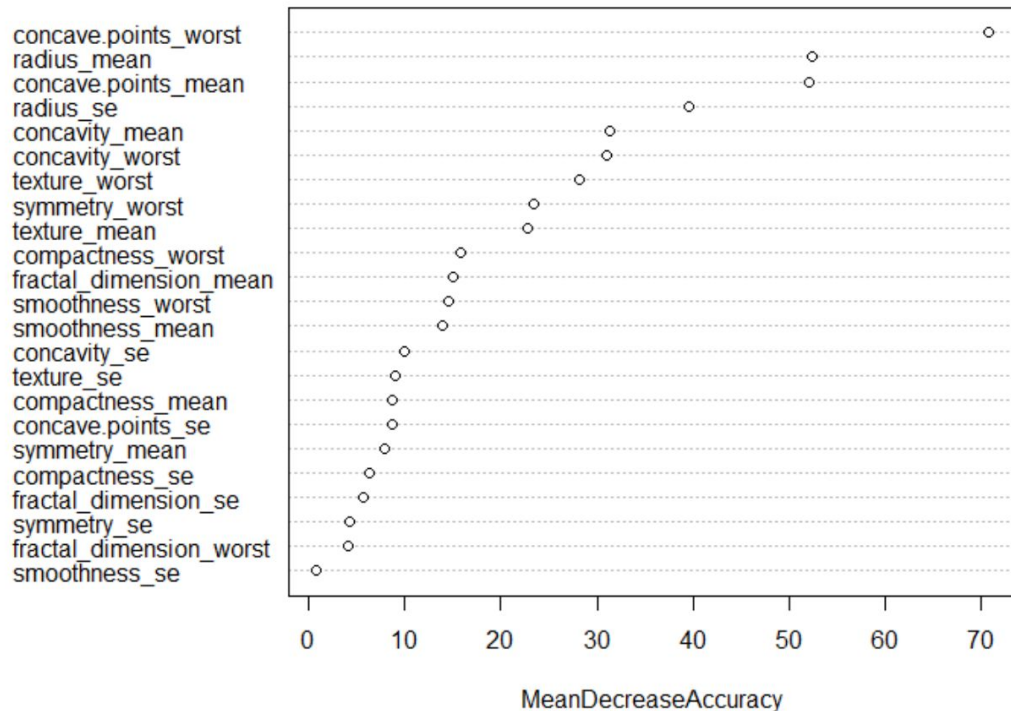
'Positive' Class : B

**Accuracy: 94.74%**

# Random Forest

Predicting malignant and benign

rf



## Confusion Matrix and Statistics

Prediction \ Reference	Reference	
	B	M
B	126	14
M	7	81

Accuracy : 0.9079  
95% CI : (0.8627, 0.9421)  
No Information Rate : 0.5833  
P-Value [Acc > NIR] : <0.0000000000000002

Kappa : 0.8085

Mcnemar's Test P-Value : 0.1904

Sensitivity : 0.9474  
Specificity : 0.8526  
Pos Pred Value : 0.9000  
Neg Pred Value : 0.9205  
Prevalence : 0.5833  
Detection Rate : 0.5526  
Detection Prevalence : 0.6140  
Balanced Accuracy : 0.9000

'Positive' Class : B

**Accuracy: 90.79%**

# Logistics Regression

*Predicting malignant and benign*

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1153.535	1239490.132	-0.001	0.999
radius_mean	15.385	16638.571	0.001	0.999
texture_mean	11.984	49514.326	0.000	1.000
smoothness_mean	6038.605	5031697.659	0.001	0.999
compactness_mean	-3267.210	2613817.993	-0.001	0.999
concavity_mean	2658.886	2062549.697	0.001	0.999
concave.points_mean	-711.589	3295629.447	0.000	1.000
symmetry_mean	-433.331	2423459.606	0.000	1.000
fractal_dimension_mean	4323.195	13897561.644	0.000	1.000
radius_se	559.541	160885.066	0.003	0.997
texture_se	-75.326	248278.675	0.000	1.000
smoothness_se	3261.872	20549085.692	0.000	1.000
compactness_se	365.675	24791849.531	0.000	1.000
concavity_se	-246.804	10518600.979	0.000	1.000
concave.points_se	585.289	47706868.456	0.000	1.000
symmetry_se	-5840.094	9237498.761	-0.001	0.999
fractal_dimension_se	-33215.225	196044085.955	0.000	1.000
texture_worst	1.427	52658.673	0.000	1.000
smoothness_worst	-2221.345	2372655.538	-0.001	0.999
compactness_worst	268.619	2101338.163	0.000	1.000
concavity_worst	-223.573	1292850.037	0.000	1.000
concave.points_worst	674.118	6145089.669	0.000	1.000
symmetry_worst	938.637	437805.246	0.002	0.998
fractal_dimension_worst	716.289	17271722.653	0.000	1.000

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 438.57771753363 on 340 degrees of freedom  
Residual deviance: 0.00000011487 on 317 degrees of freedom  
AIC: 48

Number of Fisher Scoring iterations: 25

Confusion Matrix and Statistics

	FALSE	TRUE
FALSE	126	8
TRUE	7	87

Accuracy : 0.9342

95% CI : (0.8938, 0.9627)

No Information Rate : 0.5833

P-value [Acc > NIR] : <0.00000000000000002

Kappa : 0.8645

Mcnemar's Test P-value : 1

Sensitivity : 0.9474

Specificity : 0.9158

Pos Pred Value : 0.9403

Neg Pred Value : 0.9255

Prevalence : 0.5833

Detection Rate : 0.5526

Detection Prevalence : 0.5877

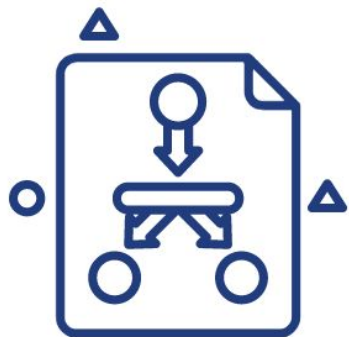
Balanced Accuracy : 0.9316

'Positive' class : FALSE

**Accuracy: 93.42%**

# Model Recommendation

*Predicting malignant and benign*



## *Pruned Decision Tree*

- The overall accuracy is 94.74%
- It's far better than a blind guess of 58%
- The number of independent variables required for predicting accurately is less, just 4, out of 24 variables
- It's highly efficient and quick in predicting the outcome with less variables





# Recommendations

# Business Recommendation

*Tangible solution for overall business development*

1

## A Diagnosis Tool in clinics

- Integrate into the clinical work
- Help the pathologist to detect cancer
- Improve the consistency of diagnosis
- decrease human error

2

## A Human-Machine Match on TV

- To educate women and community and attract donator, Merck could perform a human-machine Match on TV.
- The match will provide numerous digitized image of FNA of breast mass to compare the error rates between pathologist and the model of Pruned decision tree
- Through this kind of interesting way to appeal more and more people to pay attention to breast cancer

3

## The Economic Costs for Industry

- Merck could compare the economic costs of using the model of pruned tree and the current screening processes.
- If the model of Random Forest can save patient and clinics cost, even industry cost, there would be positive influence for medical industry.

The background of the slide features a complex, abstract network of thin, light gray lines connecting various-sized gray dots. These dots are scattered across the entire frame, creating a sense of interconnectedness and a modern, technological aesthetic. The lines vary in length and orientation, forming a web-like structure that frames the central text.

Thank You