**NAME: BEULAH NWOKOTUBO**

**STUDENT NUMBER: 041097385**

**DATE: FEBRUARY 2ND, 2024**

**COURSE: CST8390**

**TITLE: ASSIGNMENT 1**

Contents

[INTRODUCTION 2](#_Toc157786671)

[BUSINESS UNDERSTANDING 3](#_Toc157786672)

[DATA UNDERSTANDING 3](#_Toc157786673)

[3.1. Collect Initial Data 3](#_Toc157786674)

[3.2. Describe Data 3](#_Toc157786675)

[3.3. Explore Data 4](#_Toc157786676)

[3.4. Verify Data Quality 4](#_Toc157786677)

[Data Quality Issues 4](#_Toc157786678)

[DATA PREPARATION 4](#_Toc157786679)

[4.1 Select Data 4](#_Toc157786680)

[4.2. Clean Data 5](#_Toc157786681)

[4.3. Construct Data 5](#_Toc157786682)

[4.4. Integrate Data 5](#_Toc157786683)

[4.5. Format Data 5](#_Toc157786684)

[MODELING 6](#_Toc157786685)

[kNN Testing: 6](#_Toc157786686)

[Decision Tree Parameters Testing: 7](#_Toc157786687)

[Parameters for the Ideal J48 Decision Tree 8](#_Toc157786688)

[Rules for the Obtained Tree 8](#_Toc157786689)

[Explanation of Rules 9](#_Toc157786690)

[COMPARISON OF RESULTS 10](#_Toc157786691)

[CONCLUSION 10](#_Toc157786692)

[REFERENCES 10](#_Toc157786693)

# INTRODUCTION

Breast cancer is a prevalent and serious health concern that requires comprehensive analysis and classification methods to enhance early detection and treatment planning. This report explores and analyzes the Wisconsin Breast Cancer (Original) dataset, containing crucial attributes associated with breast cancer instances. The most recent parts of this dataset were collected in 1992 and reflects information from clinical cases reported by Dr. William H. Wolberg up to that point. The main objective is to employ classification techniques, namely k-Nearest Neighbors (kNN) and Decision Trees, using CRISP-DM. The goal of this exploration is to gain valuable insights into the dataset and evaluate the performance of different classification algorithms.

# BUSINESS UNDERSTANDING

The Wisconsin Breast Cancer dataset consists of 10 attributes (features of breast cancer) and 699 instances. The dataset reflects a chronological grouping of clinical cases, providing insights into the temporal distribution of the data.   
  
**Papers Citing the Dataset:**  
**1. On the Bias of Precision Estimation Under Separate Sampling**

Relevance: Logistic Regression Classification

The bias in the estimation of the precision of the classification methods can show strong bias depending on the prevalence used (sample vs. true population) [1].

**2. RBF Kernel Optimization Method with Particle Swarm Optimization on SVM**

Relevance: Support Vector Classification

Analysis performed by R. Indraswari and A. Arifin in 2017 also showed that using Particle Swarm Optimization on Support Vector Machine to optimize RBF kernel parameters provided less complexity and higher accuracy for SVM [2].

**3. Iteratively Reweighted Least Squares Algorithms for L1-Norm Principal Component Analysis**

Relevance: Neural Network Classification

A paper by S. Tsumoto, S. Hirano introduces reweighted algorithms for L1 PCA which consistently outperform existing methods in computational experiments.

**4. Widened KRIMP: Better Performance through Diverse Parallelism**

Relevance: Random Forest Classification

Another paper citing this dataset demonstrates the applicability of the Widening framework to machine learning algorithms, specifically Krimp, an itemset mining algorithm. Parallelizing the search using Widening finds better solutions in nearly the same time as the original sequential algorithm [4].

**5. Automated Empirical Selection of Rule Induction Methods Based on Recursive Iteration of Resampling Methods and Multiple Testing**

Relevance: Xgboost Classification

One more paper proposes a method for multiple testing based on recursive iteration of resampling methods for rule induction. Applied to medical databases, the method yields the best selection of estimation methods in various cases, showcasing its effectiveness [5].

