Assignment 1

24W\_CST8390 BI

Hasibullah Yosufi

041012318

Submission Date Feb 2nd

Contents

[1 Introduction 3](#_Toc157803898)

[2 Business Understanding 3](#_Toc157803899)

[3 Data Understanding 4](#_Toc157803900)

[3.1 Collect Data 4](#_Toc157803901)

[3.2 Describe Data 4](#_Toc157803902)

[3.3 Explore Data 5](#_Toc157803903)

[3.4 Verify Data Quality 9](#_Toc157803904)

[4 Data Preparation 10](#_Toc157803905)

[4.1 Select Data 10](#_Toc157803906)

[4.2 Clean Data 10](#_Toc157803907)

[4.3 Construct Data 11](#_Toc157803908)

[4.4 Integrate Data 11](#_Toc157803909)

[4.5 Format Data 11](#_Toc157803910)

[5 Modeling 12](#_Toc157803911)

[5.1 K-NN 12](#_Toc157803912)

[5.2 Decision Tree (J48 || C4.5) 14](#_Toc157803913)

[6 Comparison of Results 15](#_Toc157803914)

[7 Conclusion 15](#_Toc157803915)

# 1 Introduction

Breast cancer is one of the most common cancers among women and a leading cause of death. The objective is to build a classifier using the KNN (K-nearest neighbors) and J48 (also known as C4.5) machine learning algorithms to predict whether a tumor is cancerous (Malignant labeled as 4) or non-cancerous (Benign labeled as 2). Developing an efficient classifier with a high accuracy rate can be challenging, especially when dealing with missing values—16, in our case. The handling of these missing values is discussed in the data preparation section of this report. Classification and data pre-processing will be performed using Weka 3.8.6.

# 2 Business Understanding

This report focuses on breast cancer data obtained from the University of Wisconsin Hospitals, which is publicly available. The dataset in this report consists of periodic samples collected by Dr. Wolberg in his clinical cases which reflect chronological grouping of data. Features in our data set are computed from digitized image of FNA (fine needle aspirate) of breast mass which was obtained using Multisource Method-Tree (MSM-T) classification method uses linear programming to build the decision tree. Features in the dataset describe the characteristics of the cell nuclei.

Dr. Wolberg periodic samples

|  |  |  |
| --- | --- | --- |
| Group 1: | 367 | January 1989 |
| Group 2: | 70 | October 1989 |
| Group 3: | 31 | February 1990 |
| Group 4: | 17 | April 1990 |
| Group 5: | 48 | August 1990 |
| Group 6: | 49 | January 1991 |
| Group 7: | 31 | June 1991 |
| Group 8: | 86 | November 1991 |
| Total: 699 instances | | |

# 3 Data Understanding

## 3.1 Collect Data

We will be performing two types of classification on the caner dataset KNN (K-nearest neighbors) and Decision Tree (J48 or C4.5). For the KNN algorithm will be using numerical data to predict the class (2 for Benign, 4 for Malignant). We are going to use all nine features to perform KNN and exclude the id attribute which is irrelevant to find the Euclidean distance. For J48 we will be using all 10 attributes except id. Class attribute is included to find entropy and later on calculate information gain for other features.

## 3.2 Describe Data

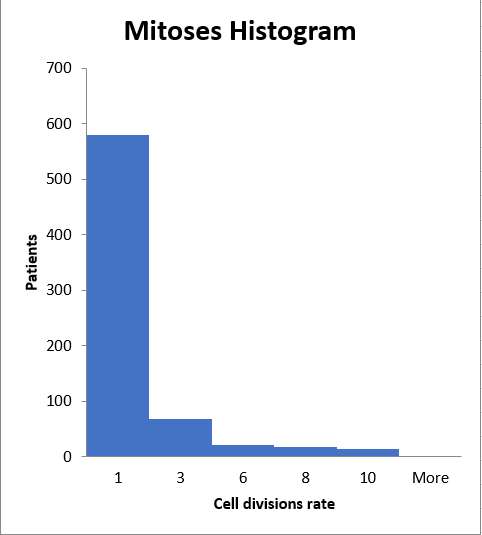
There are total of 11 attributes, nine features, one class attribute and one id attribute. Id attribute is playing any role in our calculations so we will ignore it completely.

