**A Strategic Report on the Adoption of Visualization and Classification Techniques in Healthcare sector**

# Motivation

The healthcare sector, particularly within the National Health Service (NHS) faces significant challenges in managing vast and complex data, accurately diagnosing conditions, efficiently allocating resources and developing proactive healthcare strategies (Elgendi et al., 2020). The difficulty in swiftly identifying trends, patterns and correlations from complex data can hinder decision-making thereby leading to delayed interventions and ultimately, compromised patient care. These challenges highlight the pressing need for innovative solutions that can transform data into actionable insights and enhance diagnostic accuracy (Mohanty, 2021).

The integration of visualization and classification techniques offers a transformative solution to the NHS's challenges, streamlining data analysis with tools like Power BI for quick, informed decision-making and improved resource allocation. Visualization fosters clear insights into health trends, while classification, through methods such as machine learning algorithms, enhances diagnostic precision, enabling timely, tailored interventions (Elgendi et al., 2020). Classification techniques like KNN and Random Forest provide the framework for distinguishing between complex disease patterns with high accuracy. Adopting these technologies drives the NHS towards a more effective, patient-focused model, elevating care quality and operational efficiency (Ozsahin and Ozsahin, 2021). This strategic move positions the NHS at the forefront of healthcare innovation, leveraging advanced methodologies to benefit patients and practitioners alike (Cristea et al.).

Literature Review

The dynamic intersection of visualization and classification within healthcare is a critical area of study particularly in the context of big data challenges (Kulkarni et al., 2020). This integration is pivotal for navigating the complexities inherent in healthcare data by offering nuanced insights that drive actionable interventions. Literature that explores into this intersection is rich with examples of how these combined techniques facilitate a deeper understanding and more effective management of health-related issues.

A novel approach highlighted by (Alnowaiser, 2024) demonstrates how visualization techniques, when used in tandem with classification algorithms can significantly enhance the detection and management of diabetes. Their study illustrates a model where data visualization aids in the preliminary analysis of patient data, setting the stage for more precise classification through machine learning. This synergy enables healthcare providers to rapidly identify at-risk patients and tailor interventions more effectively.

Similarly, the work of (Kristina et al., 2020) explores the use of visualization tools to interpret the outputs of classification models in lymphoma cancer research. By presenting classification results in a visually accessible manner, researchers and clinicians can better understand patterns of cancer progression and treatment efficacy. This approach not only improves clinical decision-making but also fosters a clearer communication channel between data scientists and medical practitioners.

The integration of these techniques also addresses the pressing need for efficient data handling in healthcare, as noted by (Chakraborty et al., 2021). Their analysis emphasizes the role of classification in managing the volume and velocity of healthcare data, while visualization addresses the complexity and variety, making the data actionable for decision-makers. This dual approach is essential for leveraging big data towards predictive analytics, personalized medicine, and operational efficiency in healthcare systems.

Furthermore, the work by (Jackins et al., 2021), underscores the strategic value of merging visualization and classification. They argue that this integrated approach not only facilitates immediate identification of health trends and anomalies but also empowers healthcare organizations to anticipate future challenges, thereby enhancing patient care and operational effectiveness.

By focusing on studies that explore their combined application, it becomes evident that this integration is key to unlocking the full value of big data in healthcare. Visualization and classification together represent a powerful toolkit for addressing the multifaceted challenges of modern healthcare through improved predictive analytics, personalized patient care and enhanced decision-making processes (Alnowaiser, 2024).

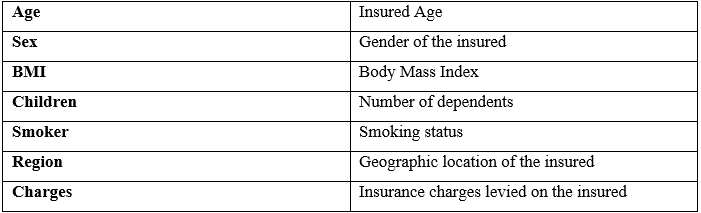
# Case Examples

## Optimizing Health Insurance Analytics through Power BI Visualizations

### Dataset Overview

In the rapidly evolving health insurance industry, understanding the dynamics of insurance premiums and the factors influencing them is crucial for insurers, policyholders and healthcare policymakers (Chiang et al., 2017). The comprehensive (Insurance) dataset from Kaggle offers an invaluable opportunity to explore into the relationships between individual attributes and insurance charges. With detailed records spanning various demographics, habits, and charges, this dataset becomes a cornerstone for analytical exploration and visualization.

This dataset lays the groundwork for a Power BI-driven analysis, aiming to uncover patterns and correlations within the health insurance landscape. It comprises the below variables:



The objective is to utilize Power BI's dynamic visualization tools to transform the insurance dataset into an interactive and enlightening set of dashboards and reports which helps in decision-making.

### Data Processing and Analysis:

Initiating the exploration with comprehensive data preparation in Power BI involved:

* Adjusting discrepancies by addressing missing values and ensuring data integrity.
* Enhancing the dataset by deriving new insightful metrics, like risk categories based on Age and smoker status.
* Carefully structuring the data to optimize its utility within Power BI, ensuring seamless analysis and visualization.