# DATA UNDERSTANDING

## 3.1. Collect Initial Data

The initial data collection was conducted using the original Breast Cancer Wisconsin dataset. This data is opensource and readily available on the UC Irvine Machine Learning Repository at <https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>.

## 3.2. Describe Data

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Datatype to be used in Weka** |
| Sample code number | ID number associated with each data instance. | Numeric |
| Clump Thickness | Thickness of cell clumps | Numeric |
| Uniformity of Cell Size | Uniformity/regularity in the size of cells | Numeric |
| Uniformity of Cell Shape | Uniformity/regularity in the shape of cells | Numeric |
| Marginal Adhesion | How adhesive cells are to adjacent cells | Numeric |
| Single Epithelial Cell Size | The size of one epithelial cell | Numeric |
| Bare Nuclei | Presence of cell nuclei with no surrounding cytoplasm | Numeric |
| Bland Chromatin | Uniformity of nucleus texture | Numeric |
| Normal Nucleoli | Presence of physiological nucleoli | Numeric |
| Mitoses | Presence of uncontrolled mitosis | Numeric |

|  |  |  |
| --- | --- | --- |
| **Class** | **Diagnosis class ((2 for benign, 4 for malignant)** | **Nominal** |

## 3.3. Explore Data

**Charts**

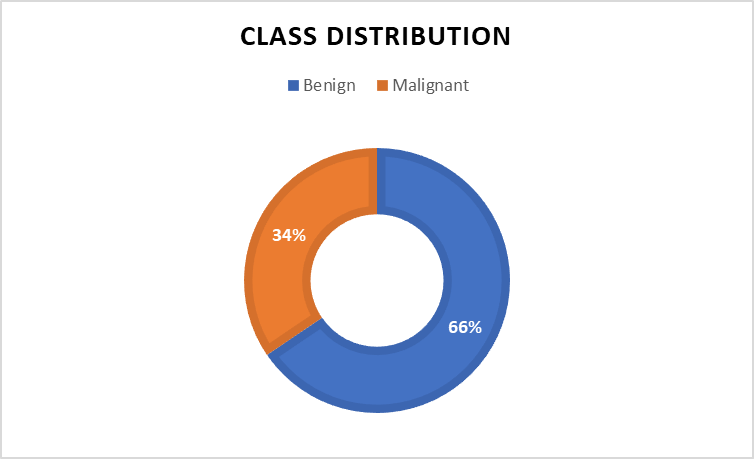
****

Figure 1: Class distribution Pie Chart

Explanation: This chart shows the prevalence of benign tumor types over malignant tumors in the dataset. Examining these proportions helps understand the balance of the classes, which is important for designing and evaluating classification models.

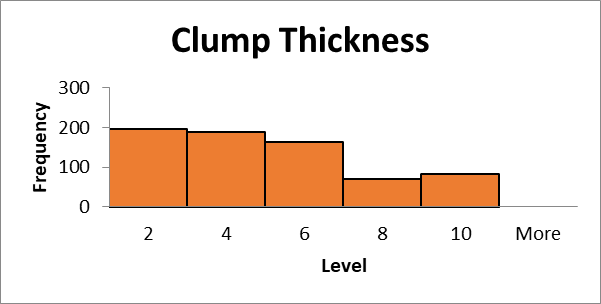
****

Figure 2: Histogram for Clump Thickness

**Explanation**: This histogram, displaying a right-skewed unimodal distribution, indicates that the majority of instances in the dataset exhibit low to moderate clump thickness. In benign cases, cells typically form monolayers, while cancerous cells tend to arrange in multilayers [6]. Considering that the dataset predominantly comprises benign diagnoses, this chart highlights a proportional correlation between clump thickness and diagnosis, emphasizing the prevalence of lower thickness values in benign instances.