|  |  |  |
| --- | --- | --- |
| Attributes | Description | Data Type |
| clump thickness | indicating grouping of cancer cells in multilayer | Numeric |
| Uniformity of Cell Size | indicating metastasis to lymph nodes | Numeric |
| Uniformity of Cell Shape | identifying cancerous cells of varying size | Numeric |
| Marginal Adhesion | suggesting loss of adhesion. a sign of malignancy. Retention of adhesion becomes an indication of malignancy as cancerous cells typically lose this property | Numeric |
| Single Epithelial Cell Size (SECS) | if the SECS become larger, it may be a malignant cell | Numeric |
| Bare Nuclei | without cytoplasm coating, found in benign tumors | Numeric |
| Bland Chromatin | usually found in benign cells | Numeric |
| Normal Nucleoli | generally, very small in benign cells | Numeric |
| Mitoses | the process in cell division by which the nucleus divides | Numeric |
| Class | Benign tumors (non-cancerous 2) malignant tumors (cancerous 4) | Nominal |

## 3.3 Explore Data

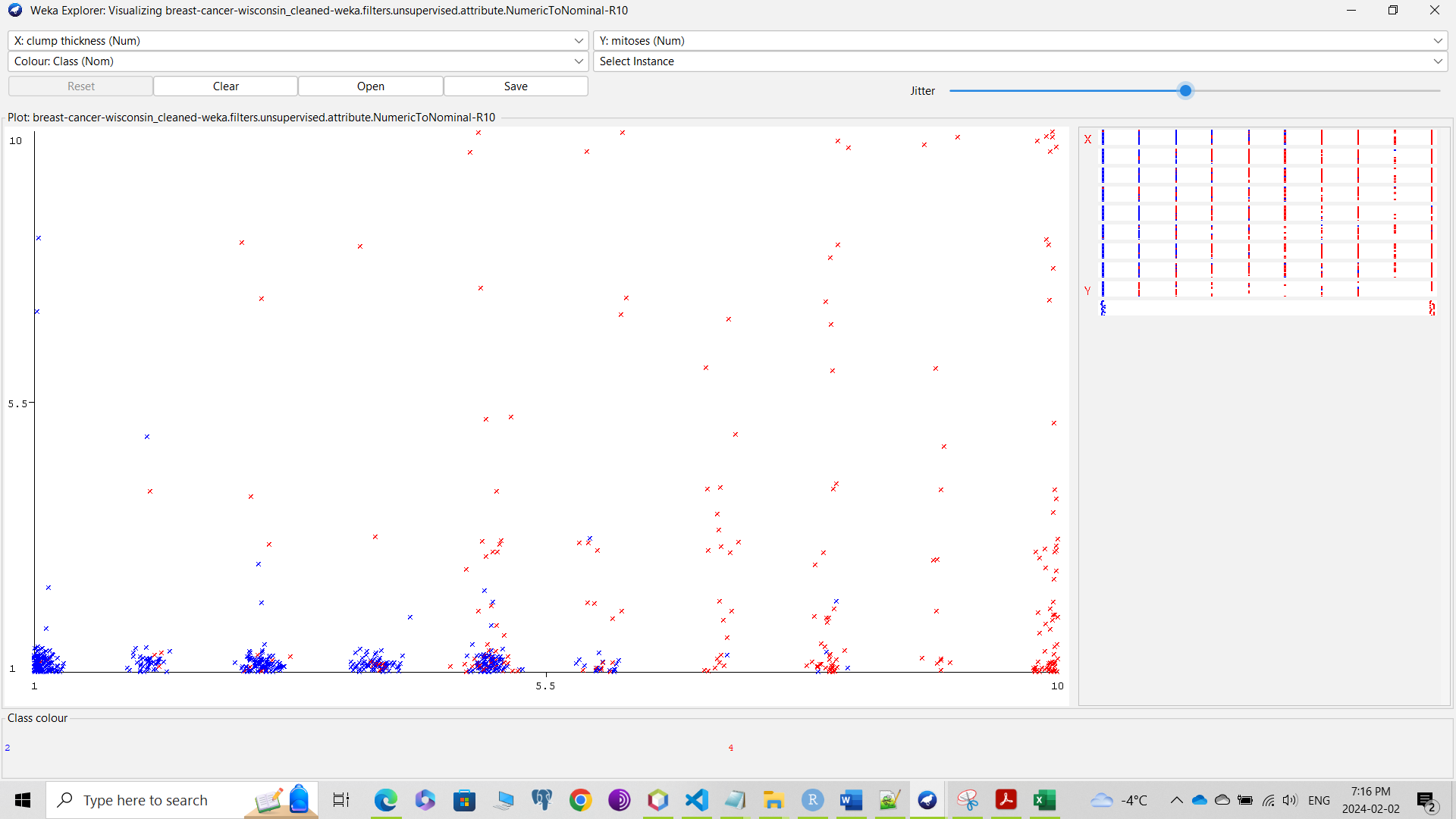
**Histogram of Mitoses:**

Mitoses is the process of cell division, cells become cancerous when they have unusual and high mitosis. The below chart shows cancerous patients with high mitosis.