### Techniques Employed:

The crafting of visualizations in Power BI (as shown in Figure 1) focused on:

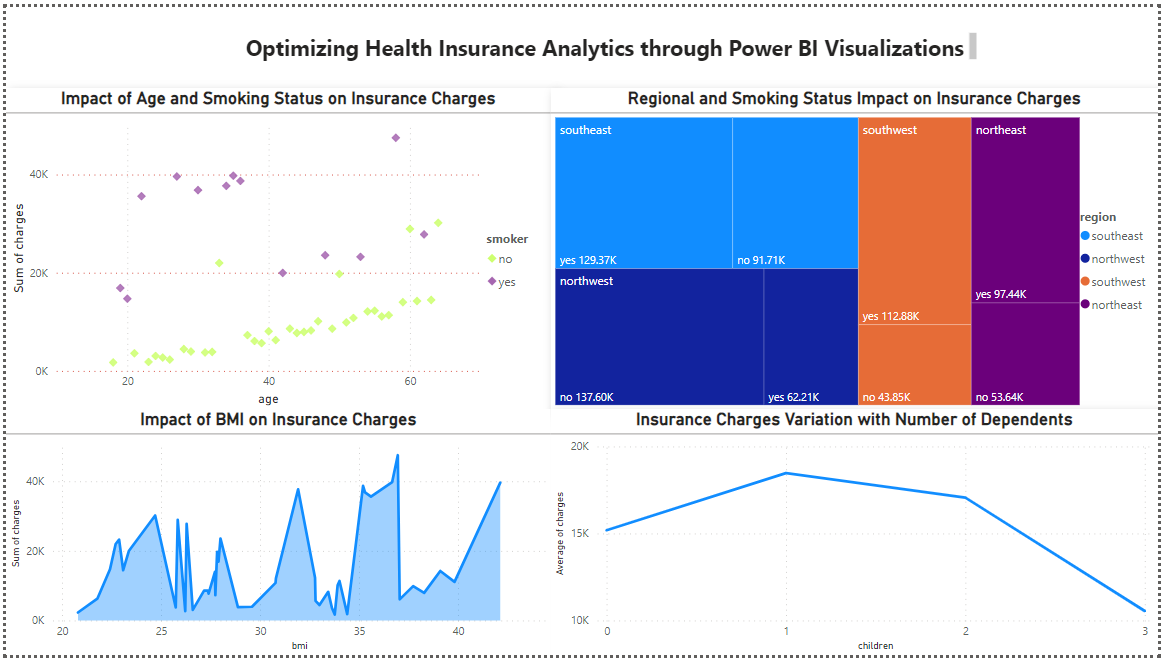
**Age and Smoking Status Correlation**: Utilized a scatter plot to elucidate the interaction between age and smoking status on insurance charges. This visual differentiation showcased how each group contributes to overall insurance costs with distinct markers for smokers and non-smokers.

**Regional and Smoking Status Impact Analysis**: Implementing stacked column charts to delineate the effects of region and smoking status on insurance charges. The segmentation within each column reveals the proportion of charges attributable to smokers versus non-smokers across different regions, providing a comparative analysis of geographic and behavioral influences on insurance expenses.

**BMI Influence Examination**: Applying an area chart to trace the association between BMI values and insurance charges. The visual representation through this area chart allows for a quick assessment of how varying BMI levels correlate with fluctuations in insurance charges, highlighting specific BMI ranges that might be of interest for further investigation.

**Dependents Variation Trend**: Adopting a line chart to outline the relationship between the number of dependents and average insurance charges. The trend line across the chart offers insight into how changes in family size can affect insurance costs, suggesting a potential economy of scale with an increasing number of dependents.

### Results and Validation:

****

**Figure 1. Optimizing Health Insurance through Power BI Visualizations**

The Power BI dashboard provided an in-depth view of health insurance data, enabling stakeholders to thoroughly analyze and navigate the complex interactions involved. Insights derived included the significant impact of smoking on charges, age-related trends in insurance costs and regional variations which will pave the way for informed decision-making and policy development.

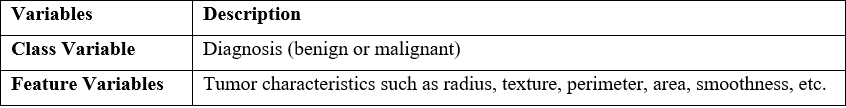
## Enhancing Breast Cancer Diagnosis with Classification Techniques

Globally, breast cancer emerges as the second leading cancer predominantly affecting younger women and is a primary cause of death from cancer (Nilashi et al., 2017). Timely and accurate diagnosis of breast cancer is critical for effective treatment and patient outcomes. The primary challenge lies in the reliable classification of tumors as either malignant or benign, a task complicated by the subtle differences in tumor characteristics. Hence, the healthcare sector continually seeks improvements in diagnostic methods to distinguish between benign and malignant tumors (Qawqzeh et al., 2023).

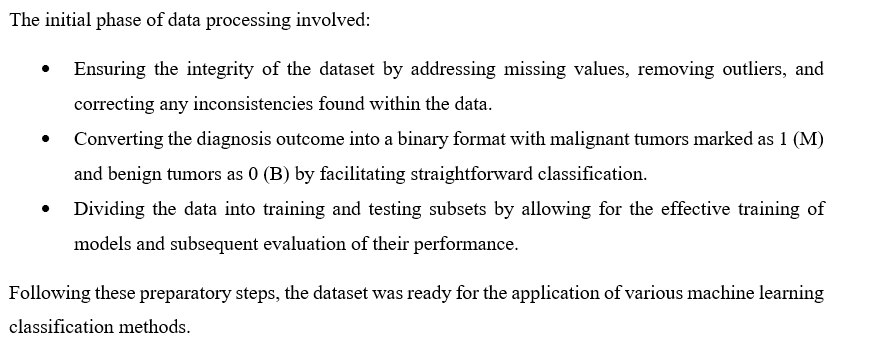
### Dataset Overview

(Breast Cancer Dataset) was taken from Kaggle. This dataset contains features derived from digital images of breast mass offering a foundation for applying machine learning to improve diagnostic accuracy.