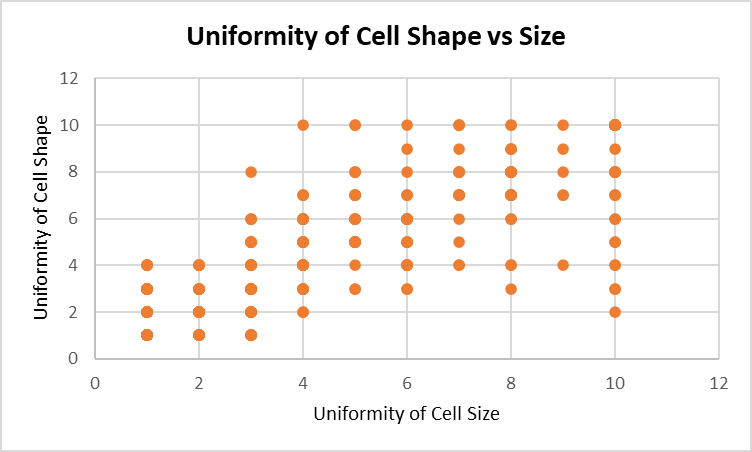
****

Figure 3: Scatter Plot for Uniformity of Cell Shape vs. Uniformity of Cell Size

**Explanation**: This scatter plot shows a relatively high positive correlation between the uniformity of cell size and uniformity of cell shape. The points on the plot show a discernible pattern, indicating that as the cell shape uniformity increases or decreases, so does that of the cell size.

## 3.4. Verify Data Quality

### Data Quality Issues

1. The dataset has 16 **missing values** for the Bare Nuclei attribute.
2. The dataset is also **outdated** as data was last collected in 1992.

# DATA PREPARATION

## 4.1 Select Data

This step included loading the data into Weka and deciding which data would be useful for data analytics. The conclusion was that the data **“Sample code number”** attribute would not be useful as it is an id column irrelevant to the diagnosis, and it was consequentially excluded from the analytics process.

## 4.2. Clean Data

Although there are missing values for the Bare Nuclei attribute in this dataset, classification methods like kNN handle missing data by estimating the missing values based on the characteristics of neighboring data points. As a result, these missing data points were neither manually imputed, nor the instances removed from the set. Instead, kNN was used for handling.

## 4.3. Construct Data

In the "Construct Data" phase, the focus is on creating new attributes or transforming existing ones to enhance the dataset's informative value for classification. For the Breast Cancer Wisconsin dataset, the original attributes were deemed sufficient for analysis, and no explicit data construction steps were performed.

However, in scenarios where additional attributes or features are required, this phase involves creating composite variables to better represent underlying patterns in the data. Techniques such as binning, scaling, or combining existing attributes may be applied to generate more meaningful insights.

In the context of the Breast Cancer Wisconsin dataset, the attributes like “Clump Thickness,” “Uniformity of Cell Size,” and others were considered as is, without requiring further construction. The decision to construct data depends on the specific characteristics of the dataset and the goals of the analysis. In cases where the original attributes do not fully capture the complexities of the problem, constructing new features becomes a valuable step in improving the predictive capabilities of the models.

## 4.4. Integrate Data

This step involves merging multiple datasets or combining data from different sources to create a comprehensive dataset for analysis. In the case of the Breast Cancer Wisconsin dataset, the data provided in the original file was considered standalone and did not require integration with additional external datasets.

Integration is crucial when information from different sources is needed to provide a more holistic view of the problem. Techniques such as merging, joining, or concatenating datasets may be employed during this phase. In my analysis, since the dataset was self-contained, no external integration was necessary.

## 4.5. Format Data

A screenshot of a computer

Description automatically generated

Figure 4: Screenshot of class distribution

All the attribute except “Class” were left as Numeric. The class attribute was converted to Nominal in the following steps:

1. Open Weka Explorer.
2. Click on the "Open File" button in the Preprocess tab.
3. Select the Breast Cancer Wisconsin dataset file.
4. In the “Attributes” section, review the types of each attribute to ensure they match expectations. The class attribute is numeric but should be nominal.
5. In the “Preprocess” tab, click on the “Choose” button next to the “Filter” section.
6. Click on the dropdown arrow for “filter” >> “unsupervised” >> “attribute”.
7. Under the “attribute” menu, select NumerictoNominal.
8. Click on the filter bar.
9. Input the index of the attribute to be changed under “attributeIndices” in dialogue box that appears (11 if the sample code number is still in the list).
10. Click “Apply” under “Filter” to apply the filter.
11. Confirm that the filter was applied.

