Universidad Autónoma de San Luis Potosí

Facultad de Ingeniería

Centro de Investigación y Estudios de Posgrado

Determinación de Factores de Riesgo Relevantes para la Predicción del Cáncer de Mama usando Selección de Características

TESIS

Que para obtener el grado de:

Maestría en Ingeniería en Computación

Presenta:

Ing. Zazil Josefina Ibarra Cuevas

Asesor:

Dr. Alberto Salvador Núñez Varela

Co-asesor:

Dr. José Ignacio Núñez Varela

San Luis Potosí, S.L.P.

Septiembre, 2022

Universidad Autónoma de San Luis Potosí

Facultad de Ingeniería

Centro de Investigación y Estudios de Posgrado

Determination of Relevant Risk Factors for Breast Cancer Prediction using Feature Selection

TESIS

Que para obtener el grado de:

Maestría en Ingeniería en Computación

Presenta:

Ing. Zazil Josefina Ibarra Cuevas

Asesor:

Dr. Alberto Salvador Núñez Varela

Co-asesor:

Dr. José Ignacio Núñez Varela

San Luis Potosí, S.L.P.

Septiembre, 2022

UASLP
Universidad Autónoma
de San Luis Potosí

FACULTAD DE
INGENIERÍA

19 de mayo de 2022

ING. ZAZIL JOSEFINA IBARRA CUEVAS
PRESENTE.

En atención a su solicitud de Temario, presentada por los **Dres. Alberto Salvador Núñez Varela y José Ignacio Núñez Varela,** Asesor y Coasesor de la Tesis que desarrollará Usted, con el objeto de obtener el Grado de **Maestra en Ingeniería de la Computación**, me es grato comunicarle que en la sesión del H. Consejo Técnico Consultivo celebrada el día 19 de mayo del presente, fue aprobado el Temario propuesto:

TEMARIO:

"Determinación de Factores de Riesgo Relevantes para la Predicción del Cáncer de Mama usando Selección de Características"

1. Introducción.
2. Antecedentes del cáncer de mama y minería de datos.
3. Conjunto de datos y preprocesamiento.
4. Selección y validación de factores de riesgo.
5. Conclusiones.

Referencias.

"MODOS ET CUNCTARUM RERUM MENSURAS AUDEBO"

ATENTAMENTE

DR. EMILIO JORGE GONZÁLEZ GALVÁN
DIRECTOR.

UNIVERSIDAD AUTONOMA
DE SAN LUIS POTOSI
FACULTAD DE INGENIERIA
DIRECCION

Copia. Archivo.
*etn.

"Rumbo al centenario de la autonomía universitaria"

Resumen

El cáncer de mama es una grave amenaza para la salud, ya que representa el tipo de cáncer más común y expandido entre las mujeres de todo el mundo. Aunque todavía se desconocen las causas exactas de esta enfermedad, estudios han identificado factores de riesgo asociados al padecimiento. Los factores de riesgo son cualquier condición genética, reproductiva, hormonal, física, biológica o de estilo de vida que aumente la probabilidad de desarrollar cáncer de mama. A lo largo de los años, el estudio de factores de riesgo de cáncer de mama, ha ayudado a tener un mayor entendimiento de la enfermedad y a crear estrategias preventivas y de control de riesgo. Esta investigación tiene como objetivo identificar los factores de riesgo más relevantes en pacientes con cáncer de mama en un conjunto de datos siguiendo el proceso de *Descubrimiento de Conocimiento en Bases de Datos* (*Knowledge Discovery in Databases*), el cuál, hace uso de algoritmos computacionales para extraer información potencialmente útil y valiosa de los datos. Para determinar la relevancia de los factores de riesgo, esta investigación implementa dos métodos de selección de características: la prueba de la *Ji al cuadrado* y el cálculo de la *Información mutua.* Además, se utilizaron siete algoritmos de clasificación para validar los resultados obtenidos. Nuestros resultados muestran que los factores de riesgo identificados como los más relevantes están relacionados con la edad de la paciente, su estatus y tipo de menopausia, y si se ha sometido a terapia hormonal.

Abstract

Breast cancer is a serious health threat, since it represents the most common and widespread type of cancer among women around the world. Although the exact causes of this disease are still unknown, studies have identified risk factors associated with the condition. Risk factors are any genetic, reproductive, hormonal, physical, biological, or lifestyle-related conditions that increase the likelihood of developing breast cancer. Over the years, the study of breast cancer risk factors has helped to have a better understanding of the disease and to create preventive and risk control strategies. This research aims to identify the most relevant risk factors in patients with breast cancer in a dataset by following the *Knowledge Discovery in Databases process*, which makes use of computational algorithms to extract potentially useful and valuable information from the data. To determine the relevance of risk factors, this research implements two feature selection methods: the *Chi-squared test* and *Mutual information.* Also, seven classification algorithms are used to validate the results obtained. Our results show that the risk factors identified as the most relevant are related to the age of the patient, her menopausal status, whether she had undergone hormonal therapy, and her type of menopause.

A mi familia, por ser una luz en mi camino.

Agradecimientos

En estas líneas quiero agradecer a todas las personas que hicieron posible esta investigación. Estas palabras son para ustedes.

Agradezco a Daniel Torres por brindarme apoyo incondicional semestre a semestre y por nunca dudar que lo lograría. Con tu apoyo, amor y paciencia cualquier cosa parece más fácil.

A mis padres, Angelica Cuevas y Martín Ibarra por su trabajo y sacrificio en todos estos años. Agradezco por los consejos, valores y principios que me han inculcado y por darme la libertad de desenvolverme como ser humano. A mi hermana, Dennise Ibarra, por todas las pláticas y risas nocturnas, tu manera de escucharme es irremplazable.

A mis amigos, en especial a Alejandro Álvarez, Griselda Espinosa y Yazmin Medina por acompañarme a crecer a lo largo de los años y por extender su mano en momentos de duda e incertidumbre que se presentaron durante mis estudios. Gracias por reafirmar que aunque a veces estemos lejos nuestros corazones siguen siendo los mismos.

A a mi asesor de tesis el Dr. Alberto Salvador Núñez Varela y a mi co-asesor el Dr. José Ignacio Núñez Varela, que con su experiencia, conocimiento y enseñanza me orientaron durante la realización de esta investigación. Terminar este trabajo no hubiera sido posible sin su paciencia y apoyo profesional.

A todos los profesores que me impartieron alguna materia, gracias por su dedicación al compartir sus conocimientos a lo largo de mi preparación. En especial, al Dr. Francisco Martínez por orientarme para ingresar a la maestría e invitarme a participar en este proyecto de investigación. A la Dra. Sandra Nava por brindarme siempre una retroalimentación muy enriquecedora. Al Dr. Héctor Pérez y al Dr. César Puente, miembros de mi comité de tesis, por sus evaluaciones a lo largo del proceso. Finalmente, al Mtro. César Ramírez por todos los consejos que me brindó a lo largo de mis estudios.

Agradezco al grupo de médicos conformado por la Dra. Virginia Canseco, el Dr. Francisco de la Rosa, el Dr. Raúl Martínez y el Dr. Alejandro Hernández que se tomaron el tiempo de compartir sus conocimientos en el área de oncología y radiología, su colaboración fue muy valiosa para este trabajo.

Finalmente, agradezco a la Universidad Autónoma de San Luis Potosí, por abrirme las puertas para crecer académicamente y tener una educación de calidad.

Índice general

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Globally, breast cancer is the most common and widespread type of cancer among women with more than 2.2 million new cases and about 680,000 deaths in 2020, according to the *Global Cancer Observatory* [1]. Researchers have studied the origin, causes, and ways to reduce the impact of this disease on society. Despite the progress that has been made, the specific causes of breast cancer incidence are still difficult to determine [2]. The early detection of breast cancer is key for increasing the chance of treatment and recovery; this is normally done by screening tests, such as a mammography. Studies have also identified what are known as *risk factors*, that are associated with the likelihood of developing breast cancer. There are a wide variety of risk factors that include genetic, reproductive, hormonal, physical, biological, and lifestyle-related, among others [2], [3].

It is important to analyze and understand the possible impact each factor can have on the development of breast cancer so that physicians could suggest preventive strategies to women who are known to have some of these risk factors. A common trend in recent years is the creation of preventive strategies from the analysis of data obtained from clinical records [4]. This has been achieved by using methodologies or processes that focus on extracting potentially useful and valuable information through computational tools. *Knowledge Discovery in Databases (KDD)* [5] is a process that follows different stages (as shown in Figure 1.1 based on [5]) with the aim of identifying new *knowledge* from datasets.

All the steps are essential for the successful application of *KDD* in practice, however,

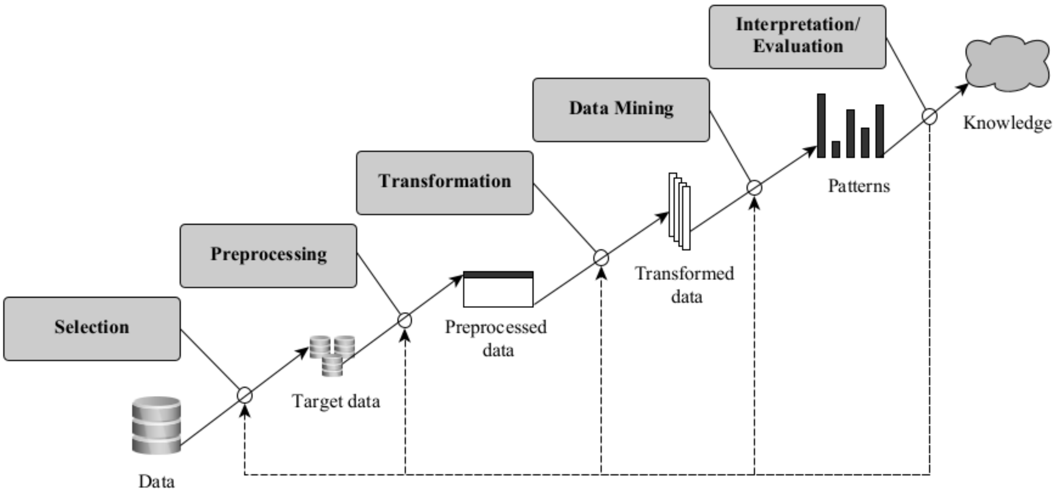


Figure 1.1: The process of Knowledge Discovery in Databases (*KDD*) (based on [5]).

the method used in the *Data Mining* step is very important, since it determines the overall discovery goal. This research uses the method called *dependency modeling*, which consists of finding a model that describes significant dependencies between variables. We are interested in determining the dependency or correlation between each of the risk factors and the variable that determines whether the person has breast cancer or not. In order to determine which risk factors are more relevant in the development of breast cancer, *Feature Selection* was used in this research. *Feature selection* is mainly focused on removing non-informative or redundant predictors from a model by methods involving the evaluation of the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest correlation with the target variable [6]. For the *feature selection* process, *Chi-squared* and *Mutual information* were used. Our results show that the risk factors identified as the most relevant are related to the age of the patient, her menopausal status, whether she had undergone hormonal therapy, and her type of menopause. Subsequently, a validation stage was implemented to evaluate the performance obtained by training seven classification algorithms: *Decision Tree*, *Random Tree*, *Decision Stump*, *Deep Learning*, *K-Nearest Neighbors (K-NN)*, *Naïve Bayes* and *Generalized Linear Model.* The purpose of the validation stage was to obtain and compare the training performance results with all the risk factors of the dataset and with the four most relevant risk factors resulting from the *feature selection* stage. Our results show that in most cases a minimum loss of performance occurs when training with the four most relevant risk factors compared to training with all the risk factors of the dataset.

1.1 Motivation

There is a large number of research works that are related to the prediction or likelihood of breast cancer. However, most of them are dedicated in identifying the risk of breast cancer by analysing mammograms, few studies are dedicated to the study of risk factors for breast cancer. In contrast to information from imaging studies, the analysis of risk factors offers a different opportunity to research breast cancer, since no specialized medical equipment is required to obtain the majority of information on risk factors (excluding information on genetic risk factors). With the study of the relevance of risk factors, it could be determined whether there is a risk of developing breast cancer or not, solely from information readily known to most people, is an important option that could be widely available without the need to have specialized equipment. Of course, this is not meant to substitute screening tests and the knowledge of medical personnel, on the contrary, these studies could provide useful information and be part of the development of breast cancer risk control strategies.

When studying risk factors it should be taken into account that it is an area that represents great challenges, for example, there are not many publicly available datasets, the imbalance of the data, the size of datasets, the unknown information in medical records, etc. Several works have proposed a variety of alternatives to solve these challenges through different computational techniques. Unlike other works, our solution integrates feature selection methods and ensamble learning algorithms to determine and validate the most relevant breast cancer risk factors of a dataset.

1.2 Research Questions

Generating guidelines for the development of control strategies focused on risk factors for breast cancer leads to an important research question addressed in this thesis:

- What are the most relevant risk factors for the determination of breast cancer in patients that already have such cancer, using the Knowledge Discovery in Databases process?

1.3 Research Goals

1.3.1 General goal

Identify and validate the most relevant risk factors of breast cancer, through the application of feature selection methods for their identification, and classification algorithms for their validation.

1.3.2 Specific goals

1. Understand the problem by reviewing the state of the art, and having meetings with radiologists and oncologists.

2. Analyze publicly available datasets of breast cancer risk factors and select the one that could suit our needs.

3. Prepare and clean the selected dataset to ensure data quality.

4. Analyze and apply feature selection methods to determine the most relevant risk factors in the selected dataset.

5. Validate the relevant risk factors by means of classification algorithms.

6. Interpret the results by comparing the results obtained with and without the feature selection process.

1.4 Research Methodology

To organize all the activities related to this research, the methodology of Figure 1.2 was created and followed. As observed, it consists of four main phases summarized next.

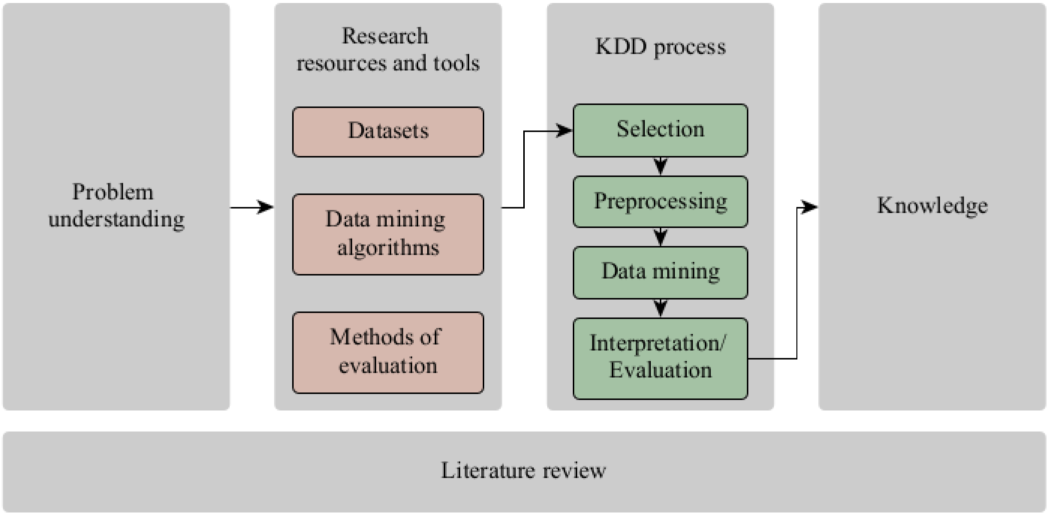


Figure 1.2: Research methodology.

Problem understanding: It consists of activities that allow us to know the problem and define the objectives of our research, both in the medical and computational fields.

Research resources and tools: The goal of this phase, as the name implies, is to gather the necessary resources and tools for the next phase where the *KDD* process is applied. Resources represent risk factors datasets and tools are the methods to be used for the data mining and interpretation/evaluation stages of the *KDD* process.

***KDD* Process:** This phase consists in applying each of the following phases of the *KDD* process:

1. *Selection*: Selecting a dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed.
2. *Preprocessing*: Perform data cleaning to ensure data quality.
3. *Data mining*: Searching for patterns of interest in a particular representational form.
4. *Interpretation/Evaluation*: This step can also involve visualization of the extracted patterns/models of the data (possibly return to previous phases).

The *KDD* process can involve significant iteration and may contain loops between those steps.

Knowledge: It is the final stage where the results obtained from the *KDD* process are concluded.

Throughout all the research there is a constant literature review.

1.5 Thesis Contribution

According to the research goals and methodology described above, the following is the contribution from this thesis:

- Determination and validation of the most relevant risk factors for breast cancer in a dataset through the integration of feature selection methods and ensemble learning algorithms.

1.6 Thesis Outline

The rest of this thesis is structured according to the following summary of chapters:

Chapter 2: This chapter presents an overview of the medical and computational background. A definition of breast cancer is presented and the problem that risk factors represent is described. Also, the computational techniques used in this thesis to address the problem are defined. The set of solution methods proposed in related works are also described and discussed.

Chapter 3: This chapter describes the activities developed for the selection and preprocessing stages of the *KDD* process. In the selection stage, the dataset selection criteria and the dataset options that were identified are described. In the preprocessing stage, a follow-up is given through the operations applied to the original dataset selected in order to give a format to data for the data mining stage. As a result, it provides a general description of the attributes and records of the final dataset.

Chapter 4: This chapter contains the description of data mining and interpretation/ evaluation stages of the *KDD* process. For the data mining stage, two feature selection methods (Chi-squared and Mutual information) are used to obtain the relevant values for each of the risk factors. In the interpretation/evaluation stage, the results obtained from the data mining stage are tested using classification algorithms to subsequently perform an analysis for the interpretation of the results. Finally, a comparison with other works is made.

Conclusions: This chapter summarizes the main findings obtained, addresses the strengths and limitations of the study and proposes areas of future research.

Chapter 2

Background on Breast Cancer and Data Mining

This chapter provides a description of the medical and computational background. First, breast cancer and the concept of risk factors are described. Next, the *Knowledge Discovery in Databases (KDD)* process is explained as well the computational methods and techniques used in this research. Finally, the related work of the solution methods proposed to find the relationship between risk factors and breast cancer is presented.