The objective is to develop a machine learning model that can accurately classify tumors, thereby assisting in early detection and improved patient outcomes. To achieve this, several classification methods were employed, each utilizing a set of specific variables. These variables play a crucial role in training the models and are categorized as follows:



### Data Processing and Analysis

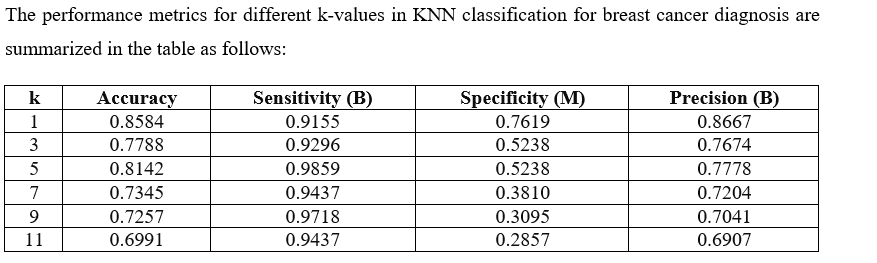
****

### Methods Employed

The analysis employed four primary classification methods executed within the R Studio environment following the import of the dataset:

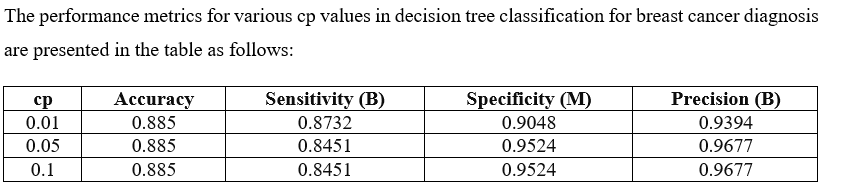
### K-Nearest Neighbors (KNN)

KNN is a simple, non-parametric method used for classification (and regression), assigns a class to a sample based on the majority class among its k nearest neighbors (Alnowaiser, 2024).



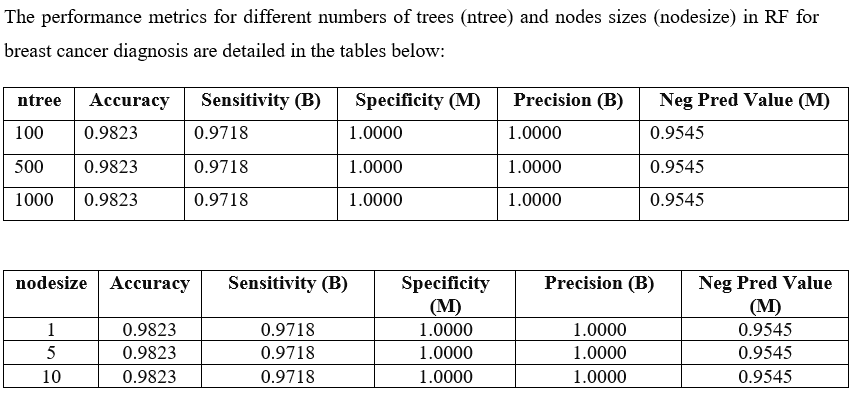
### Decision Trees (DT)

Decision Trees are flowchart-like structures where internal nodes represent tests on attributes, branches represent outcomes of those tests, and leaf nodes represent classes or class distributions (Alnowaiser, 2024).



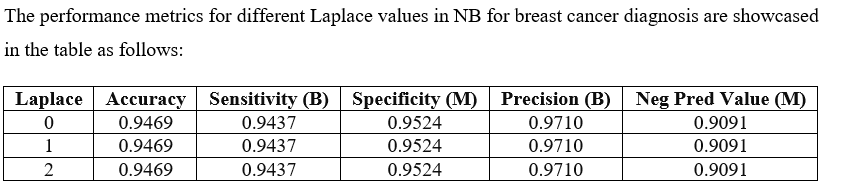
### Random Forest (RF)

An ensemble learning method that operates by constructing multiple decision trees during training time and outputting the class that is the mode of the classes of the individual trees (Jackins et al., 2021).



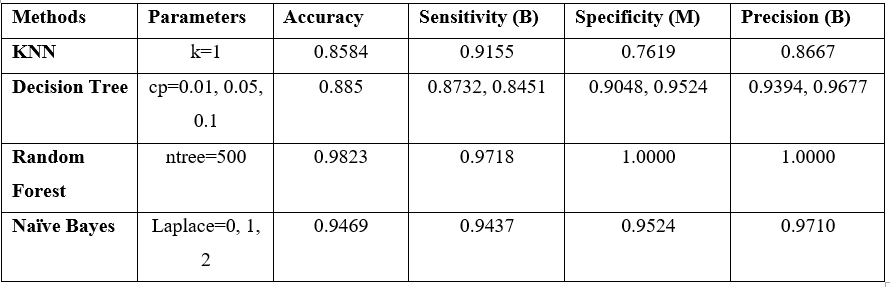
### Naive Bayes (NB)

A probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features (Jackins et al., 2021).



### Results and Discussion

The comparison of the performance metrics of the various classification methods applied in the analysis is as follows:

****

Random Forest shows the highest Accuracy and Specificity indicating its strong performance in correctly identifying all malignant cases without any false positives, thus making it particularly reliable for identifying malignant cases in this dataset.

# Barriers

The successful adoption of visualization and classification analytics in healthcare organizations is not without its challenges. From a technical perspective, the complexity of healthcare data, which often includes unstructured text, images, and complex relationships, poses significant challenges. The assumption that existing data is ready for analysis is rarely true and substantial pre-processing is often required, including data cleaning and feature extraction, which can be resource-intensive. Technical constraints also involve integrating these analytics into legacy systems, which may not be equipped to handle large-scale data analysis or real-time processing (Hemanth et al., 2020).

From a social perspective, there is often resistance to change within organizations. Healthcare professionals may be skeptical of analytics, viewing them as a potential threat to their expertise or fearing that algorithms might replace human decision-making. There is also the challenge of ensuring data privacy and meeting compliance standards, which can limit the use of sensitive data. Another barrier is the potential misinterpretation of visual data representations, leading to incorrect conclusions if not properly designed or explained (Anouncia et al., 2020).

The success of such initiatives also assumes a level of digital literacy that may not be present in all parts of the organization, necessitating extensive training and cultural adaptation. Furthermore, without a clear understanding of the benefits, stakeholders may be hesitant to allocate budget and resources towards adopting these technologies (Austin et al., 2021).

# Recommendations

Listed below are the recommendations based on case examples:

* Successful Power BI use in health insurance analysis highlights the need for accessible, in-depth tools and crucial training for effective insights.
* Breast cancer diagnosis shows classification techniques improve care. Providers could collaborate with AI experts and update models with new data for enhanced accuracy (Jackins et al., 2021).
* Could emphasize the integration of visualization tools across all departments to support data-driven decisions in real-time.
* Could strengthen data governance frameworks to ensure the quality, security, and privacy of healthcare data used in analytics (Anouncia et al., 2020).