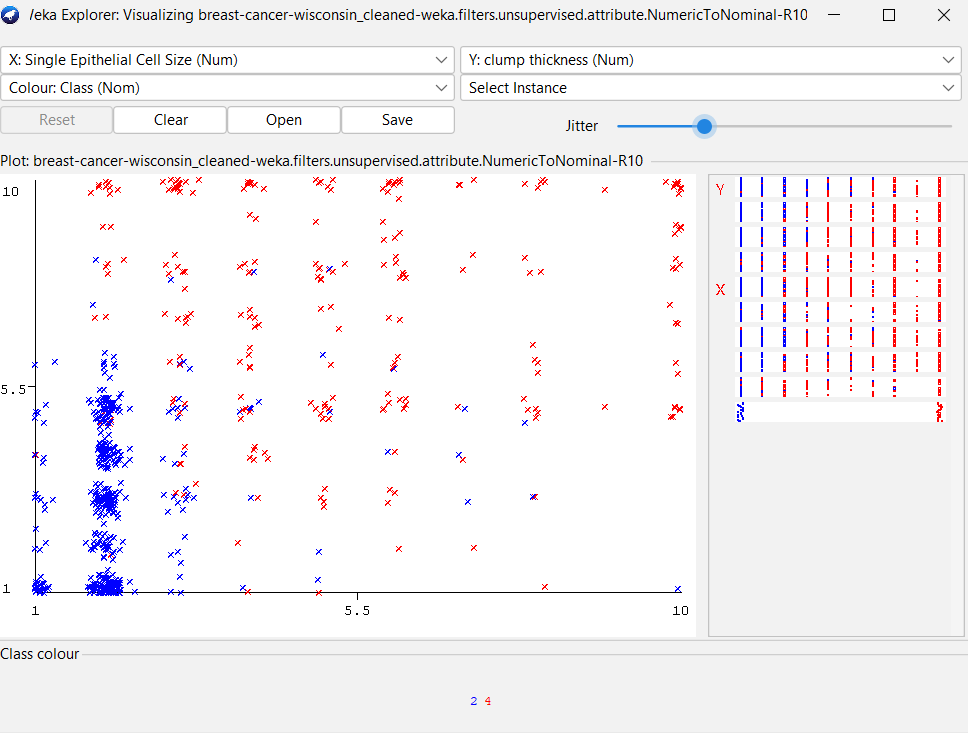
**Clump Thickness (as X) and Mitoses ( as Y):**

Figure below shows the relationship between clump thickness and Mitoses. In the picture we can see that the abnormal clumps have unusual and high mitoses indicating cancer.



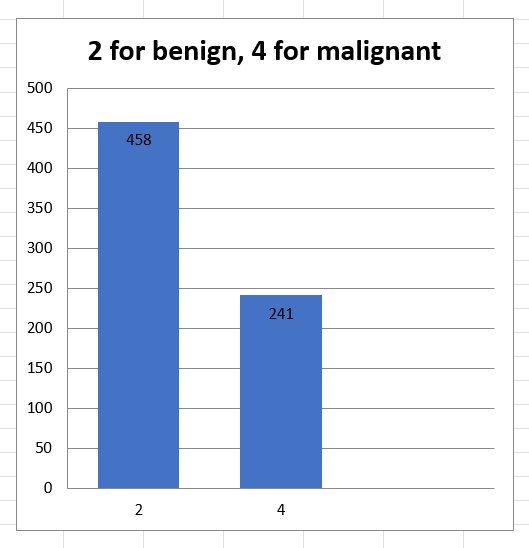
**SECS (as X) vs Clump Thickness (as Y)**

Figure below shows relationship between Single Epithelial Cell Size (SECS) and Clump Thickness. We can see thick clumps have more SECS one of malignant indication.