2.1 Breast Cancer

Breast cancer is a disease of the mammary gland that originates when breast cells begin to grow uncontrollably. These cells that divide faster than healthy cells usually build up into a lump or tumor [7], which can be detected by a physical exam or imaging tests, such as a mammogram, ultrasound, or magnetic resonance. Breast cancer is one of the world's largest health problems. This type of cancer can occur in both men and women, but it is much more common in women. In fact, it is the most diagnosed cancer in women and ranks first with the highest number of deaths for the female gender [1]. In 2020, there were 2.2 million women diagnosed with breast cancer and 685 000 deaths globally [1]. By the end of the same year, 7.8 million women who had been diagnosed with breast cancer in the previous five years were still alive, making this cancer the most prevalent in the world

[8]. Survival rates for breast cancer have increased in the last few years, and the number of deaths associated with this disease continues to decline, mostly due to factors such as early detection, a new personalized approach to treatment, and a better understanding of the disease. Finding breast cancer early and getting state-of-the-art cancer treatment are two of the most important strategies for preventing deaths from breast cancer [9]. Breast cancer that is found early, when it is small and has not yet spread, it is easier to treat successfully. Getting regular screening tests is the most reliable way to find breast cancer early. Screening refers to tests and exams used to find a disease in people who do not have any symptoms. The goal of screening tests for breast cancer is to find it early, before it causes symptoms (like a lump in the breast that can be felt). Different screening tests can be used to look for and diagnose breast cancer, for example, mammograms, breast ultrasound and breast magnetic resonance imaging. One of the most used screening test is mammograms, the results of this test uses the scale called *BI-RADS* (*Breast Imaging Reporting and Data System*), which is used in this thesis in later sections and chapters, for this reason, is explained below.

2.1.1 Mammograms

A screening mammogram is used to look for signs of breast cancer in women who do not have any breast symptoms or problems. X-ray pictures are taken by a radiologist, who categorizes the mammogram results using a numbered system [10]. This system is called the *Breast Imaging Reporting and Data System* (*BI-RADS*) sorts the results into categories numbered 0 through 6 [11]:

- 0: Incomplete, additional imaging evaluation and/or comparison to prior mammograms (or other imaging tests) is needed. This means the radiologist may have seen a possible abnormality, but it was not clear and it is necessary more tests, such as another mammogram with the use of spot compression (applying compression to a smaller area when doing the mammogram), magnified views, special mammogram views, or ultrasound.
- 1: Negative. This is a normal test result.

- 2: Benign (non-cancerous) finding. This is also a negative test result (there is no sign of cancer), but the radiologist chooses to describe a finding that is not cancer, such as benign calcifications, masses, or lymph nodes in the breast.

- 3: Probably benign finding. A finding in this category has a very low (no more than 2%) chance of being cancer.

- 4: Suspicious abnormality, biopsy should be considered. The findings in this category can have a wide range of suspicion levels. For this reason, this category is often divided further:

  - 4A: Finding with a low likelihood of being cancer (more than 2% but no more than 10%).

  - 4B: Finding with a moderate likelihood of being cancer (more than 10% but no more than 50%).

  - 4C: Finding with a high likelihood of being cancer (more than 50% but less than 95%), but not as high as Category 5.

- 5: Highly suggestive of malignancy. The findings look like cancer and have a high chance (at least 95%) of being cancer.

- 6: Known biopsy-proven malignancy. This category is only used for findings on a mammogram (or ultrasound or MRI) that have already been shown to be cancer by a previous biopsy.

With these categories, radiologists can describe what they find on a mammogram using the same words and terms. A mammogram report also include an assessment of the breast density, which is a description of how much fibrous and glandular tissue is in the breasts, as compared to fatty tissue. There are four categories of breast density. They go from almost all fatty tissue to extremely dense tissue with very little fat. The radiologist decides which of the four categories best describes how dense the breasts are:

- Category A: Breasts are almost all fatty tissue.

- Category B: There are scattered areas of dense glandular and fibrous tissue.
- Category C: More of the breast is made of dense glandular and fibrous tissue (described as heterogeneously dense). This can make it hard to see small masses in or around the dense tissue, which also appear as white areas.
- Category D: Breasts are extremely dense, which makes it harder to see masses or other findings that may appear as white areas on the mammogram.

As part of achieving early detection of breast cancer, efforts have been made to identify the causes of the disease, however, these remain unknown. Nevertheless, studies have identified some risk factors that increase the likelihood of developing breast cancer [2], [3], which are described in the next subsection.

2.1.2 Risk factors

A risk factor for breast cancer is anything that could make the disease to be more likely to occur. The literature indicates that the most important risk factors for breast cancer are advanced age and the female gender (men may also develop breast cancer, however it represents only 1% of all cases) [3]. However, there is a long list of factors related to the increased risk of developing breast cancer [2], [3], [12]–[16], which includes:

Family history of cancer: Women with a first-degree relative (mother, daughter, or sister) with breast cancer have approximately double the risk of the general population and are at particularly high risk if the cancer was premenopausal or bilateral[1].

Genetic factors: Certain mutations in the genes that increase the risk of breast cancer can be inherited. The best-known mutations are BRCA1 and BRCA2. These genes can greatly increase the risk of breast cancer and other cancers, however, they do not make the disease inevitable.

Reproductive factors: Factors such as early menarche (before age 12), late menopause (after age 55), nulliparity (when the person has never been pregnant), and first live

[1] Bilateral breast cancer occurs when cancer occurs in both breasts at the same time.

birth after age 30 bestow a slightly higher risk for breast cancer, as a result of having more menstrual cycles and longer exposure to estrogen and progesterone.

Hormonal factors: For example use of oral contraceptive or use of hormone replacement therapy that combines estrogen and progesterone to treat the signs and symptoms of menopause.

Exposure to the radiation: If radiation treatments have been done in the thorax area during childhood, the risk of breast cancer increases.

Demographic factors: Including country of origin, year of birth, specification of ethnicity and family race (Asian, Black, Hispanic, Native American and White). Women of Ashkenazi (Eastern European) Jewish heritage have a slightly higher risk of breast cancer than does the general population.

Personal history of breast diseases: A breast biopsy showing atypical ductal hyperplasia (ADH) histology increases the risk for breast cancer to four to five times that of the general population. The presence of lobular carcinoma in situ (LCIS) also increases the risk for breast cancer, but at a much higher rate than ADH (about 10 times that of the normal population). The acronym LCIS is a misnomer and not a cancer at all; rather, LCIS is a high-risk marker for developing breast cancer.

Personal history of breast cancer: If cancer has developed in one breast, there is an increased risk of developing cancer in the other.

Mammographic breast density: A large amount of fibroglandular tissue within the breast measured on the mammogram is associated with the risk of breast cancer.

Lifestyle factors: One of these is drinking alcohol. One drink per day bestows a very small risk, but two to five drinks per day increases the risk to 15 times that of women who do not drink. Being overweight or obese also increases the risk of cancer, especially if the weight gain happens after menopause and the fat is around the abdomen.

Since there are many factors that could contribute to the occurrence of breast cancer, it is very difficult to identify the exact combination of elements that cause the disease. Nevertheless, based on several researches [12]–[16] aimed on determining the relationship between risk factors and breast cancer, it has been possible to contribute to the generation of risk reduction and control strategies, including:

- Making changes in habits and lifestyle: Exercising 30 minutes per day, maintaining a healthy weight, if alcohol is consumed, doing so in moderation and limiting the amount of consumption per day, opting for a healthy diet.
- Limiting the use of hormone replacement therapy.

In cases where the risk is high, more specialized strategies have been developed:

- Preventive medications (chemoprevention): Estrogen-blocking medications, such as selective estrogen receptor modulators and aromatase inhibitors.
- Preventive surgery: Women with a high risk of breast cancer may choose to have their healthy breasts surgically removed (prophylactic mastectomy). They may also choose to have their healthy ovaries removed (prophylactic oophorectomy) to reduce the risk of both breast cancer and ovarian cancer.

Due to the challenge it represents and the importance it has to generate risk reduction and control strategies, several computational alternatives have been proposed to study the correlation between risk factors and breast cancer [4], [17], [18]. Most of these solution alternatives use a set of techniques capable of processing large dataset, identifying patterns and relationships between variables to provide useful information in the medical field. The following section describes the *Knowledge Discovery in Databases (KDD)* process used in this research to extract valuable information from the relationship between risk factors and breast cancer.

2.2 Knowledge Discovery in Databases

Fayyad, Piatetsky-Shapiro, and Smyth define *Knowledge Discovery in Databases* as *"The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"* [5]. According to the authors, the *data* is a set of facts (e.g., cases in a database) and *pattern* is an expression in some language describing a subset of the data or a model applicable to that subset. Extracting a *pattern* also designates fitting a model to data, finding structure from data, or in general, any high-level description of a set of data. The discovered patterns should be *valid* on new data with some degree of certainty. The patterns have to be *novel* and potentially useful. Finally, the patterns should be *understandable*, if not immediately, then after some post-processing. The term *process* implies that KDD is comprised of many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations. This process is characterized by being *nontrivial*, meaning that it is not a straightforward computation, but involves a more complex search or inference. Below is a detailed description of each of the steps in the *KDD* process.

2.2.1 KDD Process

Is the process of using the database along with any required selection, preprocessing, subsampling, and transformations of it; to apply data mining methods (algorithms) to enumerate patterns from it; and to evaluate the products of data mining to identify the subset of the enumerated patterns deemed as "knowledge" (as shown in Figure 2.1 based on [5]). The basic steps of the process are the following:

Understanding: Developing an understanding of the application domain and the relevant prior knowledge, to identify the goal of the *KDD* process.

Selection: Choosing a dataset or focusing on a subset of variables or data samples, on which discovery is to be performed.

Preprocessing: Basic operations such as the removal of noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for

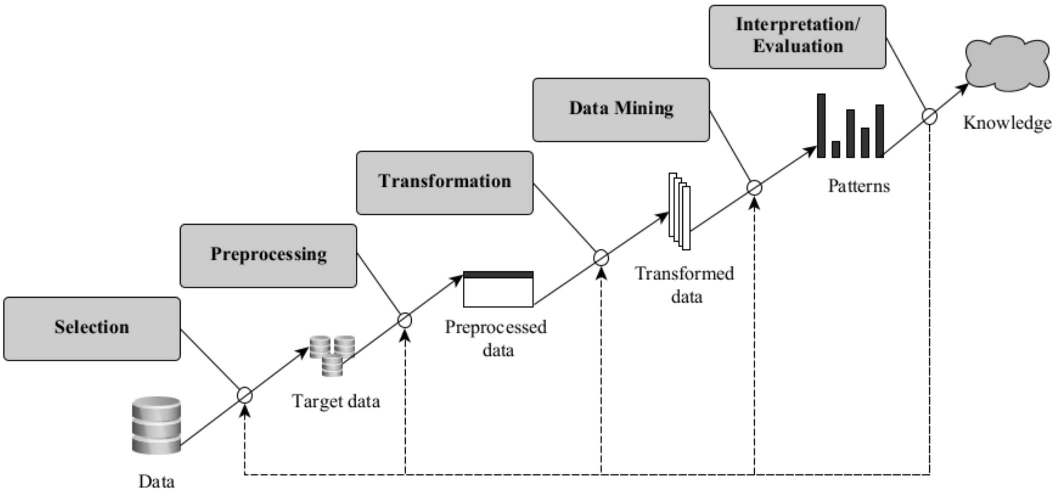


Figure 2.1: The process of Knowledge Discovery in Databases (*KDD*) (based on [5]).

handling missing data fields, accounting for time sequence information and known changes.

Transformation: Finding useful features to represent the data depending on the goal of the task. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

Data mining: Searching for patterns of interest in a particular representational form or a set of such representations: classification rules or trees, regression, clustering, and so forth. The user can significantly aid the data mining method by correctly performing the preceding steps.

Interpretation/Evaluation: This step can involve visualization of the extracted patterns/models, or visualization of the data given the extracted models (possibly return to previous steps).

Consolidating discovered knowledge: Incorporating this knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

The *KDD* process can involve significant iteration and may contain loops between any of the steps. The basic flow of steps (although not the potential multitude of iterations and loops) is illustrated in Figure 2.1. All the steps are essential for the successful application of *KDD* in practice, however, the method used in the *Data Mining* step is very important, since it determines the overall discovery goal. The knowledge discovery goals are defined by the intended use of the system [5] and are described next.

2.2.2 Data extraction methods

Two types of primary goals can be distinguished (see Figure 2.2 based on [5]): *verification*, where the system is limited to verifying the user's hypothesis, and *discovery*, where the system autonomously finds new patterns. The *discovery* goal can be subdivided into *prediction*, where the system finds patterns for the purpose of predicting the future behaviour of some entities; and *description*, where the system finds patterns for the purpose of presenting them to a user in a human-understandable form.

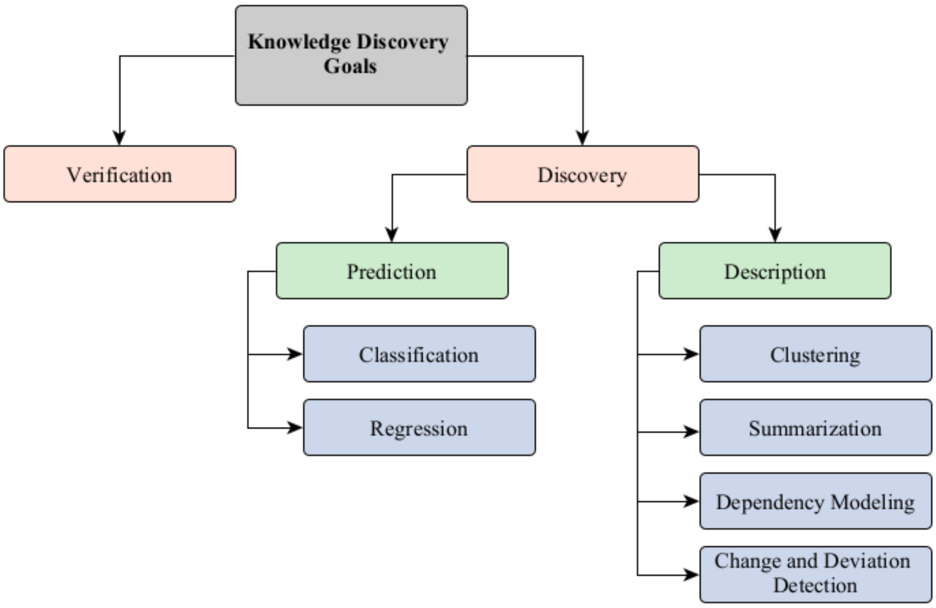


Figure 2.2: Data extraction methods (based on [5]).

The relative importance of prediction and description for particular data mining applications can vary considerably. However, in the context of KDD, description tends to be more important than prediction. This is in contrast to pattern recognition and machine learning applications where prediction is often the primary goal of the *KDD* process. The goals of prediction and description are achieved by using the following primary methods of data extraction [5]:

Classification: Learning a function that maps (classifies) a data item into one of several predefined classes.

Regression: Learning a function that maps a data item to a real-valued prediction variable and the discovery of functional correlation between variables.

Clustering: Identifying a finite set of categories or clusters to describe the data. Closely related to clustering is the method of probability density estimation which consists of techniques for estimating from data the joint multi-variate probability density function of all of the variables/fields in the database.

Summarization: Involves methods for finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.

Dependency modeling: Consists of finding a model that describes significant dependencies between variables. Dependency models exist at two levels:

- The *structural* level of the model specifies (often graphically) which variables are locally dependent on each other, and
- The *quantitative* level of the model specifies the strengths of the dependencies using some numeral scale.

Change and deviation detection: Focuses on discovering the most significant changes in the data from previously measured or normative values.

This research uses the method called *dependency modeling*, since we are interested in determining the dependency or correlation of each of the risk factors with the variable that indicates whether the person has breast cancer or not. In this way, we identify that *feature selection* is used to search features or attributes that have great contribution or most weight on the dataset. The following subsection describe the *feature selection* methods used in this research to achieve the above goal.

2.3 Feature selection methods

Feature selection is a popular technique used to find the subset of features that are relevant to build powerful learning models [19]. In the medical field, it can be used for the identification of the most crucial risk factors related to a particular disease. There are many feature selection algorithms reported in the literature, however, for this research, we are interested in those techniques that use the target variable, called *Supervised Feature Selection (SFS)* techniques. Kuhn et al. [6] indicates the approaches for *SFS* can be placed into three main categories, including filter, wrapper and embedded. In Figure 2.3 (based on [20]), the dotted line indicates the activities involved in each of the feature selection approaches.

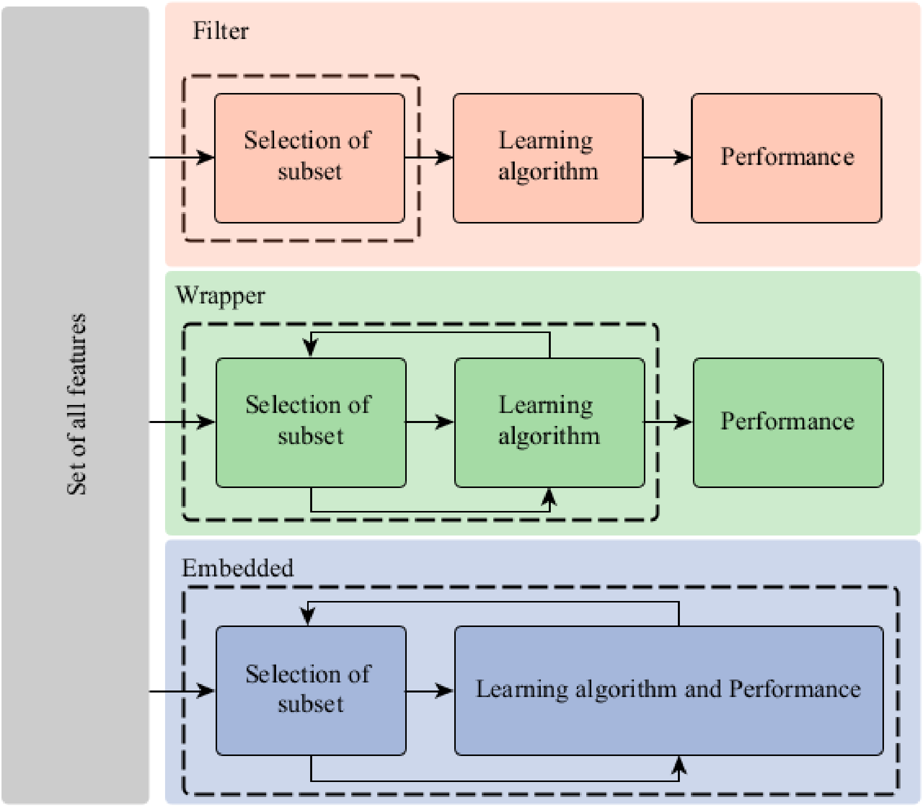


Figure 2.3: Feature selection approaches (based on [20]).