# Roadmap

**Phase 1: Establishing a Foundation for Innovation**

* Instituting a robust data management framework is key, ensuring data privacy, security, and integrity, critical for our advanced analytics ambitions (Anouncia et al., 2020).
* Targeted training on Power BI and machine learning ensures our team's proficiency for improved decision-making.

**Phase 2: Demonstrating Value through Pilot Projects**

* Launching pilot projects in high-impact areas like insurance cost analysis and diagnostic accuracy will demonstrate advanced analytics' benefits in operations and care.
* Establishing a feedback mechanism from end-users, including providers and patients, ensures our analytics tools meet real-world needs (Ozsahin and Ozsahin, 2021).

**Phase 3: Scaling Success for System-wide Integration**

* After successful pilots, analytic tools can be expanded across departments for daily operations and decision-support systems.
* Implementing continuous learning programs ensures our team stays ahead for superior patient care.

# Conclusion

Key takeaways from this report on adopting visualization and classification techniques include the potential for enhanced data-driven decision-making, improved operational efficiencies, and better patient outcomes. However, the report acknowledges that the path to successful adoption is multifaceted, requiring careful consideration of both technical and social factors.

This report provided an overview and recommendations but has limitations. It does not offer a detailed technical blueprint or account for the specific financial implications of adoption. Also, the dataset chosen are from an open source and might cover limited variables by not capturing all factors that influence the outcomes.

# References

ALNOWAISER, K. 2024. Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model. *IEEE access,* 12**,** 16783-16793.

ANOUNCIA, S. M., GOHEL, H. A. & VAIRAMUTHU, S. 2020. *Data Visualization: Trends and Challenges Toward Multidisciplinary Perception,* Singapore, Springer.

AUSTIN, R. R., MATHIASON, M. A., LINDQUIST, R. A., MCMAHON, S. K., PIECZKIEWICZ, D. S. & MONSEN, K. A. 2021. Understanding Women's Cardiovascular Health Using MyStrengths MyHealth: A Patient-Generated Data Visualization Study of Strengths, Challenges, and Needs Differences Robin R. Austin, PhD, DNP, DC, RN-BC, FAMIA, FNAP1 , Michelle A. Mathiason. *Journal of nursing scholarship,* 53**,** 634-642.

BREAST CANCER DATASET, K. Available: <https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset> [Accessed 03-03-2024].

CHAKRABORTY, C., GHOSH, U., RAVI, V. & SHELKE, Y. 2021. *Efficient Data Handling for Massive Internet of Medical Things: Healthcare Data Analytics,* Cham, Springer International Publishing AG.

CHIANG, J.-K., LIN, C.-W., WANG, C.-L., KOO, M. & KAO, Y.-H. 2017. Cancer studies based on secondary data analysis of the Taiwan's National Health Insurance Research Database: A computational text analysis and visualization study. *Medicine (Baltimore),* 96**,** e6704-e6704.

CRISTEA, D., SACĂREĂ, C. & ŞOTROPA, D.-F. 2020. Knowledge discovery and visualization in healthcare datasets using formal concept analysis and graph databases. 2020. 35-42.

ELGENDI, M., HOWARD, N., HUSSAIN, A., MENON, C. & WARD, R. 2020. From ancient times to modern: realizing the power of data visualization in healthcare and medicine. *Big data analytics,* 5**,** 1-7.

HEMANTH, J., BHATIA, M. & GEMAN, O. 2020. *Data visualization and knowledge engineering : spotting data points with artificial intelligence / Jude Hemanth, Madhulika Bhatia, Oana Geman, editors,* Cham, Springer.

INSURANCE, H. Available: <https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance> [Accessed 2024].

JACKINS, V., VIMAL, S., KALIAPPAN, M. & LEE, M. Y. 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of supercomputing,* 77**,** 5198-5219.

KRISTINA, B. C., HADI, A. F., RISKI, A., KAMSYAKAWUNI, A. & ANGGRAENI, D. 2020. The visualization and classification method of support vector machine in lymphoma cancer. *Journal of Physics: Conference Series,* 1613**,** 12065.

KULKARNI, A. J., SIARRY, P., SINGH, P. K., ABRAHAM, A., ZHANG, M., ZOMAYA, A. Y. & BAKI, F. 2020. *Big data analytics in healthcare / Anand J. Kulkarni, Patrick Siarry, Pramod Kumar Singh, Ajith Abraham, Mengjie Zhang, Albert Zomaya, Fazle Baki, editors,* Cham, Springer.

MOHANTY, S. N. 2021. *Machine learning for healthcare applications / edited by Sachi Nandan Mohanty [and three others],* Hoboken, New Jersey, Wiley.

NILASHI, M., IBRAHIM, O., AHMADI, H. & SHAHMORADI, L. 2017. A knowledge-based system for breast cancer classification using fuzzy logic method. *Telematics and informatics,* 34**,** 133-144.