****

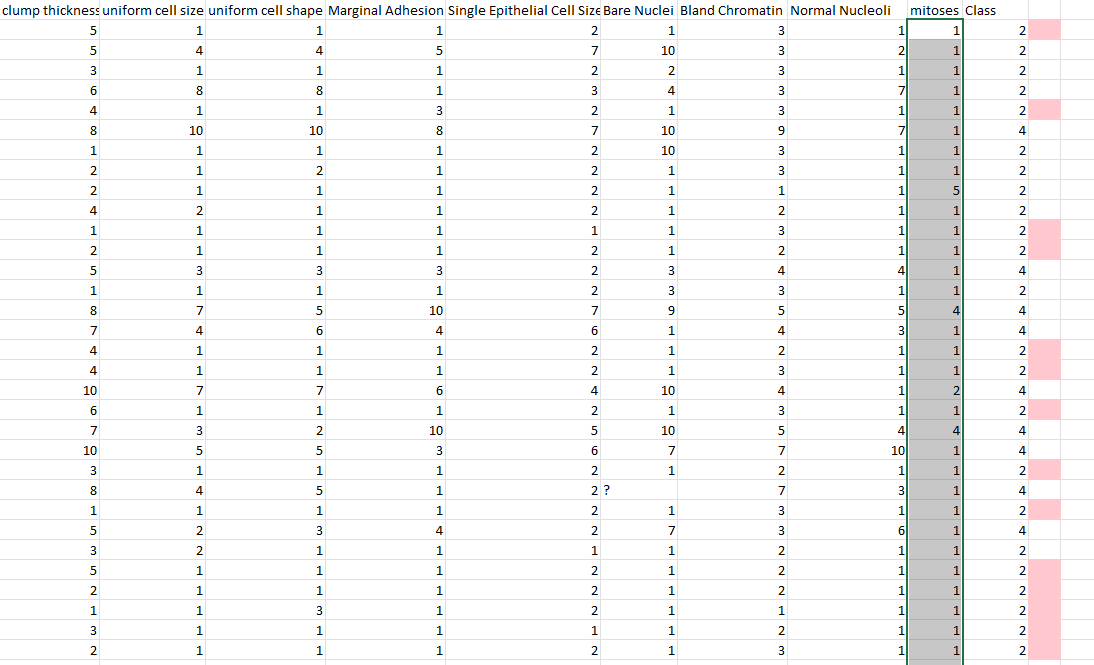
**Class bar plot:**

shows number positive (malignant labeled 4) and negative (benign labeled 2) breast cancer patients. In our case out of 699 patients 241 of them are positive and 458 of them are negative.



## 3.4 Verify Data Quality

* Missing Data: Feature Bare Nuclei contains missing data. Which is going to be an issue for KNN classification. There are serval strategies for handling numeric missing data. Which will be explained in the Data Preparation section.
* Duplicate Instances: Duplicate instances can affect classification and cause issues such as overfitting which model performs well on training set but poorly on test set or underfitting which perform poorly on both training set and testing set. We can remove duplicates in Weka using RemoveDuplicates filter. The Following figure show sample duplicate instances highlighted in light red.



* Invalid data: no invalid data found in any feature.

# 4 Data Preparation

## 4.1 Select Data

We are going to select features with missing values, or invalid values. Which in our case only Bare Nuclei feature has missing values. We are also going to normalize numeric value features for better result and to make it more compatible with K-NN classification. Duplicates instances also need to be removed to prevent an overfitting or underfitting models.

## 4.2 Clean Data

* In data cleaning, it is crucial to identify features containing missing, invalid, or incomplete values. In our dataset, most features consist of numerical discrete values, and fortunately, there are no incomplete or invalid entries. However, the "Bare Nuclei" attribute has 16 missing values. Various strategies exist to handle missing values, such as excluding instances or filling them with the mean, median, or mode. None of these strategies are deemed efficient for our dataset. Removing duplicate instances before classification is necessary, but eliminating 16 instances with missing values would significantly impact the classification algorithm. Filling missing values with the mode, mean, or median is also unsuitable, as it may lead to overfitting or underfitting the model. To address this issue, we will explore the values of other attributes, particularly Mitoses, Class, Marginal Adhesion, and Single Epithelial Cell Size (SECS). These features are directly interconnected and can aid in filling the missing values for "Bare Nuclei."