Filter methods: Uses statistical indicators to score and filter each feature, focusing on the characteristics of the data itself. As shown in Figure 2.3, the selection process is independent of the learning algorithm. Filter-based methods rank the features before the learning algorithm. In the selection process, each feature is evaluated individually to check if there is a plausible correlation between such feature and the observed classes. Only features with a relevant correlation would then be included in a learning algorithm.

Wrapper methods: Uses a learning algorithm to evaluate the feature set. This method scores the features using the learning algorithm that will ultimately be employed in classification. As shown in Figure 2.3, the feature selection process is integrated with the training of the learning algorithm, and the prediction ability of the model is used as the selection criterion to evaluate the feature subset. Wrapper-based methods evaluate multiple models using procedures that add or remove features to find the optimal combination that maximizes model performance.

Embedded methods: Compared to the other two methods, embedded feature selection is automatically built into the construction of the learning algorithm (as shown in Figure 2.3). Features with good ability of classification are selected, and then the selected feature subsets are used to perform the learning tasks.

All three approaches have advantages and drawbacks. The advantage of the *filter methods* is that the calculation is fast and does not depend on a specific model. However, the selection criteria is not directly related to the effectiveness of the model. Compared with the *filter*, the *wrapper* has better performance in generating high-quality subsets, but the data processing is computationally expensive since the learner needs to be trained many times during the feature selection process. Unlike *filter* selection, which does not consider subsequent classification algorithms, *wrapper* selection directly takes the performance of the final classification algorithms as the evaluation standard of the feature subset. In other words, *wrapper* feature selection chooses the most favorable feature subset for a given learning algorithm. Feature subset stability and adaptability are poor because each additional feature must be built as a feature subset for evaluation. Finally, *embedded* feature selection provides a trade-off solution between *filter method* and *wrapper methods*, which can solve the high redundancy of the *filter* algorithm and the computational complexity of the *wrapper* algorithm, but the design of the *embedded* method is tightly coupled with a specific learning algorithm, which in turn limits its application to other learning algorithms. *Filter methods* represent the best option for this research since unlike the other two, the results of selection are obtained from the characteristics of the data itself, unlike *wrapper methods*, selection results do not depend on the selected learning algorithm and in contrast to *embedded methods*, the learning method is not limited to a specific set.

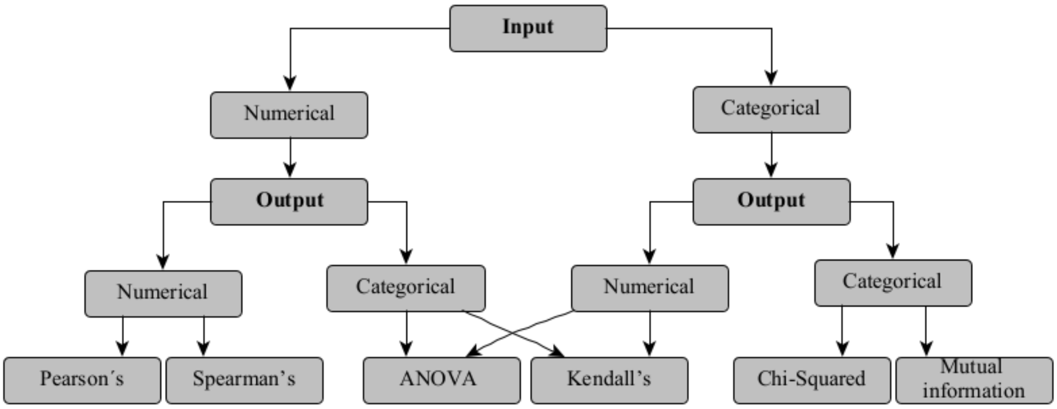


Figure 2.4: Filter-based feature selection methods (based on [21]).

Filter-based feature selection provides a variety of methods with different performance criteria for evaluating the value of information. The selection of the appropriate method usually depends on the data types of the attributes of the dataset (as seen in Figure 2.4 based on [21]). Common data types include numerical and categorical, although each may be further subdivided such as integer and floating point for numerical variables, and boolean, ordinal, or nominal for categorical variables. To select a filter-based method, the data type of the input and output attributes must be known. Input attributes are those that are provided as input to a model. Output attributes are those for which a model is intended to predict. Because the input and output attributes of the dataset used in this research are categorical (as will be described in Chapter 3), the methods selected for the feature selection process are: *Chi-Square* and *Mutual information.*The following two subsections describe these methods in detail.

2.3.1 Chi-squared test

The *Chi-squared test* [22] is a nonparametric statistical technique used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies, is one way to show a relationship between two categorical variables. The value of the *Chi-squared test* is given by Equation 2.1:

$$X_c^2 = \Sigma_i[(O_i - E_i)^2/E_i] \tag{2.1}$$

Where the subscript c represents the degrees of freedom. O is the observed frequency of variable i, and E is the expected frequency. The *Chi-squared test* summarizes the discrepancies between the expected number of times each outcome occurs (assuming that the model is true) and the observed number of times each outcome occurs, by adding the squares of the discrepancies, normalized by the expected numbers over all the categories.

To determine whether the *Chi-squared test* value indicates a statistical significance in the relationship between two categorical variable, the test results of each variable should be compared with the critical value from a chi-squared distribution table. If the chi-squared value is more than the critical value, then there is a significant relationship. To be able to make the comparison with a critical value the following information must be defined:

1. **Null hypothesis and alternate hypothesis:** The null hypothesis can be thought of as a nullifiable hypothesis. That means that can nullify it, or reject it. The alternate hypothesis is the researcher's thoughts about the experiment. For this research, they can be defined as follows:

   - Null Hypothesis (H0): Two variables are independent.
   - Alternate Hypothesis (H1): Two variables are not independent.

2. **Degrees of freedom (df):** Is the number of categories of the variable minus 1.

3. **p value:** Is used in hypothesis testing. The value P is expressed with decimals. The value must be greater than 0 and less than 1. The smaller the p-value, the more important and significant are the results.

    - If p > .10 → "not significant".
    - If p <= .10 → "marginally significant".
    - If p <= .05 → "significant".
    - If p <= .01 → "highly significant.".

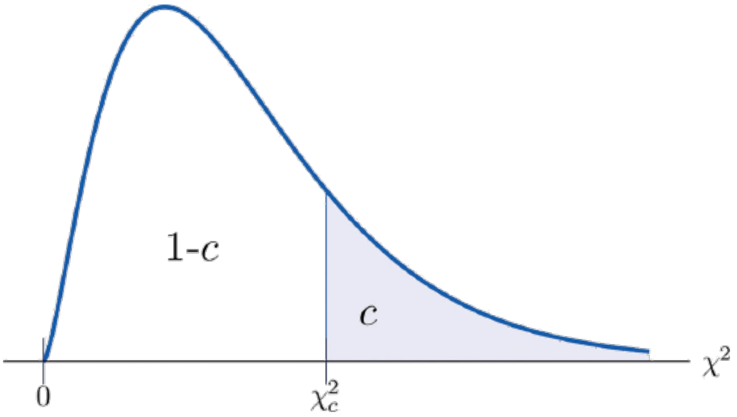
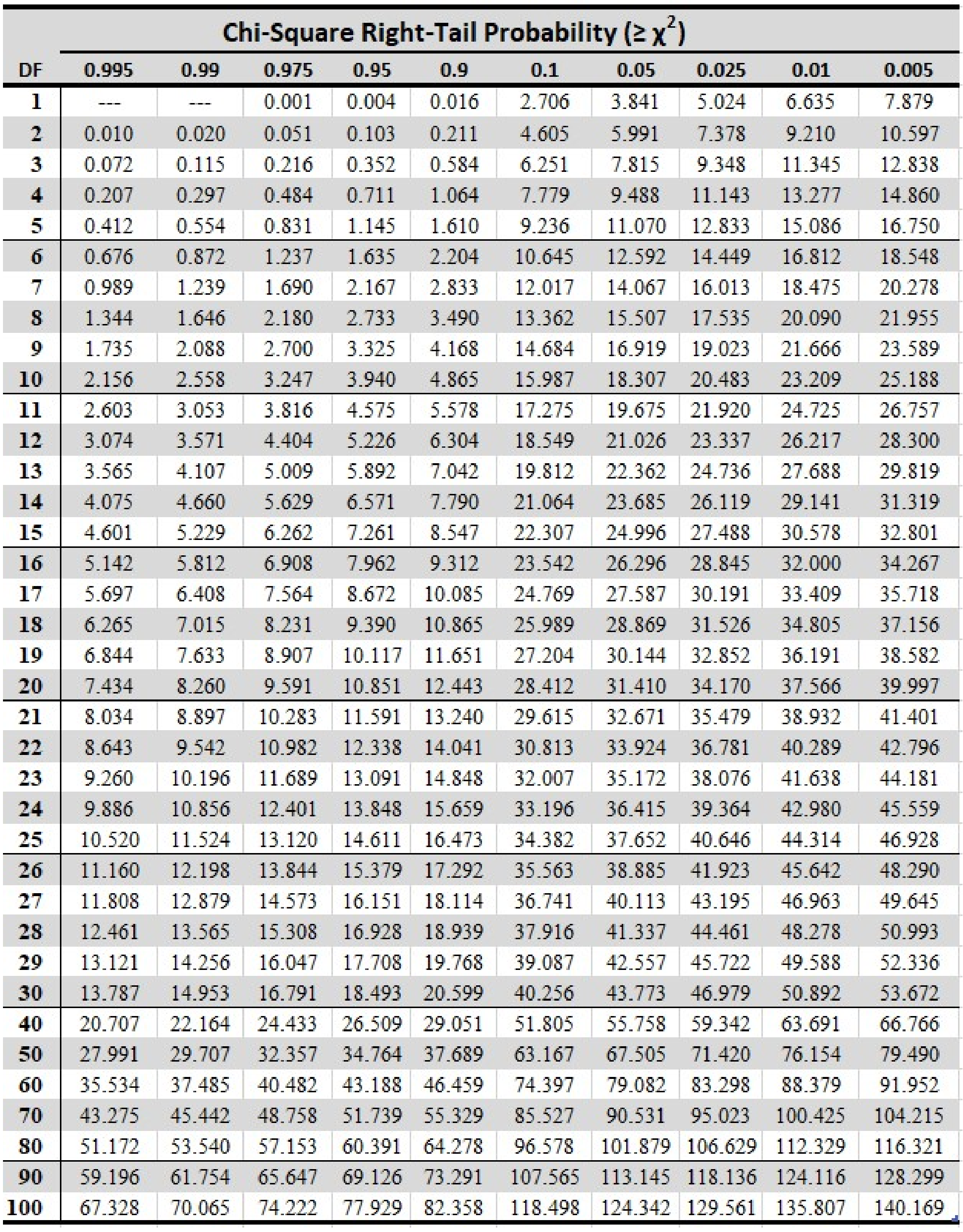


Figure 2.5: Critical value in chi-squared (taken from [23]).

The value of the chi-squared random variable X^2 with $df = k$ that cuts off a right tail of area c is denoted X^2_c and is called a critical value (Figure 2.5 taken from [23]). The critical value of each variable corresponds to the insertion between the *P value* and *degrees of freedom* in the Figure 2.6 (taken from [23]). If the chi-squared value is greater than the critical value, the *null hypothesis* is rejected and the *alternative hypothesis* is accepted, it is concluded that the variables are not independent and therefore there is a relationship of significance. Instead, if the chi-squared value of the variable is lower than the critical value, the *null hypothesis* is accepted and it is concluded that the variables are independent and therefore there is no relationship of significance.



	Chi-Square Right-Tail Probability (≥ χ^2)									
DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 2.6: Chi-squared distribution table (taken from [23]).

2.3.2 Mutual information

Mutual information [24] is usually a good measure for deciding the relevance of an attribute. *Mutual information* is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. The *Mutual information* between two random variables X and Y can be stated formally by Equation 2.2:

$$I(X;Y) = H(X) - H(X|Y) \tag{2.2}$$

where *I(X ; Y)* is the *Mutual information* of X and Y, *H(X)* is the entropy of X, and *H(X / Y)* is the conditional entropy of X given Y. The result is always greater than or equal to zero, where the greater the value, the relationship between the two variables increases. If the calculated result is zero, then the variables are independent. A threshold (cutoff) value is calculated in order to determine which attributes should be selected. As proposed in [25] the threshold value is calculated by means of the standard deviation in Equation 2.3:

$$S = \sqrt{\frac{n \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i}{n(n-1)}} \tag{2.3}$$

Where S is the standard deviation, x is the average value of the mutual information, and n is the number of attributes used in the dataset. For an attribute to be selected, its mutual information value must be greater than the threshold value S.

For the evaluation stage of the *KDD* process, the results obtained by the feature selection methods will be evaluated using classification algorithms. At the evaluation stage, the aim is to demonstrate that the most relevant risk factors resulting from the feature selection process are significant to classify breast cancer cases as if they were performing with all the risk factors of the dataset. In the subsection below, the classification algorithms for the evaluation stage are described.

2.4 Classification algorithms

A classification algorithm is a supervised learning technique that is used to identify the category of new observations on the basis of training data. Classification algorithms employ a variety of statistical, probabilistic and optimisation methods to learn from the given dataset and detect useful patterns to classify new information. In this research we opted to use different variants of supervised machine learning algorithms to evaluate the results obtained from the feature selection stage. These algorithms were selected based on the characteristics of the dataset and the capabilities of the development tool used (section 2.4 provides more information about the selected development tool, RapidMiner). Below is an overview of each of the seven algorithms used.

2.4.1 Decision tree

A decision tree [26] models the decision logic i.e., tests and corresponds outcomes for classifying data items into a tree-like structure. The nodes of a decision tree normally have multiple levels where the first or top-most node is called the root node. All internal nodes (i.e., nodes having at least one child) represent tests on input variables or attributes. Depending on the test outcome, the classification algorithm branches towards the appropriate child node where the process of test and branching repeats until it reaches the leaf node. The leaf or terminal nodes correspond to the decision outcomes.

2.4.2 Decision stump

A decision stump [27] is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input variable.

2.4.3 Random tree

Random tree [28] is a tree-based classification with the difference that the tree is only built with a random subset of variables and not with all input variables. Trees derived with traditional methods often cannot be grown to arbitrary complexity, this produce a possible loss of generation accuracy on unseen data. The limitation on complexity usually means suboptimal accuracy on training data. Random tree tries to solve this limitation with the selection of a subset of random variables to increases the accuracy for unseen data.

2.4.4 Deep learning

Deep learning [29] attempts to mimic the human brain through a combination of data inputs, weights, and bias. These elements work together to accurately recognize, classify, and describe objects within the data. Deep learning consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization. This progression of computations through the network is called forward propagation. The input and output layers of a deep neural network are called visible layers. The input layer is where the deep learning model ingests the data for processing, and the output layer is where the final prediction or classification is made. Another process called backpropagation uses algorithms, like gradient descent, to calculate errors in predictions and then adjusts the weights and biases of the function by moving backwards through the layers in an effort to train the model. Together, forward propagation and backpropagation allow a neural network to make predictions and correct for any errors accordingly.

2.4.5 Generalized linear model

The generalized linear model (GLM) [30] extends simple linear regression by allowing each outcome of the dependent variable to come from a large range of probability distributions. It is an umbrella term that encompasses many other models, which allows the response to have an error distribution other than a normal distribution. The models include Linear Regression, Logistic Regression, and Poisson Regression. In a linear regression model, the target variable is expressed as a linear function of all the predictors. The underlying relationship between the response and the predictors is linear. Also, the error distribution of the response variable should be normally distributed. Nevertheless, GLM models allow us to build a linear relationship between the response and predictors, even though their underlying relationship is not linear. This is made possible by using a link function, which links the response variable to a linear model. Unlike linear regression models, the error distribution of the response variable need not be normally distributed. The errors in the response variable are assumed to follow an exponential family of distribution (i.e. normal, binomial, Poisson, or gamma distributions).

2.4.6 K-nearest neighbor

K-Nearest Neighbors (k-NN) [31] stores all available records and predicts the class of a new instances giving attention to similarity measurements from the nearest neighbors in likelihood. This classification technique is known to be lazy learning method because it keeps the data members stored simply in efficient data structures like hash table by virtue of which computation cost becomes less to check and apply the appropriate distance function between the new observation and all k number of different data points stored and then come to any conclusion about the label of the new data point, without constructing a mapping function or internal model like other classification algorithms. Result is obtained from a simple majority support of the k number of nearest neighbors of each new data point.

2.4.7 Naïve Bayes

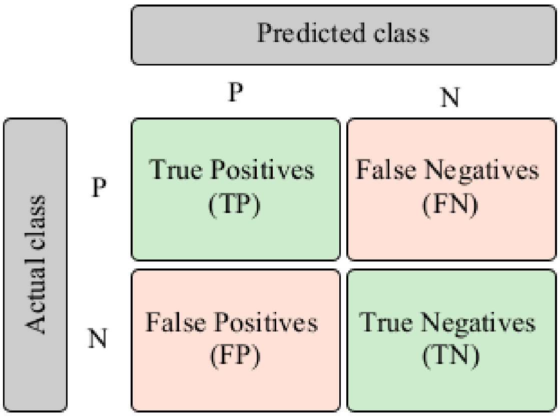
Naïve Bayes [32] is a classification technique based on the Bayes' theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayes' theorem can describe the probability of an event A given some prior probability of event B represented by $P(A|B)$ in Equation 2.4:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.4}$$

Where A and B are events, $P(A)$ and $P(B)$ are the probabilities of observing A and B independent of each other, $P(A|B)$ is the conditional probability, i.e. Probability of observing A, given B is true, and $P(B|A)$ is the probability of observing B, given A is true.