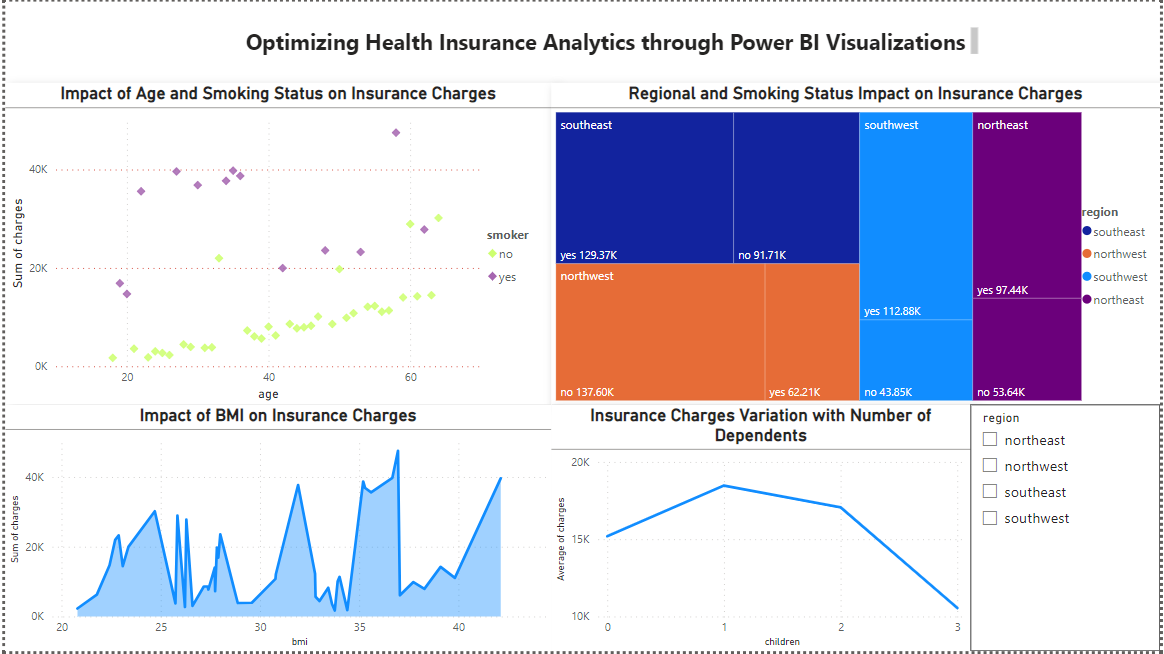
OZSAHIN, I. & OZSAHIN, D. U. 2021. *Applied Machine Learning and Multi-Criteria Decision-making in Healthcare,* UAE, Bentham Science Publishers.

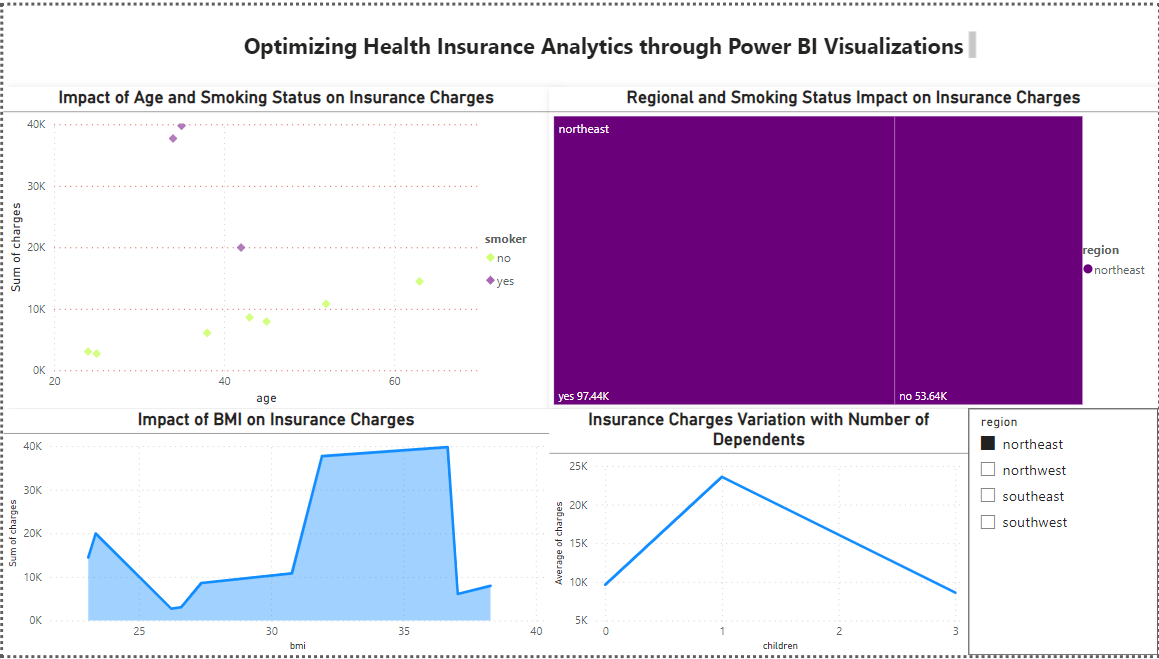
QAWQZEH, Y. K., ALOURANI, A. & GHWANMEH, S. 2023. An Improved Breast Cancer Classification Method Using an Enhanced AdaBoost Classifier. *International journal of advanced computer science & applications,* 14.