# MODELING

Modeling for this dataset has been done using kNN and Decision Tree classification methods with 10-fold Cross Validation. For the kNN method, various k’s (odd numbers from 3 - 15) were tested and the best k was chosen. For the J48 Decision Tree, various parameters were tested, and the best tree chosen. These choices were made based on accuracy, precision, sensitivity(recall), specificity, and F1 measure.

### kNN Testing:

|  |  |  |
| --- | --- | --- |
| **K** | **Accuracy** | **Number of instances misclassified in each class** |
| 3 | 96.85 % | a: 13  b: 9 |
| 5 | 96.71 % | a: 14  b: 9 |
| 7 | 96.57% | a: 13  b: 11 |
| 9 | 96.28% | a: 12  b: 14 |
| 11 | 96.28% | a: 12  b: 14 |
| 13 | 96.13% | a: 13  b: 14 |
| 15 | 96.56% | a: 12  b: 12 |

The best k is 3.

**Parameters for best kNN**

Number of Nearest Numbers: 3

Cross Validation Folds: 10

**A screenshot of a computer screen

Description automatically generated**

Figure 5: Decision Tree

### Decision Tree Parameters Testing:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Seed** | **minNumObj** | **Unpruned property** | **Correctly Classified** | **Confusion Matrix** |
| 1 | 2 | False | 94.56% | a b <-- classified as  438 20 | a = 2  18 223 | b = 4 |
| 1 | 20 | False | 94.42% | a b <-- classified as  436 22 | a = 2  17 224 | b = 4 |
| 1 | 12 | False | 93.99% | a b <-- classified as  438 20 | a = 2  22 219 | b = 4 |
| 1 | 30 | False | 92.70% | a b <-- classified as  434 24 | a = 2  27 214 | b = 4 |
| 1 | 4 | False | 95.42% | a b <-- classified as  441 17 | a = 2  15 226 | b = 4 |
| 1 | 4 | True | 94.85% | a b <-- classified as  441 17 | a = 2  19 222 | b = 4 |

The best decision tree has a seed of “1”, a minNumObj of “4”, and an Unpruned property of “False”. This tree has the best accuracy and precision.

### Parameters for the Ideal J48 Decision Tree

Seed: 1

Minimum Number of Objects: 4

Unpruned Property: False

Batch Size: 100

All other parameters were left as the default for Weka.

### Rules for the Obtained Tree

Uniformity of Cell Size <= 2

| Bare Nuclei <= 3: 2 (405.39/2.0)

| Bare Nuclei > 3

| | Clump thickness <= 3: 2 (11.55)

| | Clump thickness > 3: 4 (12.06/2.06)

Uniformity of Cell Size > 2

| Uniformity of Cell Shape <= 2

| | Clump thickness <= 5: 2 (19.0/1.0)

| | Clump thickness > 5: 4 (4.0)

| Uniformity of Cell Shape > 2

| | Uniformity of Cell Size <= 4

| | | Bare Nuclei <= 2

| | | | Normal Nucleoli <= 2: 2 (7.0)

| | | | Normal Nucleoli > 2: 4 (7.41/3.21)

| | | Bare Nuclei > 2

| | | | Clump thickness <= 6

| | | | | Uniformity of Cell Size <= 3: 4 (13.0/2.0)

| | | | | Uniformity of Cell Size > 3

| | | | | | Marginal Adhesion <= 5: 2 (5.79/1.0)

| | | | | | Marginal Adhesion > 5: 4 (5.0)

| | | | Clump thickness > 6: 4 (31.79/1.0)

| | Uniformity of Cell Size > 4: 4 (177.0/5.0)

Number of leaves: 12

Size of the tree: 23

### Explanation of Rules

The tree starts at the root node with a split on “Uniformity of Cell Size.”

Different conditions are applied at each node based on attributes like “Bare Nuclei,” “Clump Thickness,” “Uniformity of Cell Shape,” etc.

The leaf nodes represent the final classification decisions.

Rules Example:

If “Uniformity of Cell Size” is less than or equal to 2 and “Bare Nuclei” is less than or equal to 3, then the predicted class is 2 (benign).