2.5 Performance Metrics

In order to understand the outputs of classification methods and to identify whether the results are good or bad, the so-called confusion matrix [33] is often used. The basic framework of the confusion matrix has been provided in Figure 2.7 (based on [33]). For a binary classification problem the framework has two rows and two columns. Across the top is the predicted class labels and down the side are the actual class labels. Each cell contains the number of predictions made by the classifier that fall into that cell.



		Predicted class	
		P	N
Actual class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 2.7: The basic framework of the confusion matrix (based on [33]).

In this framework, true positives (TP) are the positive cases where the classifier correctly identified them. Similarly, true negatives (TN) are the negative cases where the classifier correctly identified them. False positives (FP) are the negative cases where the classifier incorrectly identified them as positive and the false negatives (FN) are the positive cases where the classifier incorrectly identified them as negative. The following measures, which are based on the confusion matrix, are commonly used in the literature an in this research are used to analyse the performance of classifiers:

Accuracy: It is the total number of correct predictions divided by the total number of predictions, defined by Equation 2.5. Accuracy is a metric that measures the balance between true positives and true negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

Precision: Precision is the number of true positives divided by the number of true positives and false Positives (Equation 2.6). Put another way, it is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV). Precision can be thought of as a measure of a classifiers exactness.

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

Recall: Recall is the number of true positives divided by the number of true positives and the number of false negatives (Equation 2.7). Put another way it is the number of positive predictions divided by the number of positive class values in the test data. It is also called sensitivity or the True Positive Rate (TPR). Recall can be thought of as a measure of a classifiers completeness.

$$Recall = \frac{TP}{TP + FN} \tag{2.7}$$

2.6 Development Tools

For the development of this research, we were looking for a software tool focused on data analysis, able to provide a set of useful methods for data preparation, implementation of machine learning algorithms and for data visualization, and whose learning curve was efficient and fast. These days there are several tools available that meet these characteristics, such as *RapidMiner* and *Weka.* These tools are known to help in cluster, regression, and classification analysis, data visualization, text mining, etc. These tools can assist in transforming a vast amount of data into useful information and knowledge.

RapidMiner[2] is an user interactive environment for machine learning and data mining processes. It is opensource, free project implemented in Java. It represents a modular approach to design even very complex problems, a modular operator concept which allows the design of complex nested operator chains for a huge number of learning prob-

[2] `https://rapidminer.com`

lems. *RapidMiner* uses XML to describe the operator trees modeling knowledge discovery processes. *RapidMiner* has flexible operators for data input and output in different file formats. It contains more than 100 learning schemes for classification, regression and clustering tasks [34].

Weka[3] is a widely used toolkit for machine learning and data mining that was originally developed at the University of Waikato in New Zealand. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for regression, classification, clustering, association rules, visualization, and data pre-processing [34].

Comparing the two tools, RapidMiner proved to be a better alternative for the following reasons:

- It has a wider variety of operators for data processing and modeling. Provides potentially useful operators at the preprocessing stage.
- In addition to the default operators, it has an extensive catalog of extensions to download for specific problems.
- It has an extension to use *Weka* operators.
- It provides a wide range of controls to facilitate dynamic visualization of data and results.
- Provides the option to create custom operators.
- The documentation is extensive and frequently updated.

2.7 Related Work

Some relevant works related to identifying relevant attributes in datasets to determine the likelihood of breast cancer [35]–[39], most of them focus on analyzing databases with clinical information from specialized imaging studies, such as mammograms. Nevertheless,

[3] `https://www.cs.waikato.ac.nz/ml/weka/`

we are interested in those works that determine the relevant risk factors and their use in the prediction of breast cancer. The following studies focus on analyze the relevance of breast cancer risk factors with multiples feature selection and machine learning techniques [40]–[43].

The authors in [40], present a prevention and control system for breast cancer by means of *Item Rule Association (IRA)* algorithms applied on a private dataset with 2,966 records and 83 attributes. An important characteristic of their work is the creation of their own dataset by interviewing patients from 22 hospitals over a one-year period and storing clinical, personal, and socioeconomic information. Three types of rules defining the more relevant risk factors were identified; 35 rules were obtained using a single factor, 19 rules were obtained combining two factors, and 9 rules were obtained combining three factors.

In [41], the authors focused on determining breast cancer risk factors for patients in Indonesia and identified differences against patients in the United States, using a private dataset with 1907 records and 21 attributes (containing demographic, pathology and therapy information). They used three features selection methods: i) *Information gain*, ii) *Fisher's discriminant*, and iii) *Chi-squared test*, to select the best attributes (risk factors). They also applied *Hierarchical K-means* clustering to remove attributes that have the lowest contribution. As a result, out of the 21 original attributes, 14 relevant attributes were obtained.

Rather than determining the relevance of risk factors, in [42], the authors analyzed the effect of caffeine consumption on the incidence of breast cancer cases. A *Bayesian network* was used on a dataset with 1,302 records from The Clinical Breast Care Project (CBCP) . The study concluded that caffeine consumption does not affect the incidence of cancer in its study population.

In [43], are generated risk factor rules by means of *Association Rule Mining (ARM)*, using the Breast Cancer Surveillance Consortium's (BCSC) Risk Factors dataset. This public dataset contains 6,318,638 cases and 13 attributes, although all records containing at least one missing value were discarded. The *Logit* model was used to select those factors that may affect the likelihood of breast cancer. A set of 5 rules was obtained for breast cancer cases and 4 rules for non-cancer cases. However, because of the class imbalance problem, they had to adjust the algorithm for the breast cancer cases.

The class imbalance is a problem that is commonly found in cancer-related datasets, since there are fewer positive cases compared to the number of negative cases. In the literature, few papers that address this problem were identified [44], [45].

In [44], the authors focused on this issue by implementing three data-level resampling approaches: *random under-sampling*, *random over-sampling*, and a *hybrid of over- and under-sampling.* These techniques were applied on the BCSC's Risk Factors dataset, after discarding all records containing at least one missing value. To evaluate the results of each of the approaches, the authors used three different classification algorithms: *Decision Tree, Random Forest*, and *XGBoost.* Their results showed that performance improves when resampling techniques are used compared to when no techniques are applied.

In [45], the authors proposed a prognosis model framework to predict Invasive Disease-Free Survival (i.e., the length of time after the primary treatment ends and no signs of cancer appear again) for early-stage breast cancer patients. They used a private dataset with 12,119 records and 89 attributes of the Clinical Research Center for Breast (CRCB) from West China Hospital of Sichuan University. The features consist of demographic, diagnosis, pathology, and therapy information. A *Stratified Feature Selection* was used by calculating the importance score using *Gradient boosting decision tree algorithm (XGBoost)*, that resulted in a selection of 23 features, including some risk factors.

2.8 Summary

This chapter provides a medical background to understand the relationship between risk factors and breast cancer, and the social problem that this disease represents. As part of the computational background, the stages of the Knowledge Discovery in Databases process that this thesis uses as part of the methodology were described. In addition, an overview of the methods and techniques used at the data mining and evaluation stages was provided. Finally, the related work to the research topic was presented, as well the alternative solutions that have been proposed.

Chapter 3

Dataset and Preprocessing

In this chapter, the selection criteria for the dataset is described. Next, a description of the characteristics of the selected dataset is provided. Finally, each of the necessary steps that were executed for the preprocessing stage are explained. This chapter covers the first two stages of the *Knowledge Discovery in Databases* process: *selection* and *preprocessing.*

3.1 Dataset Selection

This section describes what was done during the *selection* stage. The first part of this research was to gather and analyze all possible datasets that could be used. Unfortunately, the number of datasets publicly available that are related to breast cancer, and particularly, risk factors, is very small. Additionally, this initial search was complicated because the majority of public datasets of breast cancer contain information about diagnosis, pathology, and therapy information, but not enough about risk factors. This is the case for some of the most widely used public datasets in the breast cancer literature; *Beast Cancer dataset* [46] by the *Oncology Institute*, *Breast Cancer Wisconsin Diagnostic dataset* [47] and *Breast Cancer Wisconsin Prognostic dataset* [48], which only provide information about one or a maximum of two risk factors.

Although many of the public datasets did not have enough information on risk factors, four datasets provided by the *Breast Cancer Surveillance Consortium (BCSC)* were identified with a good number of risk factors and cases. The characteristics of the BCSC datasets are described below and summarized in the Table 3.1.

Risk Estimation Dataset [49]: Includes data from 2,392,998 mammograms and 16 attributes of which 11 are risk factor of breast cancer. Cancer registry and pathology data were linked to data on mammography and incident breast cancer (invasive [1] or ductal carcinoma in situ [2]). In August 2012, the BCSC added a second version of this risk estimation dataset. The second version limits observations to one per woman, as opposed to multiple observations.

Risk Factors Dataset [51]: Includes information from 6,788,436 mammograms between January 2005 and December 2017. The dataset includes participant characteristics previously shown to be associated with breast cancer ris. The dataset has 13 attributes of which 11 are risk factors.

Hormone Therapy and Breast Cancer Incidence [52]: The dataset includes information from 603,411 screening mammograms performed on women from January 1997 to December 2003. It includes data from women aged 50-69 who did not have a previous diagnosis of breast cancer and who had undergone breast mammography in the prior 9 to 30 months. The mammogram data were linked to cancer registry and pathology data to identify incident breast cancer (invasive or ductal carcinoma in situ) within one year after the screening mammogram. This aggregate dataset includes frequencies and adjusted quarterly rates of postmenopausal hormone therapy use and breast cancer (overall and by invasive[1] or DCIS[2] and estrogen receptor status).

Digital Mammography Dataset [53]: Includes data derived from a random sample of 20,000 digital and 20,000 film-screen mammograms performed between January 2005 and December 2008. Some women contribute multiple examinations to the data. The dataset includes 13 attibutes of which 5 are risk factors.

[1] Invasive breast cancer is when the cancer has spread into surrounding breast tissue [50].

[2] Ductal carcinoma in situ (DCIS) is a non-invasive or pre-invasive breast cancer [50].

Table 3.1: Main characteristics of the BCSC datasets.

Dataset	Attributes	Risk Factors	Cases	Records
Risk Estimation v.1	16	11	2,392,998	280,660
Risk Estimation v.2	16	11	1,007,660	181,903
Risk Factors	13	11	6,788,436	1,522,340
Hormone Therapy and Breast Cancer Incidence	15	1	603,411	28
Digital Mammography	13	5	40,000	40,000

For our analysis, the *Risk Estimation Dataset v.2* [3] was selected for three reasons: i) it provides an attribute indicating the presence of breast cancer, that is used to classify each case, ii) it contains information about 11 risk factors, and iii) patients had no previous diagnosis of breast cancer up until the screening test recorded in the dataset. This last point is important because we are interested in determining relevant risk factors when no cancer has been diagnosed before. For instance, the *Risk Factors Dataset* includes patient information that have had cancer at some point in their life. As presented in Table 3.1, it is important to note that the Risk Estimation dataset has two versions: the first version (v.1) contains multiple observations per patient obtained on different dates; while the second version (v.2) is limited to only one observation per patient (all other information is the same in both versions). An issue with the first version is that the dataset does not contain specific information about those multiple observations, for instance, when each observation occurred and for which patient. Thus, we preferred to use the second version to make sure the information was related to single observations. The following section provides a detailed description of the attributes and records of the selected dataset.

3.2 Dataset Description

The records of the *Risk Estimation Dataset v.2* are described by 16 attributes shown in Table 3.2, each attribute is given a name, a description and the values that can be assigned to the attribute. The *Attribute* column provides the original names of each attribute in the dataset. Although most attribute names are already understandable, the *Description*

[3] Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). `http://www.bcsc-research.org/`

column provides detailed information about the attribute. Finally, the column *Values* lists the values that can be assigned to the attribute, which are given for a numerical value from 0 to 10. In each attribute, these numeric values have different meanings, which are also presented in the last column of Table 3.2.

The first attribute *menopause* provides the person's menopause status; the possible values that may present in this attribute are 0, 1 and 9; the value 0 represents that the person's menopause status is *premenopausal*, the value 1 is assigned when the person's menopause status is *postmenopausal* or when the person is over 55 years old; finally, the value 9 is assigned when the data is *unknown*.

The attribute of *agegrp* indicates the age group to which the patient belongs; it can take values from 1 to 10, each value has a range of 5 years starting with the value 1 ranging from 35 to 39 years and ending with the value 10 ranging from 80 to 84 years.

The *density* attribute provides the person's breast density based on the *Breast Imaging Reporting and Data System (BI-RADS)* scale; values range from 1 to 4 incrementally, with 1 being the lowest density value and 4 being the value for highest density; finally, the value 9 is assigned when the data is unknown or when a different measurement system was used.

The fourth attribute *race* includes four different races: *white*, *asian* or *pacific islander*, *black* and *native american*; the value 5 is assigned when the person's race is not among these options or when the race is mixed, the last value that can be assigned is 9, it is used when the person's race is unknown.

The attribute *hispanic* works to identify those people who are hispanic, it can be assigned the value of 0 for those who are not hispanic and 1 for those who are; if it is unknown if the person is hispanic, the value 9 is assigned.

The attribute *bmi* categorizes the person's body mass index (BMI[4]) based on four options: 1 for BMI between 10 and 24.99, 2 for BMI between 25 and 29.99, 3 for BMI between 30 and 34.99, and 4 for those with BMI over 35, the value 9 is assigned when the person's BMI is unknown.

[4] According to the World Health Organization [54], for adults over 20 years old, a BMI less than 18.5 is equivalent to underweight, from 18.5 to 24.9 is a normal weight, from 25.0 to 29.9 is pre-obesity, from 30.0 to 34.9 is obesity class I, from 35.0 to 39.9 is obesity class II, and above 40 is obesity class III.

The seventh attribute *agefirst*, provides information about the age of the person's first birth; the possible values that can be assigned to this attribute are 0, 1, 2 and 9; the value 0 is assigned when the first birth occurred when the person was less than 30 years old and the value 1 when the person was 30 or more; if the person has never given birth, the value 2 is assigned; when the data is unknown, the value 9 is assigned.

The attribute *nrelbc* indicates the number of first-degree relatives with breast cancer that the person has; the value 0 is assigned if the person does not have relatives with breast cancer, the value 1 if the person have only one relative and the value 2 if the person have 2 or more relatives with breast cancer, the value 9 is assigned when the information is unknown.

The attribute *brstproc* indicates whether the person has previously undergone any breast procedure; the value 0 means that the person has not performed breast procedures before and the value 1 means that some procedure has been performed, the value 9 is set when the data is unknown.

To know the result of the person's last mammogram, the attribute *lastmamm* takes the value 0 when it is negative and 1 when it is a false positive; if the data is unknown, the value 9 is assigned.

The eleventh attribute *surgmeno* provides the person's type of menopause; the possible values that may present in this attribute are 0, 1 and 9; the value 0 represents that the person's type of menopause is *natural*, the value 1 is assigned when the type of menopause is *surgical*, finally, the value 9 is assigned when the data is *unknown* or when the values of *menopause* is 0 or 9.

The attribute *hrt* indicates if the person is currently taking hormone restitution therapy, if this is the case the value of this attribute is 1, otherwise it is 0; as in *surgmeno*, the value 9 is assigned when the data is *unknown* or when the values of *menopause* is 0 or 9.

The attribute *invasive* identifies those people who have been diagnosed with invasive breast cancer. The attribute *cancer* indicates the person's cancer diagnosis, the value 1 is assigned when the diagnosis is positive to cancer and the value 0 is assigned in the opposite case.

The dataset authors provide the attribute *training* as a suggestion to identify the records that belong to the training and validation set, assigning a value of 1 and 0 respectively.

Finally the attribute called *count* represents the number of people who presented the same combination of values of the previous attributes. The dataset *Risk Estimation v.2* is a large cross-classification of risk factors by cancer outcome, meaning that if the value of the *count* column for a particular row is 13, means that there were 13 people who reported similar conditions to obtain the same values of the attributes.

Table 3.2: Description of attributes of the *Risk Estimation Dataset v.2.*

No.	Attribute	Description	Values
1	menopause	Menopausal status	0 = premenopausal 1 = postmenopausal or $age \geq 55$ 9 = unknown
2	agegrp	Age (years) in 5-year groups	1 = 35 - 39 2 = 40 - 44 3 = 45 - 49 4 = 50 - 54 5 = 55 - 59 6 = 60 - 64 7 = 65 - 69 8 = 70 - 74 9 = 75 - 79 10 = 80 - 84
3	density	BI-RADS breast density codes	1 = almost entirely fat 2 = scattered fibroglandular densities 3 = heterogeneously dense 4 = extremely dense 9 = unknown or different measurement system
4	race	Race	1 = white 2 = asian/pacific islander 3 = black 4 = native american 5 = other/mixed 9 = unknown
5	hispanic	Patient is Hispanic	0 = no 1 = yes 9 = unknown

Table 3.2: Description of attributes of the *Risk Estimation Dataset v.2* (continued).