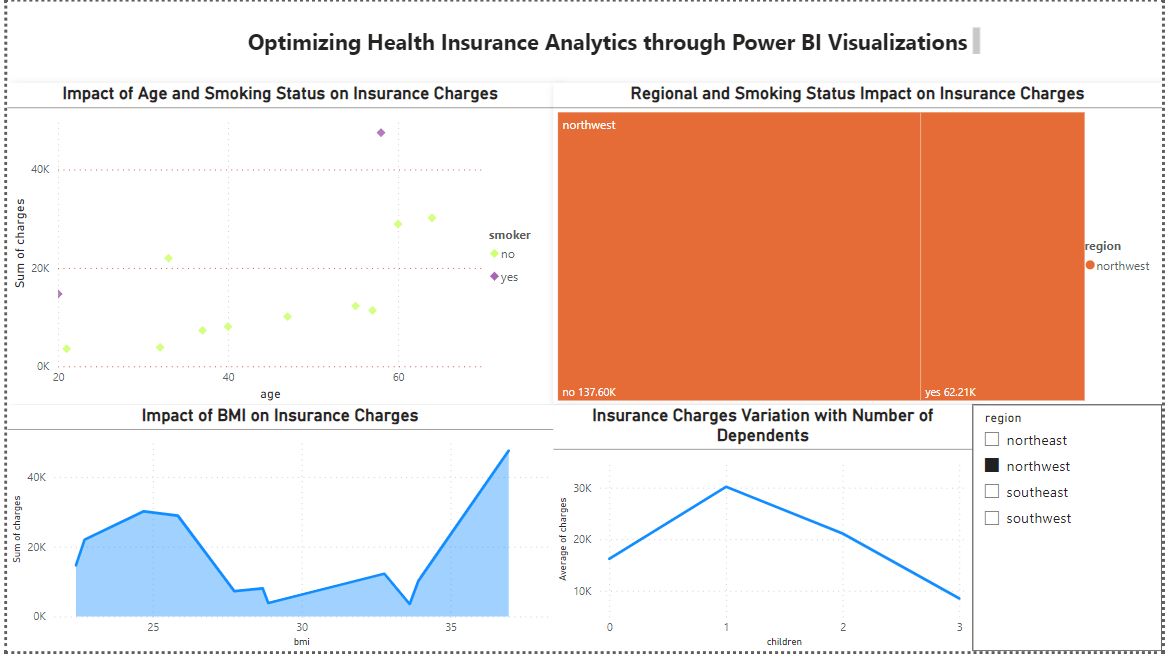
# Appendix

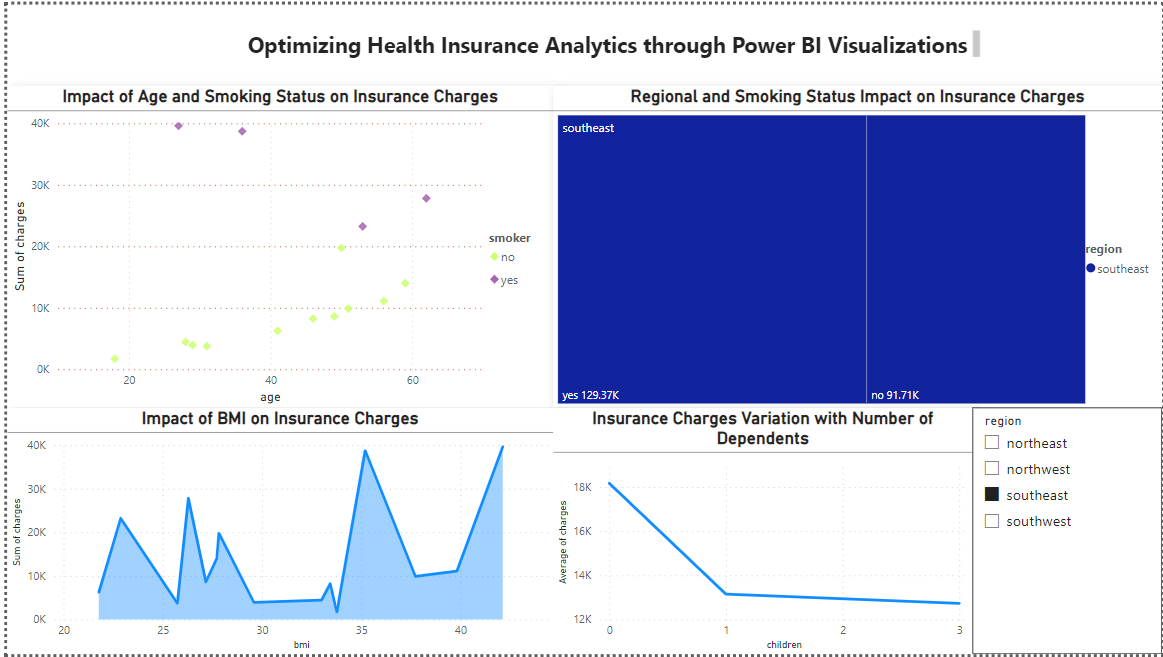
**SLICER Tool in the Power BI dashboards in the insurance dataset**

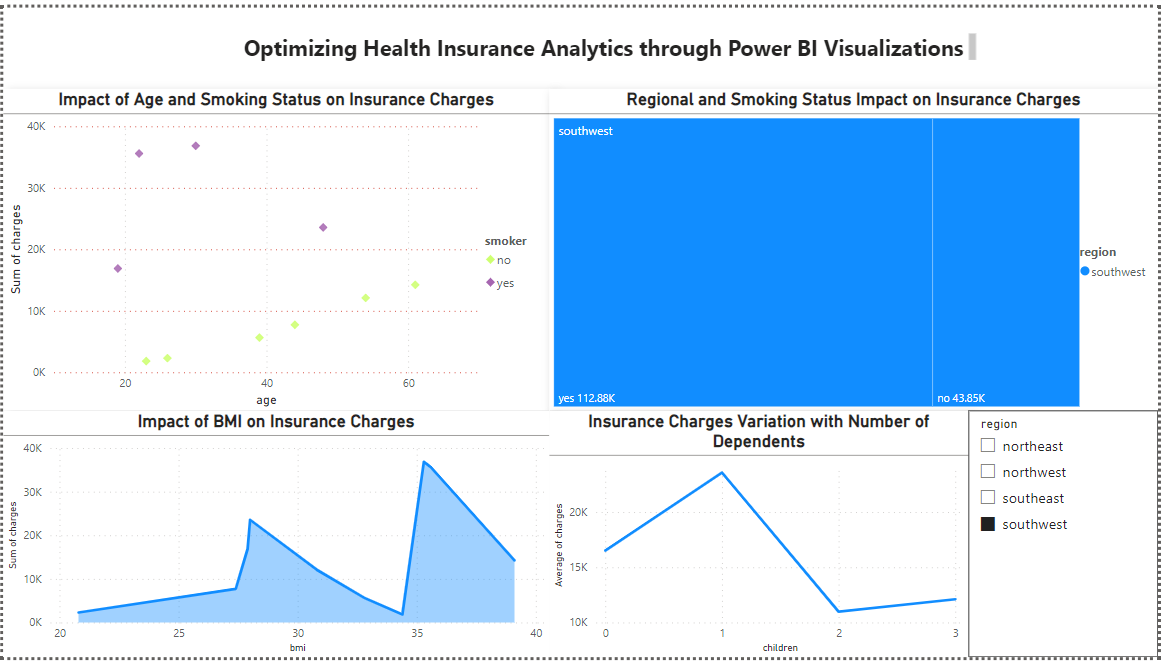
Slicer tool in Power BI provide a visual and interactive way for report viewers to slice and dice the data, enabling a dynamic and customizable analysis experience as shown below.