## 4.3 Construct Data

* To address missing values in the "Bare Nuclei" feature, my initial approach involves considering the class values. Patients with malignant tumors tend to exhibit higher values for the "Bare Nuclei" feature. Another critical factor considered is the examination of values for features such as Clump Thickness, Uniformity of Cell Shape, Mitoses, Single Epithelial Cell Size, and Marginal Adhesion. Cancerous tumors typically display elevated values in these features. Consequently, if the values for these features are high, I will assign missing "Bare Nuclei" values within the range of 6 to 10. Conversely, if the majority of values for these features are low, I will assign values between 1 to 5. For example: if the value of class is 2 and the value of Mitoses is 1, value of Single Epithelial is 2, value of Marginal Adhesion is 4 and class is 2, I am going to assign Bare Nuclei a value of 3. After filling the missing data we need to remove duplicates, which can be done in Weka through *RemoveDuplicates* filter.

## 4.4 Integrate Data

* Missing values in *Bare Nuclei* is filled based on the values of adjacent features such as Mitoses, SECS, Marginal Adhesion and class. All the 16 missing values of Bare Nuclei attribute are filled, and the data set is ready for classification.

## 4.5 Format Data

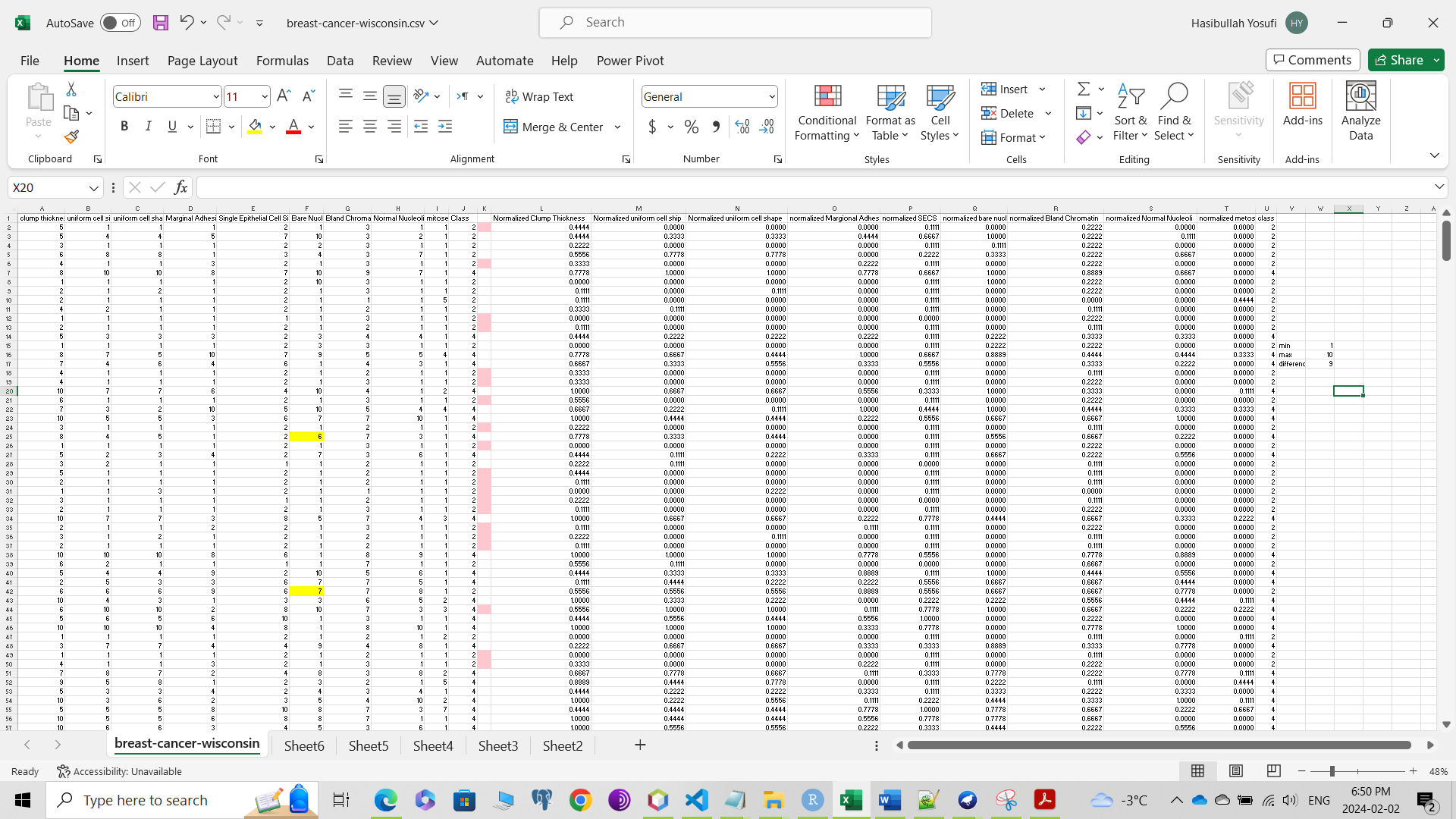
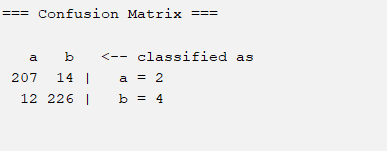
* We can perform J48 and K-NN classification algorithm on the data set. However, features might have low cardinality, it can cause algorithm such as K-NN to not perform very well, therefore I use both normalized and original data set for the classification. Figure below shows normalized value and not-normalized ones side by side.