No.	Attribute	Description	Values
6	bmi	Body mass index	1 = 10 - 24.99 2 = 25 - 29.99 3 = 30 - 34.99 4 = 35 or more 9 = unknown
7	agefirst	Age at first birth	0 = age < 30 1 = age 30 or greater 2 = nulliparous 9 = unknown
8	nrelbc	Number of first-degree relatives with breast cancer	0 = zero 1 = one 2 = two or more 9 = unknown
9	brstproc	Previous breast procedure	0 = no 1 = yes 9 = unknown
10	lastmamm	Result of last mammogram	0 = negative 1 = false positive 9 = unknown
11	surgmeno	Type of menopause	0 = natural 1 = surgical 9 = unknown or not menopausal (menopause=0 or menopause=9)
12	hrt	Current hormone therapy	0 = no 1 = yes 9 = unknown or not menopausal (menopause=0 or menopause=9)
13	invasive	Diagnosis of invasive breast cancer	0 = no 1 = yes
14	cancer	Diagnosis of invasive or ductal carcinoma in situ breast cancer	0 = no 1 = yes
15	training	Training data	0 = no (validation) 1 = yes (training)
16	count	Frequency count of this combination of covariates and outcomes (all variables 1 to 15)	

Table 3.3 shows the distribution of positive and non-cancer numbers of both cases and records within the dataset. To clarify the difference between *case* and *record* terms, it is important to define that a *case* corresponds to a person, and a *record* corresponds to a row in the dataset, which, contains in the *count* attribute the number of cases that reported the same values of the attributes 1 to 15 of Table 3.2. This means that the number of

records in the dataset is less than the number of cases because each record corresponds to several cases. In total, the dataset contains 1,007,660 cases with 7,319 (0.73%) positive cancer and 1,000,341 (99.27%) non-cancer cases. The difference between the number of positive and non-cancer cases is clearly evident. Regarding to the number of records, the overview is similar, the dataset contains 181,903 records with 6,274 (3.45%) positive cancer and non-cancer 175,629 (96.55%) records. This imbalance in the data is an issue commonly present in this type of problems and will be further discussed in Chapter 4.

Table 3.3: Distribution of positive and non-cancer cases within the Risk Estimation v.2 dataset.

Breast Cancer Diagnosis	Cases	Cases (%)	Records	Records (%)
Yes	7,319	0.73	6,274	3.45
No	1,000,341	99.27	175,629	96.55
Total	1,007,660	100	181,903	100

After the selection of the dataset, it was necessary to apply a set of preprocessing operations in order to improve the quality of the data. The preprocessing stage in the *KDD* process is important for the success of later stages. The following section describes the preprocessing operations applied to the *Risk Estimate v.2* dataset.

3.3 Preprocessing

Data preprocessing techniques generally refer to the addition, deletion, or transformation of training set data [6]. Data preprocessing is an important stage in the *KDD* process, because it can be handle various types of dirty data on large datasets. Dirty data consist of data noise, incomplete, inconsistent and missing values.

To evaluate data quality during preprocessing, we use the tool called *Quality Measures* that *RapidMiner* provides. *Quality Measures* is a way of seeing at a glance typical data quality problems. They are encoded with the colors specified below. The *Quality Measures* (located at the top left) of the *agegrp* attribute are shown as an example in Figure 3.1. Here are the details about how those quality measurements are calculated and what they mean:

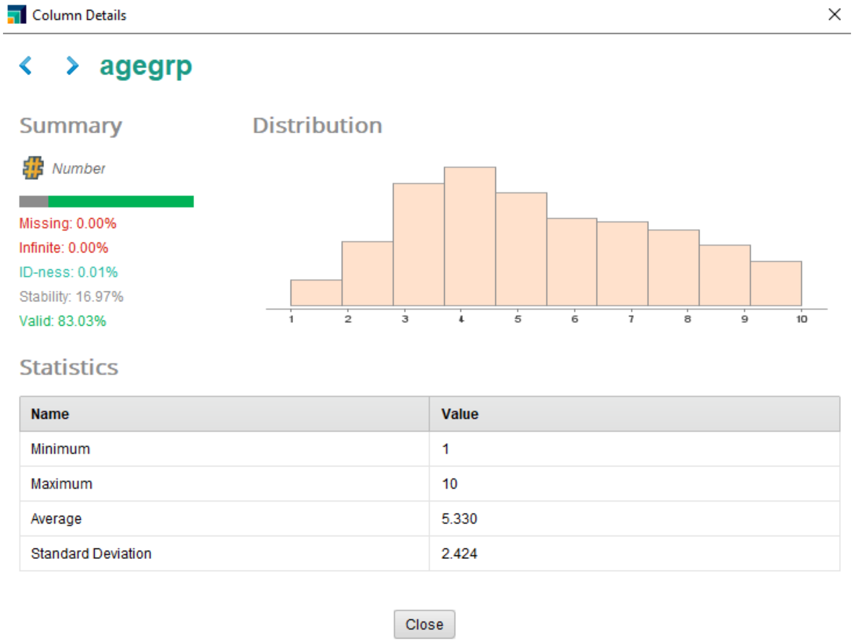
- **Missing** *(red)*: The number of missing values in this column divided by the number of rows.

- **Infinite** *(red)*: The number of infinite values in this column divided by the number of rows.

- **ID-ness** *(blue)*: The number of different values for this column divided by the number of rows.

- **Stability** *(gray)*: The count for the most frequent non-missing value for this column divided by the number of rows.

- **Valid** *(green)*: The fraction of values of this column which are not counted as missing, infinite, id, or stable.



Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 16.97%
Valid: 83.03%

Name	Value
Minimum	1
Maximum	10
Average	5.330
Standard Deviation	2.424

Figure 3.1: Column details in RapidMiner.

Table 3.4 shows the quality measures of the attributes in the original *Risk Estimation v.2* dataset, before preprocessing. The rows correspond to each of the attributes and the columns to the quality measures. Although it seems that initially all attributes have a good percentage of valid values, some irregularities are detected. For example, all unknown values are not properly detected, these should be reflected in the percentage of the *Missing* measure. In order to resolve these irregularities and obtain a dataset suitable for its use in later stages, we applied four different preprocessing operations: i) simple conversion operations, ii) transformation of attributes, iii) removal of attributes and iv) elimination of records with unknown values.

Table 3.4: Initial quality measurements of the *Risk Estimation v.2* dataset.

No.	Attribute	Missing	Infinite	ID-ness	Stability	Valid
1	menopause	0.00%	0.00%	0.00%	77.67%	22.33%
2	agegrp	0.00%	0.00%	0.01%	16.97%	83.03%
3	density	0.00%	0.00%	0.00%	29.91%	70.09%
4	race	0.00%	0.00%	0.00%	57.20%	42.79%
5	hispanic	0.00%	0.00%	0.00%	57.31%	42.69%
6	bmi	0.00%	0.00%	0.00%	41.05%	58.95%
7	agefirst	0.00%	0.00%	0.00%	39.77%	60.23%
8	nrelbc	0.00%	0.00%	0.00%	62.25%	37.75%
9	brstproc	0.00%	0.00%	0.00%	60.94%	39.06%
10	lastmamm	0.00%	0.00%	0.00%	56.31%	43.69%
11	surgmeno	0.00%	0.00%	0.00%	45.04%	54.96%
12	hrt	0.00%	0.00%	0.00%	35.49%	64.51%
13	invasive	0.00%	0.00%	0.00%	97.11%	2.89%
14	training	0.00%	0.00%	0.00%	65.99%	34.01%
15	cancer	0.00%	0.00%	0.00%	96.34%	3.65%
16	count	0.00%	0.00%	0.08%	57.07%	42.85%

3.3.1 Simple conversion operations

Two conversion operations were applied to the original dataset. The first operation was to convert all data types from numerical to categorical, except the count attribute which remained as a numerical attribute. As mentioned in Chapter 2, the data type of the predictor and outcome varibles are important to define the feature selection and classification algorithms that are useful to work with. Initially, all the attribute values within the dataset are numbers from 0 to 10, due to this fact Rapid Miner assigned them the data type *Integer* by default, however, the actual meaning of those values corresponds to a data type *Category*. It is important to perform this operation so that the data is interpreted properly.

The second operation was to convert all 9 values to the categorical value of *unknown* in all attributes that contain this value (i.e., attributes 1 and 3 to 12, in Table 3.2).

It is important to perform this operation with the aim that the feature selection and classification algorithms do not evaluate it as a categorical value, instead the algorithms evaluate it as what it is, *unknown* data.

Table 3.5 shows the quality measures after the two previous operations, the table only lists attributes 1 and 3 to 12 because the quality measures for attributes 2 and 13 to 16 are the same as in Table 3.4. It can be noticed that all attributes have decreased in the percentage of *Valid* measure due to the increase in the percentage of *Missing.* While these values in the quality measures could be considered worse than the initial ones, it should be noted that the data is being assigned the correct meaning for its use in later stages of the *KDD* process.

Table 3.5: Quality measurements after simple conversion operations.

No.	Attribute	Missing	Infinite	ID-ness	Stability	Valid
1	menopause	6.33%	0.00%	0.00%	82.96%	10.71%
3	density	25.45%	0.00%	0.00%	40.05%	34.50%
4	race	20.54%	0.00%	0.00%	71.95%	7.51%
5	hispanic	29.34%	0.00%	0.00%	81.15%	0.00%
6	bmi	40.38%	0.00%	0.00%	37.96%	21.66%
7	agefirst	35.00%	0.00%	0.00%	60.92%	4.08%
8	nrelbc	13.70%	0.00%	0.00%	72.16%	14.14%
9	brstproc	12.08%	0.00%	0.00%	69.33%	18.59%
10	lastmamm	38.41%	0.00%	0.00%	91.22%	0.00%
11	surgmeno	45.93%	0.00%	0.00%	59.21%	0.00%
12	hrt	35.45%	0.00%	0.00%	54.58%	9.97%

3.3.2 Attribute transformation

After analyzing the values of three attributes, specifically, value *1* of the *menopause* attribute, value *unknown* of the *surgmeno* attribute, and value *unknown* of the *hrt* attribute (attributes 1, 11 and 12 in Table 3.2 respectively); we decided to transform these three attributes to clarify the information given by those values.

For the *menopause* attribute, value *1* refers to postmenopausal women or women of more than 55 years old. It is possible to identify true postmenopausal cases by means of the *surgmeno* attribute. If the *surgemeno* attribute contains a *0* or *1*, it means that the record refers to a postmenopausal woman, and these records are assigned a value of

1 in the *menopause* attribute. A new value *2* was created and assigned to those cases where it is not possible to define whether a woman is postmenopausal or is older than 55 years. The attribute was renamed as *menopause_new* to differentiate from the original (see attribute 1 in Table 3.7). Originally, value *1* was assigned to 140,843 records; after the transformation 107,810 records were detected as true postmenopausal records (that were left with a value of *1*), and the rest (33,033 records) were assigned the new value of *2*. As can be seen in Table 3.6, this change produces an increase in the *Valid* measure from 10.71% to 29.39% because a new value is introduced to the variety of values of the attribute and this produces that the *stability* measure decreases from 82.96% to 64.28%.

For the *surgmeno* attribute, value *unknown* is given to women that have not undergone menopause yet or the status of menopause is unknown. To clearly identify cases that have not undergone menopause and separate them from those that are unknown, a new value *2* was created to refer to cases that are still not menopausal by checking if the *menopause* attribute is *0*. The attribute was renamed as *surgmeno_new* to differentiate from the original (see attribute 10 in Table 3.7). Originally, value *unknown* was assigned to 83,545 records; after this operation 29,542 records were given the value of *2*, and 54,003 remained as *unknown*. As presented in Table 3.6, this change produces an significant increase in the *valid* measure from 0.00% to 24.42% because the values for *missing* and *stability* are reduced from 45.93% to 29.69% and 59.21% to 45.89% respectively.

Similarly, for the *hrt* attribute, the same value *unknown* is assigned to cases that have not presented menopause or to cases where the use of hormone restitution therapy is unknown. To clearly identify cases that have not undergone menopause and separate them from those that are unknown, a new value *2* was created to refer to cases that are still not menopausal by checking if the *menopause* attribute is *0*. The attribute was renamed as *hrt_new* to differentiate from the original (see attribute 11 in Table 3.7). Originally, value *unknown* was assigned to 64,489 records; after this operation 29,542 records were given the value of *2*, and 34,947 remained as *unknown*. As presented in Table 3.6, this change produces an significant increase in the *valid* measure from 9.97% to 36.96% because the values for *missing* and *stability* are reduced from 35.45% to 19.21% and 54.58% to 43.83%

respectively.

Table 3.6: Quality measurements after attribute transformation.

No.	Attribute	Missing	Infinite	ID-ness	Stability	Valid
1	menopause	6.33%	0.00%	0.00%	64.28%	29.39%
11	surgmeno	29.69%	0.00%	0.00%	45.89%	24.42%
12	hrt	19.21%	0.00%	0.00%	43.83%	36.96%

3.3.3 Attribute removal

There are potential advantages to removing predictors prior to modeling [6]. First, fewer predictors means decreased computational time and complexity. Second, if two predictors are highly correlated, this implies that they are measuring the same underlying information. Removing one should not compromise the performance of the model and might lead to a more parsimonious and interpretable model. Third, some models can be crippled by predictors with degenerate distributions. In these cases, there can be a significant improvement in model performance or stability without the problematic variables.

Four attributes were removed from the dataset. The *invasive* attribute, that refers to the diagnosis of invasive or ductal carcinoma, was not considered due to the causality of correlation with the cancer attribute of interest. The *training* attribute suggests whether that record in the dataset is to be considered for training or validation. However, because of the next transformations to be described we cannot use this division of records, thus the attribute is removed. The *last_ mammogram* attribute indicates the result of the last mammogram taken before the index mammogram that relates to the cancer attribute. Since it only contains information about negative and false positive results, then, it can be removed without affecting our analysis. Finally, the *count* attribute is removed to keep only one example of the record and remove duplicates. After this operation the number of records and cases is the same. Table 3.7 shows the list of resulting attributes after attribute removal.

Table 3.7: Description of attributes of the *Risk Estimation Dataset v.2* after preprocessing.

No.	Attribute	Description	Values
1	menopause_new	Menopausal status	0 = premenopausal 1 = postmenopausal 2 = postmenopausal or age>=55
2	agegrp	Age (years) in 5-year groups	1 = 35 - 39 2 = 40 - 44 3 = 45 - 49 4 = 50 - 54 5 = 55 - 59 6 = 60 - 64 7 = 65 - 69 8 = 70 - 74 9 = 75 - 79 10 = 80 - 84
3	density	BI-RADS breast density codes	1 = almost entirely fat 2 = scattered fibroglandular densities 3 = heterogeneously dense 4 = extremely dense
4	race	Race	1 = white 2 = asian/pacific islander 3 = black 4 = native american 5 = other/mixed
5	hispanic	Patient is Hispanic	0 = no 1 = yes
6	bmi	Body mass index	1 = 10 - 24.99 2 = 25 - 29.99 3 = 30 - 34.99 4 = 35 or more
7	agefirst	Age at first birth	0 = age < 30 1 = age 30 or greater 2 = nulliparous
8	nrelbc	Number of first-degree relatives with breast cancer	0 = zero 1 = one 2 = two or more
9	brstproc	Previous breast procedure	0 = no 1 = yes
10	surgmeno_new	Type of menopause	0 = natural 1 = surgical 2 = not menopausal
11	hrt_new	Current hormone therapy	0 = no 1 = yes 2 = not menopausal

Table 3.7: Description of attributes of the *Risk Estimation Dataset v.2* after preprocessing (continued).

No.	Attribute	Description	Values
12	cancer	Diagnosis of invasive or ductal carcinoma in situ breast cancer	0 = no 1 = yes

3.3.4 Elimination of records with unknown values

Most of the attributes, as shown in Table 3.2, contain the unknown value. In general, we had three options for handling missing or unknown data: i) leave the records as they are, ii) apply imputation methods by substituting the missing data to complete the dataset or, iii) discard all records containing missing values. If we leave the missing data in the dataset, the feature selection methods and machine learning algorithms could get confused with these unknown values and be used as true values for the attributes. This would result in classification estimates with information that is not really known. The second option could add some bias to the data, and in health-related datasets this could lead to unreliable results. Even though the third option would discard a large number of records, after careful analysis we decided to remove all records containing one or more unknown values and work only with records containing true values. After this operation, out of the 181,903 records in the dataset (see Table 3.3), we are left with 25,251 records after this operation (see Table 3.8).

Table 3.8: Distribution of positive and non-cancer cases within the *Risk Estimation v.2* dataset after elimination of records with *unknown* values.

Breast Cancer Diagnosis	Records	Records (%)
Yes	1,053	4.17
No	24,198	95.83
Total	25,251	100

After removing records with unknown values, the resulting quality measures are shown in Table 3.9. It can be observed that if the quality measurement *missing* is decreased to 0.00%, then the *valid* values increase.

Table 3.9: Quality measurements after elimination of records with unknown values.

No.	Attribute	Missing	Infinite	ID-ness	Stability	Valid
1	menopause_new	0.00%	0.00%	0.01%	75.89%	24.10%
2	agegrp	0.00%	0.00%	0.04%	19.17%	80.79%
3	density	0.00%	0.00%	0.02%	36.68%	63.30%
4	race	0.00%	0.00%	0.02%	71.23%	28.75%
5	hispanic	0.00%	0.00%	0.01%	88.48%	11.51%
6	bmi	0.00%	0.00%	0.02%	35.10%	64.88%
7	agefirst	0.00%	0.00%	0.01%	56.96%	43.03%
8	nrelbc	0.00%	0.00%	0.01%	67.41%	32.58%
9	brstproc	0.00%	0.00%	0.01%	65.79%	34.20%
10	surgmeno_new	0.00%	0.00%	0.01%	44.12%	55.87%
11	hrt_new	0.00%	0.00%	0.01%	40.82%	59.17%
12	cancer	0.00%	0.00%	0.01%	95.92%	4.07%

3.4 Summary

Following the *KDD* process, in this chapter we searched and selected the dataset used in this thesis. The *Risk Estimation v.2* dataset proved to be the best option compared to other datasets found during the search. In addition, the prepocessing stage was performed, at the end of this stage the attributes' data types were converted to categorical; the values of three attributes were transformed and renamed (i.e., *menopause_new*, *surgmeno_new*, and *hrt_new*); four attributes were removed from the dataset (i.e., *invasive*, *training*, *last_mammogram* and *count*); and all records with *unknown* values were removed. The final list of attributes is presented in Table 3.7. The number of final records are found in Table 3.8. The quality measures of the attributes shown in Table 3.9. The resulting dataset will be used for the next chapter.

Chapter 4

Selection and Validation of Risk Factors

The previous chapter described the activities carried out to select and preprocess the dataset used in this research. This chapter describes the activities carried out for the *Data mining* and *Evaluation* stage of the *Knowledge Discovery in Databases* process. The methods identified to perform the feature selection and validation are implemented.