****

****

****

****

****

**Outputs of the Breast Cancer Dataset in R Studio**

> View(breast.cancer)

> head(breast.cancer)

id diagnosis radius\_mean texture\_mean perimeter\_mean area\_mean

1 842302 M 17.99 10.38 122.80 1001.0

2 842517 M 20.57 17.77 132.90 1326.0

3 84300903 M 19.69 21.25 130.00 1203.0

4 84348301 M 11.42 20.38 77.58 386.1

5 84358402 M 20.29 14.34 135.10 1297.0

6 843786 M 12.45 15.70 82.57 477.1

smoothness\_mean compactness\_mean concavity\_mean concave.points\_mean

1 0.11840 0.27760 0.3001 0.14710

2 0.08474 0.07864 0.0869 0.07017

3 0.10960 0.15990 0.1974 0.12790

4 0.14250 0.28390 0.2414 0.10520

5 0.10030 0.13280 0.1980 0.10430

6 0.12780 0.17000 0.1578 0.08089

symmetry\_mean fractal\_dimension\_mean radius\_se texture\_se perimeter\_se area\_se

1 0.2419 0.07871 1.0950 0.9053 8.589 153.40

2 0.1812 0.05667 0.5435 0.7339 3.398 74.08

3 0.2069 0.05999 0.7456 0.7869 4.585 94.03

4 0.2597 0.09744 0.4956 1.1560 3.445 27.23

5 0.1809 0.05883 0.7572 0.7813 5.438 94.44

6 0.2087 0.07613 0.3345 0.8902 2.217 27.19

smoothness\_se compactness\_se concavity\_se concave.points\_se symmetry\_se

1 0.006399 0.04904 0.05373 0.01587 0.03003

2 0.005225 0.01308 0.01860 0.01340 0.01389

3 0.006150 0.04006 0.03832 0.02058 0.02250

4 0.009110 0.07458 0.05661 0.01867 0.05963

5 0.011490 0.02461 0.05688 0.01885 0.01756

6 0.007510 0.03345 0.03672 0.01137 0.02165

fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst area\_worst

1 0.006193 25.38 17.33 184.60 2019.0

2 0.003532 24.99 23.41 158.80 1956.0

3 0.004571 23.57 25.53 152.50 1709.0

4 0.009208 14.91 26.50 98.87 567.7

5 0.005115 22.54 16.67 152.20 1575.0

6 0.005082 15.47 23.75 103.40 741.6

smoothness\_worst compactness\_worst concavity\_worst concave.points\_worst

1 0.1622 0.6656 0.7119 0.2654

2 0.1238 0.1866 0.2416 0.1860

3 0.1444 0.4245 0.4504 0.2430

4 0.2098 0.8663 0.6869 0.2575

5 0.1374 0.2050 0.4000 0.1625

6 0.1791 0.5249 0.5355 0.1741

symmetry\_worst fractal\_dimension\_worst

1 0.4601 0.11890

2 0.2750 0.08902

3 0.3613 0.08758

4 0.6638 0.17300

5 0.2364 0.07678

6 0.3985 0.12440

**k-Nearest Neighbor**

[1] "Results for k= 1"

Confusion Matrix and Statistics

Reference

Prediction B M

B 65 10

M 6 32

Accuracy : 0.8584

95% CI : (0.7803, 0.9168)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 5.334e-08

Kappa : 0.6908

Mcnemar's Test P-Value : 0.4533

Sensitivity : 0.9155

Specificity : 0.7619

Pos Pred Value : 0.8667

Neg Pred Value : 0.8421

Prevalence : 0.6283

Detection Rate : 0.5752

Detection Prevalence : 0.6637

Balanced Accuracy : 0.8387

'Positive' Class : B

[1] "Results for k= 3"

Confusion Matrix and Statistics

Reference

Prediction B M

B 66 20

M 5 22

Accuracy : 0.7788

95% CI : (0.691, 0.8514)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 0.0004449

Kappa : 0.4891

Mcnemar's Test P-Value : 0.0051103

Sensitivity : 0.9296

Specificity : 0.5238

Pos Pred Value : 0.7674

Neg Pred Value : 0.8148

Prevalence : 0.6283

Detection Rate : 0.5841

Detection Prevalence : 0.7611

Balanced Accuracy : 0.7267

'Positive' Class : B

[1] "Results for k= 5"

Confusion Matrix and Statistics

Reference

Prediction B M

B 70 20

M 1 22

Accuracy : 0.8142

95% CI : (0.7301, 0.8811)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 1.419e-05

Kappa : 0.5616

Mcnemar's Test P-Value : 8.568e-05

Sensitivity : 0.9859

Specificity : 0.5238

Pos Pred Value : 0.7778

Neg Pred Value : 0.9565

Prevalence : 0.6283

Detection Rate : 0.6195

Detection Prevalence : 0.7965

Balanced Accuracy : 0.7549

'Positive' Class : B

[1] "Results for k= 7"

Confusion Matrix and Statistics

Reference

Prediction B M

B 67 26

M 4 16

Accuracy : 0.7345

95% CI : (0.6432, 0.8132)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 0.011256

Kappa : 0.3635

Mcnemar's Test P-Value : 0.000126

Sensitivity : 0.9437

Specificity : 0.3810

Pos Pred Value : 0.7204

Neg Pred Value : 0.8000

Prevalence : 0.6283

Detection Rate : 0.5929

Detection Prevalence : 0.8230

Balanced Accuracy : 0.6623

'Positive' Class : B

[1] "Results for k= 9"

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 29

M 2 13

Accuracy : 0.7257

95% CI : (0.6337, 0.8054)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 0.01889

Kappa : 0.3239

Mcnemar's Test P-Value : 3.016e-06

Sensitivity : 0.9718

Specificity : 0.3095

Pos Pred Value : 0.7041

Neg Pred Value : 0.8667

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.8673

Balanced Accuracy : 0.6407

'Positive' Class : B

[1] "Results for k= 11"