# COMPARISON OF RESULTS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Sensitivity (Recall)** | **Specificity** | **F1 Measure** | **Confusion Matrix** |
| kNN | 96.9% | 96.9% | 96.9% | 96.9% | 96.9% | a b <-- classified as  445 13 | a = 2  9 232 | b = 4 |
| Decision Tree | 95.4% | 95.4% | 95.4% | 95.4% | 95.4% | a b <-- classified as  441 17 | a = 2  15 226 | b = 4 |

Where Specificity was calculated as:

The evaluation metrics considered for this comparison include accuracy, precision, sensitivity, specificity, and F1 measure.

Consideration of accuracy helps identify the overall correctness of predictions, while precision and sensitivity provide insights into class-specific performance. Specificity highlights the ability to correctly identify instances of the negative class.

The F1 measure, which combines precision and sensitivity, offers a balanced assessment of a model's overall performance. Comparing these metrics allows us to make informed decisions about the suitability of each method for the Breast Cancer Wisconsin dataset.

Based on the results displayed in the above table, kNN is the best choice for classification between the two as it provides better results across all measures of performance.

# CONCLUSION

In conclusion, this analysis of the Wisconsin Breast Cancer dataset using kNN and Decision Tree classification methods has provided valuable insights into the characteristics and behavior of breast cancer instances. Through the CRISP-DM methodology, I successfully navigated through business understanding, data understanding, data preparation, and modeling and evaluation phases. The kNN algorithm with a k value of 3 demonstrated much better performance, outperforming other k values. Meanwhile, Decision Tree classification revealed intricate rules governing the classification process. The comparison of results highlighted the strengths and limitations of each method, emphasizing the importance of algorithm selection based on specific dataset characteristics.

# REFERENCES

[1] S. Xie and U. Braga-Neto, “On the bias of precision estimation under separate sampling,” Semantics Scholar, [https://www.semanticscholar.org/paper/On-the-Bias-of-Precision-Estimation-Under-Separate-Xie-Braga-Neto/80cb856ae3437b249c62fe52502c9e5117a24f1b](https://www.semanticscholar.org/paper/On-the-Bias-of-Precision-Estimation-Under-Separate-Xie-Braga-Neto/80cb856ae3437b249c62fe52502c9e5117a24f1b%20) (accessed Jan. 29, 2024).

[2] R. Indraswari and A. Arifin, “RBF kernel optimization method with particle swarm optimization on SVM ...,” Semantics Scholar, <https://www.semanticscholar.org/paper/RBF-KERNEL-OPTIMIZATION-METHOD-WITH-PARTICLE-SWARM-Indraswari-Arifin/dced0d28800e162f8e48cb3fdb52aa731ae61892> (accessed Jan. 29, 2024).

[3] Y. W. Park and D. Klabjan, “iteratively reweighted least squares algorithms for L1-norm ...,” Semantics Scholar, <https://www.semanticscholar.org/paper/Iteratively-Reweighted-Least-Squares-Algorithms-for-Park-Klabjan/53cbfd21b9e0bfd0c0a869d411f920fbfeb1d383> (accessed Jan. 29, 2024).

[4] O. R. Sampson and M. Berthold, “Widened Krimp: Better performance through diverse parallelism,” Semantics Scholar, <https://www.semanticscholar.org/paper/Widened-KRIMP%3A-Better-Performance-through-Diverse-Sampson-Berthold/9e54ed633ebf64bd7d8ec082e2b84cd38008b2a7> (accessed Jan. 30, 2024).

[5] S. Tsumoto and S. Hirano, “Automated empirical selection of rule induction methods based on ...,” Semantics Scholar, <https://www.semanticscholar.org/paper/Automated-Empirical-Selection-of-Rule-Induction-on-Tsumoto-Hirano/5b82a7e1a7f1a1c4d23aa09f9bcd5dcca5f898bc> (accessed Jan. 30, 2024).

[6] R. López, “Diagnose breast cancer using machine learning,” Neural Designer, [https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis/#:~:text=Epithelial%20cells%20that%20are%20significantly,typically%20seen%20in%20benign%20tumors](https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis/%23:~:text=Epithelial%20cells%20that%20are%20significantly,typically%20seen%20in%20benign%20tumors%20) (accessed Jan. 30, 2024).