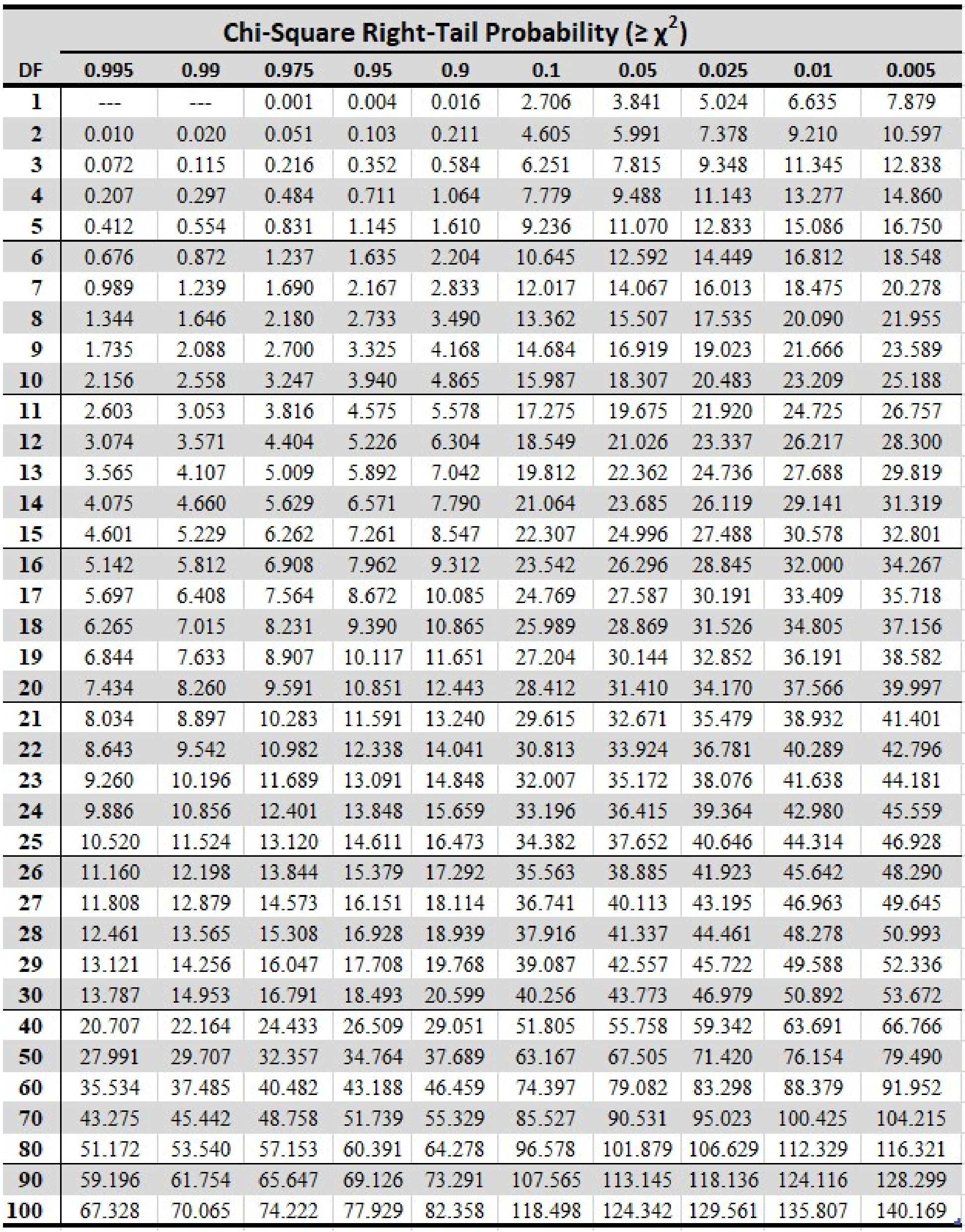
4.1 Feature Selection

The main purpose of this research work is to identify the relevant risk factors that could accurately predict whether a woman could get breast cancer or not. To determine which risk factors are more relevant, feature selection methods are used that involve statistical evaluations that calculate how strong the relationship between each attribute and the target variable is (where the target variable is the cancer attribute). To determine the ranking of attributes, this research makes use of two feature selection methods: *Chi-squared test* and *Mutual Information.*

4.1.1 Chi-squared test

As described in Chapter 2, in order to perform the *Chi-squared* test it is necessary to define three things; the hypothesis of the test, the degrees of freedom of each attribute and the value of p, which indicate the significance of the hypothesis test. The hypotheses for the test are defined below:

- Null Hypothesis (H0): Two variables are independent.
- Alternate Hypothesis (H1): Two variables are not independent.



	Chi-Square Right-Tail Probability (≥ χ^2)									
DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 4.1: Critical value in chi-squared (taken from [23]).

The degrees of freedom of each attribute are shown in Table 4.1 in the column named *D.F.* The value p is defined with a low value $p = 0.05$ in order for the results to be significant. After knowing the previous information, the critical values of each attribute were identified using Figure 4.1 (taken from [23]) and are shown in the column *C.V.* in Table 4.1. The column X^2 in Table 4.1 presents the chi-squared values obtained for each of the 11 risk factors within the dataset. The values are sorted in descending order. The higher the value of an attribute the more relevant it is considered. The last two columns in Table 4.1 show the test results for hypotheses *H0* and *H1*. Which are evaluated by the following two conditions:

$$X^2 > C.V \rightarrow Reject(H0) \wedge Accept(H1) \tag{4.1}$$

$$X^2 < C.V \rightarrow Accept(H0) \wedge Reject(H1) \tag{4.2}$$

Condition 4.1 indicates that if the value X^2 is greater than the critical value ($C.V.$) of the attribute then null hypothesis (*H0*) is rejected and alternative hypothesis (*H1*) is accepted. This means that the variables are not independent and there is a correlation between them, the correlation is established between the attribute of the risk factor evaluated and the variable that indicates that the person has breast cancer. Otherwise, condition 4.2 indicates that if the value X^2 is less than the critical value ($C.V.$) of the attribute then the null hypothesis (*H0*) is accepted and the alternative hypothesis (*H1*) is rejected. This means that the variables are independent and there is no relationship between the attribute of the risk factor evaluated and the variable that indicates if the person has breast cancer.

Table 4.1: Chi-squared results.

No.	Attribute	X^2	D.F	C.V.	H0	H1
1	agegrp	170.285	9	16.92	Reject	Accept
2	hrt_new	84.666	2	5.99	Reject	Accept
3	surgmeno_new	82.351	2	5.99	Reject	Accept
4	menopause_new	82.305	1	3.84	Reject	Accept
5	brstproc	49.162	1	3.84	Reject	Accept
6	density	40.555	3	7.81	Reject	Accept
7	nrelbc	21.018	2	5.99	Reject	Accept
8	hispanic	16.403	1	3.84	Reject	Accept
9	agefirst	6.721	2	5.99	Reject	Accept
10	race	4.455	4	9.49	Accept	Reject
11	bmi	1.373	3	7.81	Accept	Reject

The Table 4.1 shows that attributes from 1 to 9 are statistically significant at the 0.05 level. Only attributes 10 and 11 are not statistically significant. According to the obtained values the first four attributes could be considered as more relevant, i.e., the patient's age (*agegrp*), whether she had undergone hormone therapy (*hrt_ new*), her type of menopause (*surgmeno_ new*), and her menopausal status (*menopause_ new*). The next two attributes are also interesting, whether the patient have had a breast procedure (brstproc) and the patient's breast density (*density*). The rest of the attributes could be considered less relevant for this specific dataset.

4.1.2 Mutual information

Table 6 presents the values obtained from the *Mutual Information.* Again, the values are sorted in descending order. The higher the value of an attribute the more relevant it is considered. Here, as explained in Chapter 2, a threshold (cutoff) value was calculated in order to determine which attributes should be selected. Our threshold value was calculated by means of the standard deviation. For an attribute to be selected, its *Mutual Information* value must be greater than the threshold value S. In this case, only the first four attributes are greater than our calculated $S = 0.00022$. Notice that these four selected attributes are the same most relevant calculated by the *Chi-squared test.* The rest of the attributes have a similar ranking as given by the *Chi-squared test.*

Table 4.2: Mutual Information results.

No.	Attribute	Mutual Information	Test for independence
1	agegrp	0.000739	Accept
2	hrt_new	0.000398	Accept
3	surgmeno_new	0.000389	Accept
4	menopause_new	0.000389	Accept
5	brstproc	0.000202	Reject
6	density	0.000196	Reject
7	hispanic	0.000092	Reject
8	nrelbc	0.000085	Reject
9	agefirst	0.000031	Reject
10	race	0.000021	Reject
11	bmi	0.000006	Reject

4.1.3 Definition of subsets of relevant attributes

To synthesize and validate the results obtained by the *Chi-squared test* and *Mutual Information*, three subsets are defined based on the values given in the rankings of both methods as seen in Table 4.3.

Table 4.3: Description of the new attributes after being transformed.

Attribute	Values
Subset(4)	{agegrp, hrt_new, surgmeno_new, menopause_new}
Subset(7)	{*Subset(4)*, brstproc, density, nrelbc}
Subset(11)	{*Subset(7)*, Hispanic, agefirst, race, bmi}

Subset(4) contains the four risk factors ranked as the most relevant in both *Chi-squared test* and *Mutual Information*, corresponding to the patient's age (*agegrp*), whether she had undergone hormone therapy (*hrt_new*), her type of menopause (*surgmeno_new*), and her menopausal status (*menopause_new*).

Subset(7) contains all attributes of *Subset(4)* plus the next three attributes given by the *Chi-squared test*; whether the patient has had a breast procedure (*brstproc*), the patient's breast density (*density*), and whether she has first-degree relatives with breast cancer (*nrelbc*).

Finally, *Subset(11)* contains all the risk factors of the dataset. We need to consider all risk factors to validate the previous two subsets as will be seen in the next sections. The sets defined in this section shall be used at the *Validation* stage.

4.2 Risk Factors Validation

This section covers the *Validation* process, corresponding to the *Evaluation* stage of the *KDD* process. Because the two feature selection methods used in this research are statistical indicators focusing on the characteristics of the data itself, it is necessary to evaluate whether the attributes identified as the most relevant are actually significant and related to the effectiveness of a classification model. To achieve the evaluation, in this research it is proposed to perform the training of the seven classification algorithms selected (*Decision tree*, *Random tree*, *Decision stump*, *Deep learning*, *k-Nearest Neighbors (K-NN)*, *Naïve Bayes* and *Generalized linear model*) with the preprocessed dataset (with 11 attributes and an target attribute that indicates the incidence of breast cancer) to obtain the values of the performance metrics of each classifier. Subsequently, perform the training of the same classification algorithms but now, only with the attributes of the subsets generated in the previous section. The aim is to demonstrate that the most relevant risk factors resulting from the feature selection process are significant to obtain the values of performance metrics as if they were training with all the risk factors of the preprocessed dataset.

4.2.1 Imbalance classification problem

Before we try to validate our selected risk factors, it is necessary to tackle what is known as the *imbalance classification problem.* This type of problem occurs when the number of records of some class label is much larger than the other class. In other words, classes are not represented equally. Table 4.4 shown that the dataset used in our research has 99.27% of non-cancer cases versus 0.73% of positive cancer cases. This problem remains after the preprocessing phase described in Chapter 3, where all records with an unknown value were eliminated. The resulting dataset ended up with 95.83% of non-cancer records versus 4.17% of positive cancer records.

Table 4.4: Distribution of positive and non-cancer cases within the *Risk Estimation v.2* dataset before and after preprocessing.

Breast Cancer Diagnosis	Records before preprocessing	Records after preprocessing
Yes	7,319 (0.73%)	1,053 (4.17%)
No	1,000,341 (99.27%)	24,198 (95.83%)
Total	1,007,660 (100%)	25,251 (100%)

Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of a similar number of examples for each class. As a consequence of an unbalanced dataset, the models obtained have poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class. The imbalance class classification problem brings with it the *paradox of accuracy*, and it happens for example when in a dataset with 990 samples of a certain class "A" and only 10 of another class "B", an algorithm will learn that the best assumption you can make will be that every element is of class "A", since in this way you will get a 99% hit rate, this is known as *paradox of accuracy.* These results are not at all accurate considering that it has failed in predicting 100% of the elements that were class "B". Table 4.5 shows the results of training the seven classifiers selected with the complete preprocessed dataset without having solved the data imbalance problem.

Table 4.5: Performance metrics with the complete preprocessed dataset.

Model	Accuracy	Precision	Recall
Decision stump	95.83%	unknown	0.00%
Decision tree	95.83%	unknown	0.00%
Random tree	95.83%	unknown	0.00%
Deep learning	91.32%	19.82%	34.86%
Generalized linear model	95.83%	unknown	0.00%
Naïve Bayes	95.83%	unknown	0.00%
k-NN	95.65%	13.34%	0.57%

If only accuracy is observed in Table 4.5, it could be deduced that the results of the classifiers are good, since all values are greater than 90%. However, this is a classic example of the *paradox of accuracy*; the models classified all or most of the positive records as negative, that is, the model has failed in predicting 100% of the positive cancer records, and still obtained high accuracy. This can be easily confirmed by looking at the values of the other two metrics beginning with *precision.* Precision is defined by the Equation 4.3:

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

When models register a precision of *unknown* is because the sum of $TP + FP$ was equal to 0, that means that the model did not identify any record as positive, neither true positive *(TP)* nor false positive *(FP)*. Having no true positive *(TP)*, Recall gets a value of 0.00% since the dividend *(TP)* in Equation 4.4 is 0. A low recall indicates many false negatives *(FN)*.

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

The *precision* and *recall* values different to *unknown* and 0.00% obtained by *Deep learning* and *k-NN* indicate that these two classifiers were able to detect a small number of true positives (TP) despite the existing imbalance.

As shown by the above results, data imbalance is an obvious problem for the validation process and needs to be addressed. The key approaches presented in the literature to solve this problem are presented below. Then, the best option for this research is selected.

Key approaches for resolution of imbalanced problem

The problem of class imbalance has been actively addressed and several techniques to deal with this problem have been proposed. In the literature, three key approaches to the learning for resolution of imbalanced problem are defined [55]–[57]:

Data-level methods: This approach is geared towards matching the class distributions. The class distribution are being balanced using the sampling methods by resizing the training datasets. The sampling methods can be categorized into techniques for *under-sampling* and *over-sampling* [56].

- *Over-sampling:* The basic idea of over-sampling is to increase the size of the minority class to obtain balanced classes. Duplication of samples is done in random over-sampling in which samples are randomly selected. Thus, class size increases due to duplication of samples, as shown in Figure 4.2 (based on [57]).
- *Under-sampling:* Taking a random set of samples from the majority class to balance the classes and rest of the samples are ignored. The size of the data space is measured to obtain desirable class distribution ratio. Thus, under-sampling helps in gaining the equal number of class samples (as shown in Figure 4.3 based on [57]) and makes training phase faster.

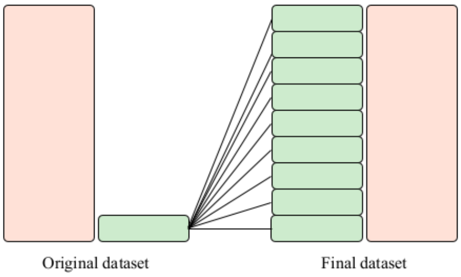


Figure 4.2: Example of over-sampling (based on [57]).

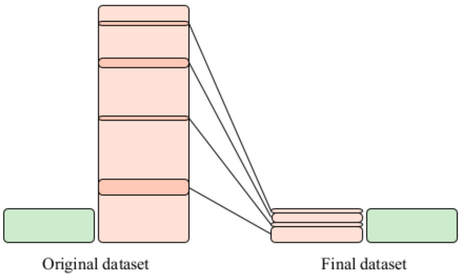


Figure 4.3: Example of under-sampling (based on [57]).

Algorithm-level methods: Improving the ability of current classifier algorithms for learning from minority classes. For example, adjustment of the estimation of probability or modification of cost per class may be favorable to the minority class [56].

Hybrid methods: Is the combination of data-level and algorithm-level approaches. The main idea behind this is to delete the noisy and unreliable samples to extract useful and consistent samples using methods of data-level and then, use algorithm-level methods to achieve good classification accuracy [56].

Data-level methods are easy to implement and more popular as compared to algorithm-level methods. But algorithm-level methods are more effective computational techniques [57]. Because in this research it is important to maintain the integrity of our dataset, we follow an algorithm-level approach by implementing an ensemble learning method, since in the other two approaches it is necessary to make modifications to the original data. Combining multiple classifiers into an ensemble is one of the most powerful approaches in modern machine learning, leading to improved predictive performance, generalization capabilities, and robustness. The following section describes the main assembly methods.

4.2.2 Ensemble learning methods

Ensemble learning [58] is a general approach to machine learning that seeks better predictive performance by combining predictions from multiple models. Although there is a wide variety of methods, there are three that dominate the field of ensemble learning: *Boosting* [59], *Stacking* [60] and *Bagging* [61].

Boosting

The Boosting method [59] involves the incremental and sequential construction of sub-classifiers based on a single machine learning algorithm. The training of each sub-classifier is executed by generating weighted samples emphasizing the misclassified instances in the previous sub-classifier. For the final prediction, a weighted average vote occurs. Figure 4.4 (based on [59]) shows a diagram of the boosting method.

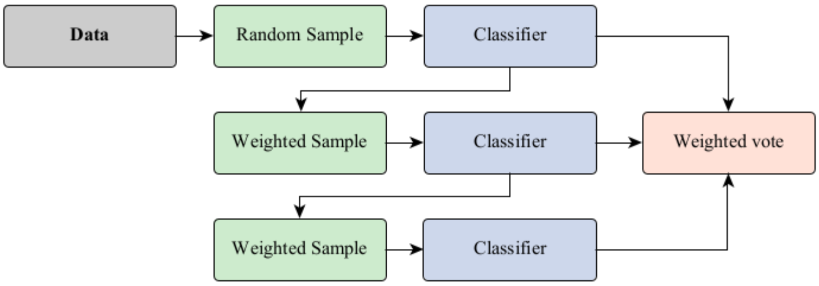


Figure 4.4: Boosting method (based on [59]).