Confusion Matrix and Statistics

Reference

Prediction B M

B 67 30

M 4 12

Accuracy : 0.6991

95% CI : (0.6057, 0.7818)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 0.07062

Kappa : 0.2626

Mcnemar's Test P-Value : 1.807e-05

Sensitivity : 0.9437

Specificity : 0.2857

Pos Pred Value : 0.6907

Neg Pred Value : 0.7500

Prevalence : 0.6283

Detection Rate : 0.5929

Detection Prevalence : 0.8584

Balanced Accuracy : 0.6147

'Positive' Class : B

**Decision Tree**

Results for cp = 0.01

Confusion Matrix and Statistics

Reference

Prediction B M

B 62 4

M 9 38

Accuracy : 0.885

95% CI : (0.8113, 0.9373)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 8.286e-10

Kappa : 0.7595

Mcnemar's Test P-Value : 0.2673

Sensitivity : 0.8732

Specificity : 0.9048

Pos Pred Value : 0.9394

Neg Pred Value : 0.8085

Prevalence : 0.6283

Detection Rate : 0.5487

Detection Prevalence : 0.5841

Balanced Accuracy : 0.8890

'Positive' Class : B

Results for cp = 0.05

Confusion Matrix and Statistics

Reference

Prediction B M

B 60 2

M 11 40

Accuracy : 0.885

95% CI : (0.8113, 0.9373)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 8.286e-10

Kappa : 0.764

Mcnemar's Test P-Value : 0.0265

Sensitivity : 0.8451

Specificity : 0.9524

Pos Pred Value : 0.9677

Neg Pred Value : 0.7843

Prevalence : 0.6283

Detection Rate : 0.5310

Detection Prevalence : 0.5487

Balanced Accuracy : 0.8987

'Positive' Class : B

Results for cp = 0.1

Confusion Matrix and Statistics

Reference

Prediction B M

B 60 2

M 11 40

Accuracy : 0.885

95% CI : (0.8113, 0.9373)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 8.286e-10

Kappa : 0.764

Mcnemar's Test P-Value : 0.0265

Sensitivity : 0.8451

Specificity : 0.9524

Pos Pred Value : 0.9677

Neg Pred Value : 0.7843

Prevalence : 0.6283

Detection Rate : 0.5310

Detection Prevalence : 0.5487

Balanced Accuracy : 0.8987

'Positive' Class : B

**Random Forest**

Results for ntree = 100

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 0

M 2 42

Accuracy : 0.9823

95% CI : (0.9375, 0.9978)

No Information Rate : 0.6283

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9625

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.9718

Specificity : 1.0000

Pos Pred Value : 1.0000

Neg Pred Value : 0.9545

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9859

'Positive' Class : B

Results for ntree = 500

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 0

M 2 42

Accuracy : 0.9823

95% CI : (0.9375, 0.9978)

No Information Rate : 0.6283

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9625

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.9718

Specificity : 1.0000

Pos Pred Value : 1.0000

Neg Pred Value : 0.9545

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9859

'Positive' Class : B

Results for ntree = 1000

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 0

M 2 42

Accuracy : 0.9823

95% CI : (0.9375, 0.9978)

No Information Rate : 0.6283

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9625

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.9718

Specificity : 1.0000

Pos Pred Value : 1.0000

Neg Pred Value : 0.9545

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9859

'Positive' Class : B

Varying node size:

Results for nodesize = 1

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 0

M 2 42

Accuracy : 0.9823

95% CI : (0.9375, 0.9978)

No Information Rate : 0.6283

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9625

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.9718

Specificity : 1.0000

Pos Pred Value : 1.0000

Neg Pred Value : 0.9545

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9859

'Positive' Class : B

Results for nodesize = 5

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 0

M 2 42

Accuracy : 0.9823

95% CI : (0.9375, 0.9978)

No Information Rate : 0.6283

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9625

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.9718

Specificity : 1.0000

Pos Pred Value : 1.0000

Neg Pred Value : 0.9545

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9859

'Positive' Class : B

Results for nodesize = 10

Confusion Matrix and Statistics

Reference

Prediction B M

B 69 0

M 2 42

Accuracy : 0.9823

95% CI : (0.9375, 0.9978)

No Information Rate : 0.6283

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9625

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.9718

Specificity : 1.0000

Pos Pred Value : 1.0000

Neg Pred Value : 0.9545

Prevalence : 0.6283

Detection Rate : 0.6106

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9859

'Positive' Class : B

**Naïve Bayes:**

Laplace 0

Confusion Matrix and Statistics

Reference

Prediction B M

B 67 2

M 4 40

Accuracy : 0.9469

95% CI : (0.888, 0.9803)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 1.866e-15

Kappa : 0.8874

Mcnemar's Test P-Value : 0.6831

Sensitivity : 0.9437

Specificity : 0.9524

Pos Pred Value : 0.9710

Neg Pred Value : 0.9091

Prevalence : 0.6283

Detection Rate : 0.5929

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9480

'Positive' Class : B

Laplace 1

Confusion Matrix and Statistics

Reference

Prediction B M

B 67 2

M 4 40

Accuracy : 0.9469

95% CI : (0.888, 0.9803)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 1.866e-15

Kappa : 0.8874

Mcnemar's Test P-Value : 0.6831

Sensitivity : 0.9437

Specificity : 0.9524

Pos Pred Value : 0.9710

Neg Pred Value : 0.9091

Prevalence : 0.6283

Detection Rate : 0.5929

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9480

'Positive' Class : B

Laplace 2

Confusion Matrix and Statistics

Reference

Prediction B M

B 67 2

M 4 40

Accuracy : 0.9469

95% CI : (0.888, 0.9803)

No Information Rate : 0.6283

P-Value [Acc > NIR] : 1.866e-15

Kappa : 0.8874

Mcnemar's Test P-Value : 0.6831

Sensitivity : 0.9437

Specificity : 0.9524

Pos Pred Value : 0.9710

Neg Pred Value : 0.9091

Prevalence : 0.6283

Detection Rate : 0.5929

Detection Prevalence : 0.6106

Balanced Accuracy : 0.9480

'Positive' Class : B