Figure 1 Original data set and normailzed veriosn of it

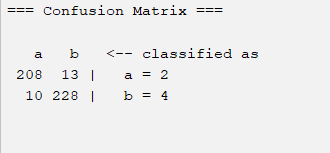
# 5 Modeling

## 5.1 K-NN

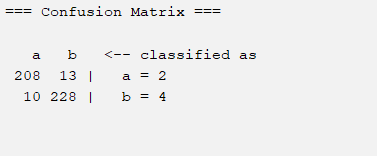
K = 3



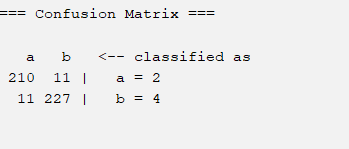
K = 5



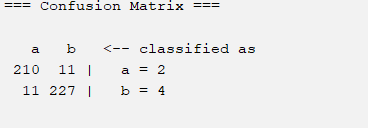
K = 7



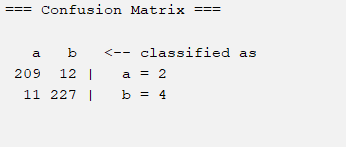
K = 9



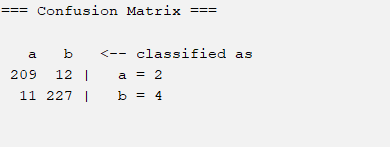
K = 11



K = 13



K = 15



For the K-NN classification with K (number of neighbors) equal to 9 gives the highest percentage of correctly classified instances. With 95.207% accuracy and with relative absolute accuracy error of 15.7238%.

## 5.2 Decision Tree (J48 || C4.5)

The highest accuracy achieved with the J48 algorithm is 92.3747 percent. This accuracy can be attained by setting the seed value to 1, the minNumObj property to 22, and the unpruned property to true. The figure below depicts the uniform cell size as the root node with the highest information gain. The node undergoes further splitting; if the uniform cell size is greater than 3 (right branch), it indicates a malignant tumor (4). If the value is less than or equal to three, the algorithm examines the Bare Nuclei node. If the Bare Nuclei value is less than or equal to two, it is identified as a benign tumor (2). If the value of Bare Nuclei is greater than two, the algorithm then considers the clump thickness feature. If the value of clump thickness is less than or equal to three, it is labeled as a benign tumor (2); otherwise, it is classified as a malignant tumor (4).

A network of lines and dots

Description automatically generated with medium confidence

# 6 Comparison of Results

J48 and K-NN comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Sensitivity | Specificity | F1 measure |
| K-NN  K = 9 | 95.207 | 0.95 | 0.95 | 0.954 | 0.95 |
| J48  minNumObj = 22  unpruned = True | 92.3747 | 0.931 | 0.91 | 0.94 | 0.920 |

# 7 Conclusion

In this report, we have followed the CRISP-DM (Cross Industry Standard Process - Data Mining) methodology to execute two machine learning classification algorithms (J48 and KNN) for classifying the Wisconsin breast cancer dataset. The K-Nearest Neighbor algorithm demonstrated the highest accuracy at 95.2%. It is crucial to preprocess (clean) the data before applying the algorithm.

In our Wisconsin breast cancer dataset, there are nine features with one class attribute. The "Bare Nuclei" feature was the only attribute that contained 16 missing values. To address these missing values, I examined its adjacent attributes and compared all of them with the class. It was determined that Malignant (4) tumors exhibit elevated values for clump thickness, Mitoses, and SECS, as well as the same pattern for other features. Based on this analysis, missing values were filled, and the dataset was modeled using both K-NN and J48 algorithms.