Stacking

Stacking [60] is based on the creation of parallel sub-classifiers using distinct types of machine learning algorithms to strategically maximize the individual strengths of each of them. Penalized logistic regression is used to combine the sub-classifiers. Figure 4.5 (based on [60]) shows a diagram of the stacking method.

Bagging

The Bagging method [61] creates independent and parallel sub-classifiers with a single machine learning algorithm. First, from the initial data, several subsets of the same

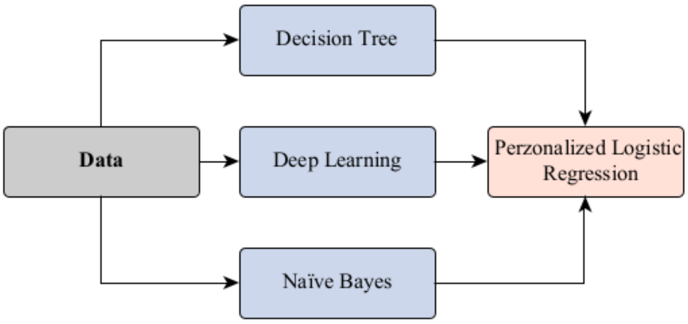


Figure 4.5: Stacking method (based on [60]).

size are generated, thus ensuring diversity and independence. Then, for each sample, a sub-classifier is constructed and, finally, using a majority vote the final classification is obtained. Figure 4.6 (based on [61]) depicts the diagram of how the bagging method operates.

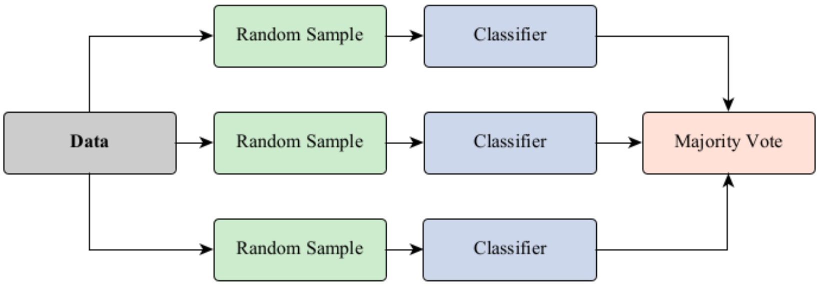


Figure 4.6: Bagging method (based on [61]).

Although the three ensemble methods try to improve the predictive performance of the classification problem, not all three methods could solve the class imbalance problem. *Stacking* uses the dataset as it currently is, thus the class imbalance problem would remain the same. *Boosting* could be used to solve class imbalance since it takes a random sample of data that could be constructed with balanced data. However, this method would have to be used several times to consider the remaining data, and the method does not consider this kind of situation. *Bagging*, on the other hand, provides a solution to the imbalance problem by using independent sub-classifiers trained with random samples that we can make sure they are balanced. For this reason, in this research *bagging* is implemented as a solution to data imbalance.

4.2.3 Bagging implementation

Following the *bagging* method, it was necessary to create a resampling of the data according to the cancer attribute. Figure 4.7 shows the process used to perform such resampling. From the dataset, after being preprocessed, twenty-three sample groups were randomly generated, combining all the positive cancer records with a subset of the same number of randomly selected non-cancer records. Since the dataset ended up with 1,053 positive cancer records after the preprocessing phase, each sample group contains that number of records plus a random selection of 1,053 non-cancer records, that results in 2,106 records per sample group.

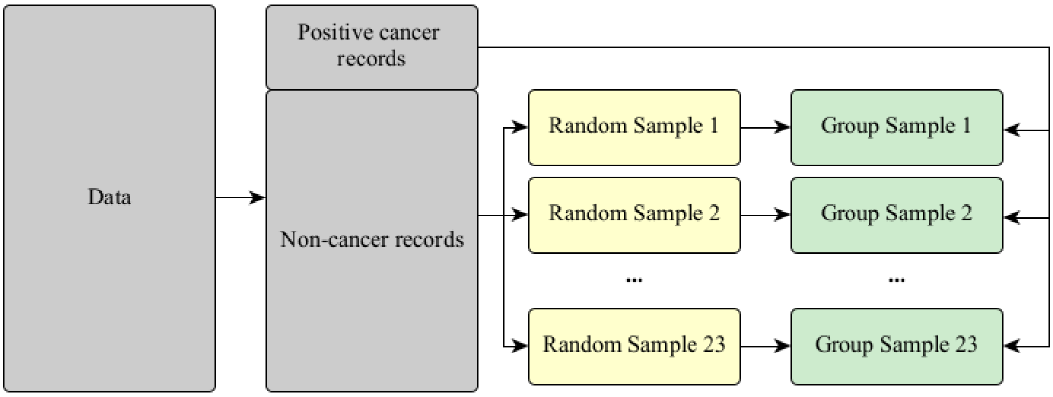


Figure 4.7: Resampling process for the class imbalance problem.

Subsequently, each of the samples generated were used to train the seven classification algorithms selected in this research. Through the training, a 10 fold cross validation was carried out to obtain the performance metrics of accuracy, precision, and recall. To obtain the final result, all classifiers of the same type generated from each of the sample groups were assembled, as shown in Figure 4.8.

Table 4.6 shows the results of the *bagging* implementation in the dataset with all attributes. The performance improvement in *precision* and *recall* metrics can be observed compared to those obtained in Table 4.5, where no ensemble technique was applied. Related to *precision*, before the *bagging* implementation, the models could not identify any record as positive; in contrast, after the bagging implementation was obtained a minimum *precision* of 87.79% with *random tree* and a maximum of 99.91% with *k-NN*. Regarding

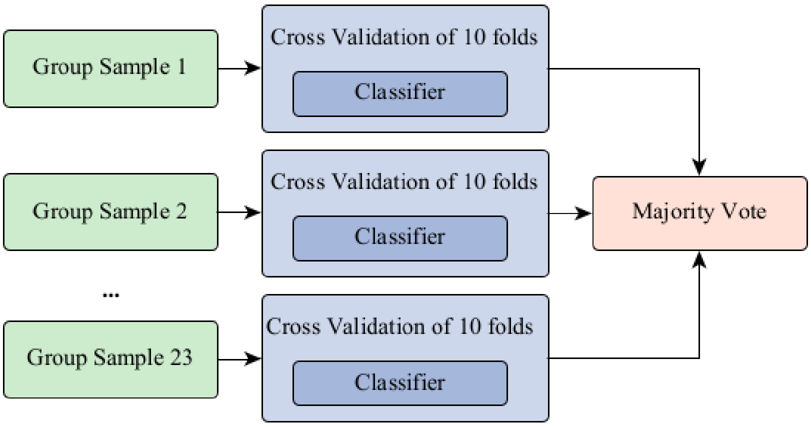


Figure 4.8: Ensemble process for the class imbalance problem.

recall, the value of 0.00% obtained before the *bagging* implementation was improved to obtain a minimum value of 62.67% with *k-NN* and a maximum of 95.12% with *Decision tree*. Obtaining average values above 90% with *bagging* implementation indicates that the models dramatically improved in the identification of positive breast cancer registries without sacrificing the identification of negative records, thus, it is shown that *bagging* is a good solution to data imbalance in this research. Now that the imbalance problem has been dealt with, the results of the risk factor validation with the defined subsets will be presented in the next section.

Table 4.6: Performance metrics with the complete preprocessed dataset after *bagging* implementation.

Model	Accuracy	Precision	Recall
Decision stump	86.32%	99.79%	72.83%
Decision tree	97.45%	99.77%	95.12%
Random tree	78.38%	87.79%	70.08%
Deep learning	97.21%	99.52%	94.88%
Generalized linear model	96.62%	99.71%	93.51%
Naïve Bayes	93.93%	98.70%	89.16%
k-NN	81.30%	99.91%	62.67%

4.2.4 Risk factor validation with the defined subsets

Section 4.1 (Feature Selection) defined two similar rankings for the risk factors within the dataset. The aim of identifying which risk factors are more relevant than others, is to use those relevant attributes to determine breast cancer cases, or at least, to pay more attention to those specific factors; in case not all attributes are available or could not be obtained. In this section, experiments will be performed to determine the predictive performance of the attribute subsets as defined in Table 4.3, i.e., *Subset(4)*, *Subset(7)*, and *Subset(11)*, where the latter will be used as baseline for the previous two subsets.

For our experiments, seven different algorithms were selected to cover multiple machine learning techniques: *Decision tree*, *Decision stump*, *Random tree*, *Deep learning*, *Generalized linear model*, *Naïve Bayes*, and *k-Nearest Neighbors (k-NN)*. All algorithms were executed considering the default settings given by *RapidMiner*. To validate each subset of attributes, the seven classification algorithms were trained only with the attributes that belong to the subset being evaluated. Also, a 10-fold cross validation was used to obtain the performance metrics of accuracy, precision, and recall.

Table 4.7 presents the results using bagging for the three subsets of attributes as defined in Table 4.3. Recall that *Subset(4)* contains the four most relevant risk factors found, *Subset(7)* contains those four attributes plus the next three risk factors as ranked by the *Chi-squared test*. Finally, *Subset(11)* contains all risk factors within the dataset, correspond to the baseline results obtained in Table 4.6.

The first thing to note is the column that refers to *Subset(11)*; this is our baseline, as it considers all attributes. The classifiers with the highest accuracy (Acc.) are Decision tree and Deep learning with 97.45% and 97.21% respectively, while the least accurate is Random tree with 78-38%. It is important to also consider the metrics of precision (Prec.) and recall (Rec.), that provide more information with regard of the classification of positive cancer cases. The higher the precision value the fewer false positives being classified. On the other hand, the higher the recall value the more positive records are classified correctly. In our experiments for *Subset(11)*, the precision values for all algorithms are high. However, the recall value for k-NN is low, which means that only 62.67% of the positive cancer cases were correctly classified. In terms of the three metrics, Decision tree, Deep learning, and Generalized linear model obtained the best results for all attributes.

In order to validate whether the selected attributes could be truly relevant in our study, we need to compare the results against those obtained as the baseline (*Subset(11)*). First,

Table 4.7: Performance metrics of the defined subsets of risk factors.

Algorithm	Metric	*Subset(4)*	*Subset(7)*	*Subset(11)*
Decision stump	Acc. Prec. Rec.	86.32% 99.79% 72.83%	86.32% 99.79% 72.83%	86.32% 99.79% 72.83%
Decision tree	Acc. Prec. Rec.	86.32% 99.79% 72.83%	96.18% 99.82% 92.53%	97.45% 99.77% 95.12%
Random tree	Acc. Prec. Rec.	85.24% 98.29% 72.15%	80.67% 94.39% 67.16%	78.38% 87.79% 70.08%
Deep learning	Acc. Prec. Rec.	93.32% 99.41% 87.19%	96.16% 99.65% 92.65%	97.21% 99.52% 94.88%
Generalized linear model	Acc. Prec. Rec.	92.87% 99.62% 86.09%	95.56% 99.68% 91.44%	96.62% 99.71% 93.51%
Naïve Bayes	Acc. Prec. Rec.	92.51% 98.77% 86.31%	93.87% 98.64% 89.11%	93.93% 98.70% 89.16%
k-NN	Acc. Prec. Rec.	93.10% 100.00% 86.20%	87.27% 99.94% 74.59%	81.30% 99.91% 62.67%

notice that Decision stump reported the same metrics for the three subsets. This is because the algorithm generates a Decision tree with only one division obtained from the evaluation of one of the most significant attributes. In our case, the algorithm chose the attributes of *agegrp* and *menopause_ new* as a single node, since both attributes are part of the three subsets then the results are the same. Although the results do not provide important information, the algorithm supports the relevance of these two attributes as stated in Section 4.1.

The Decision tree algorithm obtained good results with all the attributes (*Subset(11)*), and maintains its precision for all subsets. However, it is drastically affected using *Subset(4)*, as it loses 11% in accuracy and 22% in recall. On the other hand, *Subset(7)* maintains practically the same performance compared to the baseline.

In the case of Random tree we can see that *Subset(4)* performs somehow better than the baseline, with a 6% gain in accuracy and 10% gain in precision. Nevertheless, the recall in the three subsets is low, which means it would be better to consider other classifiers.

Deep learning is one of the best alternatives of the classifiers used in this research. It can be observed that it is quite stable for all subsets. Using *Subset(4)*, there is only around 4% loss in accuracy and 7% loss in recall. Although there is some loss, the results are better than other classifiers. For *Subset(7)*, the results are almost the same as the baseline. Similar results are obtained with the Generalized linear model, where *Subset(4)* loses the same percentage in accuracy and recall, and *Subset(7)* is very close to the baseline.

Although Naïve Bayes does not have results such as Deep learning, it could also be considered a stable model, since training the algorithm with *Subset(4)* has a loss of only 1% in accuracy and 3% in recall. Finally, the performance metrics of *Subset(4)* in k-NN are higher than the baseline, with an increase of 11% in accuracy and 23% gain in recall. This increment could mean that the selected attributes are indeed relevant.

After analyzing these results, it is possible to conclude that the four selected risk factors: the patient's age (*agegrp*), whether she had undergone hormone therapy (*hrt_new*), her type of menopause (*surgmeno_new*), and her menopausal status (*menopause_new*); are relevant for the classification of positive cancer cases. Also, the next three risk factors: whether the patient has had a breast procedure (*brstproc*), the patient's breast density (*density*), and whether she has first-degree relatives with breast cancer (*nrelbc*); should also be further analyzed.

4.3 Results Comparison

Based on the results obtained in this research, this section presents a comparison of the results and methods presented in five related works. Such a comparison is focused on key aspects of each work, in terms of the overall goal, the dataset, the feature selection methods, the classification methods, resampling techniques, and the obtained results. The following is a summary of the five papers selected in terms of the key aspects to be evaluated:

Fahrudin et al. [41]: Focused on determining breast cancer risk factors for patients in Indonesia and identified differences against patients in the United States, using a private dataset with 1907 records and 21 attributes (containing demographic and pathology and therapy information). They used three features selection methods: i) *Information gain*, ii) *Fisher's discriminant*, and iii) *Chi-squared test*, to select the best attributes (risk factors). They also applied *Hierarchical K-means* clustering to remove attributes that have the lowest contribution. Do not use classification methods or resampling techniques. As a result, out of the 21 original attributes, 14 relevant attributes were obtained .

Fu et al. [45]: Proposed a prognosis model framework to predict Invasive Disease-Free Survival (i.e., the length of time after the primary treatment ends and no signs of cancer appear again) for early-stage breast cancer patients. They used a private dataset with 12,119 records and 89 attributes of the Clinical Research Center for Breast (CRCB) from West China Hospital of Sichuan University. The attributes consist of demographic, diagnosis, pathology, and therapy information. A *Stratified Feature Selection* was used by calculating the importance score using five methods based on the type of the individual feature: i) Kolmogorov-Smirnov (KS) statistical test to feature with Interval scale, ii) The independent sample T-test to feature with notable influence on the 5-year iDFS is separated from others, iii) The Wilcoxon Mann-Whitney statistical test used to feature with Ordinal scale, but whose distribution is not normally distributed, and, iv) *Chi-squared test* for the Nominal scale feature. To predict the 5-year iDFS of breast cancer, the ensemble learning algorithm *Gradient boosting decision tree (XGBoost)* is used to construct the prediction model. Do not use resampling techniques. One of their results is a selection of 23 attributes, including some risk factors.

Kabir and Ludwig [44]: Focused on improve the classification performance of the standard machine learning algorithms towards the prediction of the important or minority class by using different resampling techniques (*random under-sampling*, *random over-sampling*, and a *hybrid of over- and under-sampling*) on a real-world breast cancer risk factors data set. They used the public Risk Factors dataset by the BCSC with 6,318,638 cases and 13 attributes. To evaluate the results of each of the resampling approaches, the authors used three different classification algorithms: *Decision tree, Random forest*, and *XGBoost*. Their results showed that performance improves when resampling techniques are used compared to when no techniques are applied.

Kabir et al. [43]: Generated risk factor rules by means of *Association Rule Mining (ARM)*, using the Breast Cancer Surveillance Consortium's (BCSC) Risk Factors dataset. This public dataset contains 6,318,638 cases and 13 attributes. The *Logit* model was used to select those factors that may affect the likelihood of breast cancer. Do not use resampling techniques. A set of 5 rules was obtained for breast cancer cases and 4 rules for non-cancer cases.

Li et al. [40]: Present a prevention and control system for breast cancer by means of *Item Rule Association (IRA)* algorithms applied on a private dataset with 2,966 records and 83 attributes. An important characteristic of their work is the creation of their own dataset by interviewing patients from 22 hospitals over a one-year period and storing clinical, personal, and socio-economical information. Do not use feature selection methods or resampling techniques. Three types of rules defining the more relevant risk factors were identified; 35 rules were obtained using a single factor, 19 rules were obtained combining two factors, and 9 rules were obtained combining three factors.

Table 4.8: Results Comparison.

Aspect	**Fahrudin et al. [41]**	**Fu et al. [45]**	**Kabir and Ludwig [44]**	**Kabir et al. [43]**	**Li et al. [40]**
Goal	*Similar*	Different	Different	Different	Different
Dataset	Different	Different	*Similar*	*Similar*	Different
Feature Selection	*Similar*	*Similar*	Not used	Different	Not used
Resampling	Not used	Not used	Different	Not used	Not used
Classification	Not used	*Similar*	*Similar*	Different	Different
Results	*Similar*	Different	Different	Different	Different

Table 4.8 summarizes the similarities and differences between the previous related works and this research. The main difference of this research with respect to all five is

that in this research makes use of feature selection methods, resampling techniques, and classification for validation. In particular, the use of ensemble methods, since Kabir and Ludwig [44] are the only ones that perform resampling but at the data-level, while we resample at the algorithmic-level. Our goal and that of Fahrudin et al. [41] are similar, in the sense that, based on risk factors we both try to predict the likelihood of cancer. They apply association rules, and we apply seven different classifiers. Also, they do not directly handle the class imbalance problem, they had to adjust the algorithm to try to compensate the positive cancer class. No classification nor resampling was performed in the process. Our work is similar in that we also use *Chi-squared test* and *Mutual information* for feature selection; however, we use resampling and classification methods for validation. The main difference between our work and [40] is the creation of their own dataset, that provides more information and control. Since rule association algorithms were implemented, there is no need to apply feature selection methods. Although the risk factors appearing in the obtained rules could be defined as being the most relevant, the authors did not explicitly specify their relevance.

It is difficult to make a comparison in terms of results, not just because of the methods being used, but mainly because the datasets and the type of information they contain. Datasets may contain clinical, personal, demographical, or pathological information. The availability of this information and the number of attributes of each type will affect the results we might obtain.

Chapter 5

Conclusions

Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has a direct influence on their daily practice and clinical service. In this research, it is proposed the study of risk factors for breast cancer as an alternative that has been investigated to create control and risk assessment strategies in women. The main objective of this research is to identify relevant risk factors that could accurately predict whether a woman can develop breast cancer or not. To construct the solution in this research, compared to other work done so far, this research analyzed three different elements: i) feature selection, ii) ensemble learning, and iii) classification algorithms. Our research explores feature selection techniques, namely *Chi-square test* and *Mutual information*, combined with an ensemble method (*Bagging*) to detect breast cancer cases with information on risk factors.

During the course of this investigation, we were able to see that the study of risk factors brings with it different advantages and disadvantages from a computationally point of view. One of the most notable advantages is that it is less expensive both medically and computationally compared to other types of breast cancer research. Because most of these investigations are dedicated to the processing and analysis of medical images (such as mammograms, ultrasounds or magnetic resonances), these studies represent a higher cost since specialized medical equipment is required to generate them. Computationally, it also represents a higher cost in terms of time and resources.

In contrast, most breast cancer risk factor information can be easily collected through an interview or a form, for the generation of this data, no specialized medical equipment is required. In computational terms, during this research we identified different challenges that arise. One of the main challenges identified from the early stages of this research was the limited number of public datasets on breast cancer risk factors. Additionally, the few datasets that were available contain information of a small number of factors, considering that the literature provides a long list of them. The computational challenges presented in this research are partly related to the deficiencies of datasets. One of them was to identify the optimal way to handle the dirty data (data noise, incomplete, inconsistent and missing values) in the dataset. In this sense, the *Knowledge Discovery in Databases* process is a methodology suitable for this type of problems, where it is not only sought to obtain knowledge of the data, but it is also important to ensure that the knowledge obtained comes from a reliable source with certain parameters of data quality.

Another major challenge is the unbalance of existing data in diseases such as breast cancer, this problem has attracted the attention of researchers in many other contexts, not only medical. In this research an ensemble method is proposed as a solution to the imbalance in breast cancer risk factor data, this alternative is one of the most supported and used in recent years in the literature. However, little has been used in similar works. Based on the results obtained in this research, it represents a good alternative when using algorithms of classification in breast cancer risk factor data.

Main Findings

Throughout this thesis, we found that the most relevant risk factors in breast cancer cases, according to the dataset analyzed, are the patient's age, whether she had undergone hormone therapy, her type of menopause, and her menopausal status. These four risk factors were validated by means of seven classification algorithms. We conclude that is possible to obtain a predictive performance similar to that obtained using all the 11 attributes of the dataset. It is still necessary to validate these results with medical experts in the field.

Future Work

Diverse aspects of the risk factors of breast cancer problem have been analyzed in this thesis, revealing future research opportunities to extend this research. The future work below arises as a response to the main challenges identified in this thesis:

1. Work with physicians to create a more complete risk factors dataset. This is probably one of the most important issues, if not the most important, to further advance our understanding in topics as relevant such as this. In order to have a better understanding of how certain risk factors affect certain populations. It is important to start generating the data sets with the appropriate characteristics to perform a deep and complete analysis and in this way, to create prevention and risk control strategies appropriate to the population.

2. To complement the previous point, it would later be sought to perform an analysis of the data similar to the one presented in this thesis and even explore the different objectives of discovery raised by KDD, as well as the methods used for the *data mining* stage.

3. Investigate, apply, and evaluate other data imbalance solution approaches (at the data level or hybrids) to determine which methods are useful for the breast cancer context.

References

[1] Global Cancer Observatory, *Cancer Today.* [Online]. Available: https://gco.iarc.fr/today/online-analysis-pie (visited on 06/25/2021).

[2] Cancer.Net, *Breast Cancer: Risk Factors and Prevention*, 2021. [Online]. Available: https://cancer.net/cancer-types/breast-cancer/risk-factors-and-prevention (visited on 11/25/2021).

[3] D. M. Ikeda and M. Kanae K., *Breast Imaging*, Third edit. Elsevier, 2017, p. 479, ISBN: 978-0-323-32904-0.

[4] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting Breast Cancer Recurrence Using Machine Learning Techniques", *ACM Computing Surveys*, vol. 49, no. 3, pp. 1–40, Dec. 2016, ISSN: 0360-0300. DOI: 10.1145/2988544. [Online]. Available: https://dl.acm.org/doi/10.1145/2988544.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework", in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, pp. 82–88.

[6] M. Kuhn and K. Johnson, *Applied Predictive Modeling.* New York, NY: Springer New York, 2013, ISBN: 978-1-4614-6848-6. DOI: 10.1007/978-1-4614-6849-3. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-6849-3.

[7] The American Cancer Society Medical and Editorial Content Team, *What is breast cancer?*, 2021. [Online]. Available: http://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html (visited on 06/25/2021).

[8] World Health Organization, *Breast cancer*, 2021. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/breast-cancer` (visited on 05/11/2022).

[9] The American Cancer Society Medical and Editorial Content Team, *ACS Breast Cancer Screening Guidelines*, 2022. [Online]. Available: `https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html` (visited on 05/11/2022).

[10] ——, *What Is a Mammogram?*, 2022. [Online]. Available: `https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms.html` (visited on 05/11/2022).

[11] ——, *Mammogram Results*, 2022. [Online]. Available: `https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/understanding-your-mammogram-report.html` (visited on 05/11/2022).

[12] G. Parmigiani, D. A. Berry, and O. Aguilar, "Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2", *American Journal of Human Genetics*, vol. 62, no. 1, pp. 145–158, 1998, ISSN: 00029297. DOI: `10.1086/301670`.

[13] D. J. Schaid, "Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history.", *Journal of the National Cancer Institute*, vol. 89, no. 21, pp. 1632–1634, 1997, ISSN: 00278874. DOI: `10.1093/jnci/89.21.1632-a`.

[14] A. C. Antoniou, A. P. Cunningham, J. Peto, *et al.*, "The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions", *British Journal of Cancer*, vol. 98, no. 8, pp. 1457–1466, 2008, ISSN: 15321827. DOI: `10.1038/sj.bjc.6604305`.

[15] M. H. Gail, L. A. Brinton, D. P. Byar, *et al.*, "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually", *JNCI: Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–1886,

Dec. 1989, ISSN: 0027-8874. DOI: 10.1093/jnci/81.24.1879. [Online]. Available: https://doi.org/10.1093/jnci/81.24.1879.

[16] J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors", *Statistics in Medicine*, vol. 23, no. 7, pp. 1111–1130, 2004. DOI: https://doi.org/10.1002/sim.1668. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1668.

[17] G. Meenalochini and S. Ramkumar, "Survey of machine learning algorithms for breast cancer detection using mammogram images", *Materials Today: Proceedings*, vol. 37, no. Part 2, pp. 2738–2743, 2021, ISSN: 22147853. DOI: 10.1016/j.matpr.2020.08.543. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214785320364257.

[18] I. Sechopoulos, J. Teuwen, and R. Mann, "Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art", *Seminars in Cancer Biology*, vol. 72, no. November 2019, pp. 214–225, Jul. 2021, ISSN: 1044579X. DOI: 10.1016/j.semcancer.2020.06.002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1044579X20301358.

[19] I. Syarif, "Dimensionality Reduction Algorithms on High Dimensional Datasets", *EMITTER International Journal of Engineering Technology*, vol. 2, no. 2, pp. 28–38, 2014. DOI: 10.24003/emitter.v2i2.24. [Online]. Available: https://emitter.pens.ac.id/index.php/emitter/article/view/24.

[20] L. Xie, Z. Li, Y. Zhou, Y. He, and J. Zhu, "Computational Diagnostic Techniques for Electrocardiogram Signal Analysis", *Sensors*, vol. 20, no. 21, p. 6318, Nov. 2020, ISSN: 1424-8220. DOI: 10.3390/s20216318. [Online]. Available: https://www.mdpi.com/1424-8220/20/21/6318.

[21] J. Brownlee, *How to Choose a Feature Selection Method For Machine Learning*, 2020. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ (visited on 02/25/2022).

[22] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes", in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, 1995, pp. 388–391. DOI: 10.1109/TAI.1995.479783.

[23] D. S. Shafer and Z. Zhang, *Introductory Statistics.* Saylor.org, 2010.

[24] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms.* USA: Cambridge University Press, 2002, ISBN: 0521642981.

[25] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest", *Journal of Big Data*, vol. 8, no. 1, p. 84, Dec. 2021, ISSN: 2196-1115. DOI: 10.1186/s40537-021-00472-4. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00472-4.

[26] J. R. Quinlan, "Induction of decision trees", *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986. DOI: https://doi.org/10.1007/BF00116251.

[27] W. Iba and P. Langley, "Induction of One-Level Decision Trees", in *Machine Learning Proceedings 1992*, Elsevier, 1992, pp. 233–240. DOI: 10.1016/B978-1-55860-247-2.50035-8. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9781558602472500358.

[28] T. K. Ho, "Random decision forests", in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, IEEE Comput. Soc. Press, 1995, pp. 278–282. DOI: 10.1109/ICDAR.1995.598994. [Online]. Available: http://ieeexplore.ieee.org/document/598994/.

[29] IBM Cloud Education, *What is deep learning?*, 2020. [Online]. Available: https://www.ibm.com/cloud/learn/deep-learning (visited on 11/10/2021).

[30] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models", *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972, ISSN: 00359238. DOI: https://doi.org/10.2307/2344614. [Online]. Available: http://www.jstor.org/stable/2344614.

[31] T. Cover and P. Hart, “Nearest neighbor pattern classification”, *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967. DOI: 10.1109/TIT.1967.1053964.

[32] D. V. Lindley, “Fiducial distributions and bayes’ theorem”, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 102–107, 1958. DOI: 10.1111/J.2517-6161.1958.TB00278.X.

[33] R. Kohavi and F. Provost, “Glossary of terms”, *Machine Learning*, vol. 2, pp. 271–274, Jan. 1998. DOI: 10.1023/A:1017181826899.

[34] A. Naik and L. Samant, “Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime”, *Procedia Computer Science*, vol. 85, pp. 662–668, 2016, ISSN: 18770509. DOI: 10.1016/j.procs.2016.05.251. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050916306019.

[35] A. K. Das, S. K. Biswas, A. Mandal, and M. Chakraborty, “A Neural Expert System to Identify Major Risk Factors of Breast Cancer”, in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, IEEE, Nov. 2020, pp. 1–4, ISBN: 978-1-7281-9744-9. DOI: 10.1109/INOCON50539.2020.9298261. [Online]. Available: https://ieeexplore.ieee.org/document/9298261/.

[36] Y. Khourdifi and M. Bahaj, “Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms”, in *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, IEEE, Nov. 2018, pp. 1–6, ISBN: 978-1-5386-7328-7. DOI: 10.1109/ISAECT.2018.8618688. [Online]. Available: https://ieeexplore.ieee.org/document/8618688/.

[37] H. Kutrani, S. Eltalhi, and N. Ashleik, “Predicting factors influencing survival of breast cancer patients using logistic regression of machine learning”, in *The 7th International Conference on Engineering & MIS 2021*, New York, NY, USA: ACM,

Oct. 2021, pp. 1–6, ISBN: 9781450390446. DOI: 10.1145/3492547.3492590. [Online]. Available: https://dl.acm.org/doi/10.1145/3492547.3492590.

[38] R. Dhanya, I. R. Paul, S. Sindhu Akula, M. Sivakumar, and J. J. Nair, "A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection", in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, May 2019, pp. 1049–1055, ISBN: 978-1-5386-8113-8. DOI: 10.1109/ICCS45141.2019.9065563. [Online]. Available: https://ieeexplore.ieee.org/document/9065563/.

[39] D. Jain and V. Singh, "Diagnosis of Breast Cancer and Diabetes using Hybrid Feature Selection Method", in *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, Dec. 2018, pp. 64–69, ISBN: 978-1-7281-0646-5. DOI: 10.1109/PDGC.2018.8745830. [Online]. Available: https://ieeexplore.ieee.org/document/8745830/.

[40] A. Li, L. Liu, A. Ullah, *et al.*, "Association Rule-Based Breast Cancer Prevention and Control System", *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 1106–1114, Oct. 2019, ISSN: 2329-924X. DOI: 10.1109/TCSS.2019.2912629. [Online]. Available: https://ieeexplore.ieee.org/document/8713389/.

[41] T. M. Fahrudin, I. Syarif, and A. R. Barakbah, "The determinant factor of breast cancer on medical oncology using feature selection based clustering", in *2016 International Conference on Knowledge Creation and Intelligent Computing (KCIC)*, IEEE, Nov. 2016, pp. 232–239, ISBN: 978-1-5090-5231-8. DOI: 10.1109/KCIC.2016.7883652. [Online]. Available: http://ieeexplore.ieee.org/document/7883652/.

[42] S. Maskery, Yonghong Zhang, Hai Hu, C. Shriver, J. Hooke, and M. Liebman, "Caffeine Intake, Race, and Risk of Invasive Breast Cancer Lessons Learned from Data Mining a Clinical Database", in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, vol. 2006, IEEE, 2006, pp. 714–718, ISBN: 0769525172. DOI: 10.1109/CBMS.2006.64. [Online]. Available: http://ieeexplore.ieee.org/document/1647655/.

[43] M. F. Kabir, S. A. Ludwig, and A. S. Abdullah, “Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining”, in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2018, pp. 2433–2441, ISBN: 978-1-5386-5035-6. DOI: 10.1109/BigData.2018.8622028. [Online]. Available: https://ieeexplore.ieee.org/document/8622028/.

[44] M. F. Kabir and S. Ludwig, “Classification of Breast Cancer Risk Factors Using Several Resampling Approaches”, in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2018, pp. 1243–1248, ISBN: 978-1-5386-6805-4. DOI: 10.1109/ICMLA.2018.00202. [Online]. Available: https://ieeexplore.ieee.org/document/8614227/.

[45] B. Fu, P. Liu, J. Lin, L. Deng, K. Hu, and H. Zheng, “Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data”, *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 7, pp. 2053–2064, Jul. 2019, ISSN: 0018-9294. DOI: 10.1109/TBME.2018.2882867. [Online]. Available: https://ieeexplore.ieee.org/document/8543186/.

[46] Z. Matjaz and S. Milan, *Breast cancer data set*, University Medical Center, Institute of Oncology, 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer.

[47] W. H. Wolberg, N. Street, and O. L. Mangasarian, *Breast cancer wisconsin (diagnostic) data set*, University of Wisconsin, General Surgery Dept, Computer Sciences Dept, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

[48] ——, *Breast cancer wisconsin (prognostic) data set*, University of Wisconsin, General Surgery Dept, Computer Sciences Dept, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(Prognostic).

[49] W. E. Barlow, E. White, R. Ballard-Barbash, *et al.*, “Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography”, *JNCI: Journal of the National Cancer Institute*, vol. 98, no. 17, pp. 1204–1214, Sep. 2006, ISSN:

1460-2105. DOI: 10.1093/jnci/djj331. [Online]. Available: http://academic.oup.com/jnci/article/98/17/1204/2521747/Prospective-Breast-Cancer-Risk-Prediction-Model.

[50] The American Cancer Society Medical and Editorial Content Team, *Types of Breast Cancer*, 2021. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html (visited on 06/25/2021).

[51] Breast Cancer Surveillance Consortium (BCSC), *Risk factors dataset*, 2017. [Online]. Available: https://www.bcsc-research.org/data/rf.

[52] ——, *Hormone therapy and breast cancer incidence*, 2003. [Online]. Available: https://www.bcsc-research.org/data/ht.

[53] ——, *Digital mammography dataset*, 2008. [Online]. Available: https://www.bcsc-research.org/data/mammography_dataset.

[54] World Health Organization, *A healthy lifestyle*, 2010. [Online]. Available: https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations (visited on 06/13/2022).

[55] G. Rekha, A. K. Tyagi, N. Sreenath, and S. Mishra, "Class Imbalanced Data: Open Issues and Future Research Directions", in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, Jan. 2021, pp. 1–6, ISBN: 978-1-7281-5875-4. DOI: 10.1109/ICCCI50826.2021.9402272. [Online]. Available: https://ieeexplore.ieee.org/document/9402272/.

[56] K. M. Hasib, M. S. Iqbal, F. M. Shah, *et al.*, "A survey of methods for managing the classification and solution of data imbalance problem", *Journal of Computer Science*, vol. 16, no. 11, pp. 1546–1557, Nov. 2020. DOI: 10.3844/jcssp.2020.1546.1557. [Online]. Available: https://thescipub.com/abstract/jcssp.2020.1546.1557.

[57] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning", *ACM Computing Surveys*, vol. 52, no. 4, pp. 1–36, Jul. 2020, ISSN: 0360-0300. DOI: 10.1145/3343440. [Online]. Available: https://dl.acm.org/doi/10.1145/3343440.

[58] N. Nilsson, *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*, 1965.

[59] R. E. Schapire, “Using Output Codes to Boost Multiclass Learning Problems”, in *Proceedings of the Fourteenth International Conference on Machine Learning*, ser. ICML ’97, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 313–321, ISBN: 1558604863. [Online]. Available: `https://www.researchgate.net/publication/2453554_Using_Output_Codes_to_Boost_Multiclass_Learning_Problems`.

[60] D. H. Wolpert, “Stacked generalization”, *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, ISSN: 08936080. DOI: `10.1016/S0893-6080(05)80023-1`. [Online]. Available: `https://linkinghub.elsevier.com/retrieve/pii/S0893608005800231`.

[61] L. Breiman, “Bagging Predictors”, *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996, ISSN: 1573-0565. DOI: `10.1023/A:1018054314350`. [Online]. Available: `https://doi.org/10.1023/A:1018054314350`.