UNIVERSITY NAME

MASTER THESIS

Data mining: a study of the factors of distribution and prevalence of cancer diseases in Annaba and East Algeria.

Author:
Chouia HOUSSEM

Supervisor:
Dr. Boulemdan AHMED

A thesis submitted in fulfillment of the requirements for the degree of Master's degree

in the
Department of Computer Science

July 10, 2021

Declaration of Authorship

I, Chouia HOUSSEM, declare that this thesis titled, "Data mining: a study of the factors of distribution and prevalence of cancer diseases in Annaba and East Algeria." and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"add some quote here"

Chouia Houssem

UNIVERSITY NAME

Abstract

Faculty of Engineering

Department of Computer Science

Master's degree

Data mining: a study of the factors of distribution and prevalence of cancer diseases in Annaba and East Algeria.

by Chouia HOUSSEM

Cancer is one of the major causes of death around the world. There were 18.1 million new cancer cases and 9.5 million cancer-related deaths globally in 2018. By 2040, the annual number of new cancer patients is estimated to reach 29.5 million, including 16.4 million cancer-related deaths. For effective clinical decision making in therapeutic and care methods, as well as reducing risks of undertreatment or overtreatment, accurate prognosis prediction for cancer patients is critical. In this work we have set a first objective to improve and optimize accuracy for the death prediction model made by "..." Our second work objective was to utilise data analysis techniques on breast cancer patients data collected by "..." as an individual work, Data analysis is the systematic use of statistical and/or logical methods to explain and demonstrate, summarize and analyze data. The precise and adequate interpretation of research results is a critical component of maintaining data integrity. ...

Acknowledgements

I would also like to thank my teachers for their support and the efforts made for my benefit, in particular my supervisor, “Dr. Boulemden Ahmed” for the support and for the encouragement during the realization of this work. I also thank “Pr. BOUZBID” and “Pr. ” and the CHU Cancer service of Annaba for their help and cooperation and their very valuable time. I would also like to thank members of the Jury and heads of our department. . . .

Contents

List of Figures

List of Tables

dedication . . .

Chapter 1

CANCER

1.1 Introduction

Cancer is one of the major causes of death around the world. There were 18.1 million new cancer cases and 9.5 million cancer-related deaths globally in 2018. By 2040, the annual number of new cancer patients is estimated to reach 29.5 million, including 16.4 million cancer-related deaths. In terms of new cancer cases in 2020, the most common were: 2.26 million cases of breast cancer; 2.21 million cases of lung cancer; 1.93 million cases of colon and rectum cancer; 1.41 million cases of prostate cancer; 1.20 million cases of skin cancer (non-melanoma); and 1.20 million cases of stomach cancer (1.09 million cases). For effective clinical decision making in therapeutic and care methods, as well as reducing risks of undertreatment or overtreatment, accurate prognosis prediction for cancer patients is critical. In this work we have set a first objective to improve and optimize accuracy for the death prediction model made by "..." Our second work objective was to utilise data analysis techniques on breast cancer patients data collected by "..." as an individual work, Data analysis is the systematic use of statistical and/or logical methods to explain and demonstrate, summarize and analyze data. The precise and adequate interpretation of research results is a critical component of maintaining data integrity.

1.2 Definitions

Cancer

A group of more than 100 different diseases that can begin almost anywhere in the body, characterized by abnormal cell growth and the ability to invade nearby tissues.

Disease-free survival (DFS)

The measure of time after treatment during which no sign of cancer is found. This term can be used for an individual or for a group of people within a study. This term is usually used in the context of scientific research.

Absolute risk

A disparity between two threats that is normally less than a relative risk in Cancers.

Acute

Symptoms that appear suddenly and escalate easily, but do not last for a long time.

Adjuvant therapy

Care provided after the main treatment to eliminate any remaining cancer cells and minimize the risk of cancer recurrence. Chemotherapy, radiation therapy, hormone therapy, and/or immunotherapy are commonly used after surgery.

Benign

Refers to a tumor that is not cancerous. The tumor does not usually invade nearby tissue or spread to other parts of the body.

Bone marrow

The soft, spongy tissue found in the center of large bones where blood cells are formed.

Bone marrow transplant

A medical procedure in which diseased bone marrow is replaced by healthy bone marrow from a volunteer donor.

Carcinoma

Cancer that starts in skin or tissues that line the inside or cover the outside of internal organs.

Case manager

A health care professional, often a nurse with experience in cancer, who helps coordinate the care of a person with cancer before, during, and after treatment. At a medical center, a case manager may provide a wide range of services for patients that may include managing treatment plans, coordinating health insurance approvals, and locating support services. Insurance companies also employ case managers.

Cells

The basic units that make up the human body, cancer happens on a cellular level.

Chemoprevention The use of natural, synthetic (made in a laboratory), or biologic (from a living source) substances to reverse, slow down, or prevent the development of cancer.

1.3 What makes a cell cancerous?

Cancer is unchecked cell growth. Mutations in genes can cause cancer by accelerating cell division rates or inhibiting normal controls on the system, such as cell cycle arrest or programmed cell death. As a mass of cancerous cells grows, it can develop into a tumor.

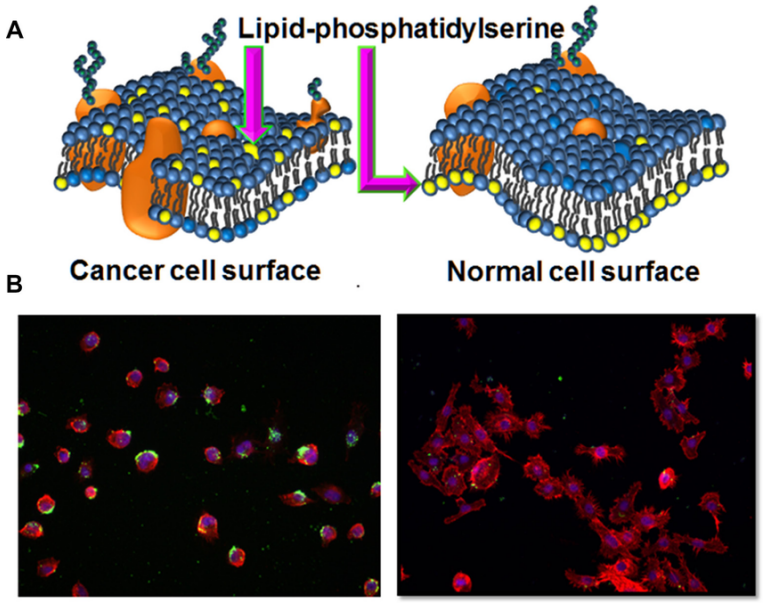

FIGURE 1.1: visual representation of the difference between normal and Cancerous cells [10]

1.4 Cancer cells vary from ordinary cells in lots of ways

For instance, most cancers cells:

- develop within the absence of indicators telling them to develop. Normal cells best develop after they acquire such indicators.
- forget about indicators that usually inform cells to forestall dividing or to die (a method referred to as programmed cell death, or apoptosis).
- invade into close by regions and unfold to different regions of the body. Normal cells forestall developing after they come upon different cells, and maximum ordinary cells do now no longer circulate across the body.
- inform blood vessels to develop closer to tumors. These blood vessels deliver tumors with oxygen and vitamins and eliminate waste merchandise from tumors. conceal from the immune machine. The immune machine usually removes broken or strange cells.

- trick the immune machine into supporting most cancers cells live alive and develop. For instance, a few most cancers cells persuade immune cells to shield the tumor in preference to attacking it.
- acquire a couple of modifications of their chromosomes, including duplications and deletions of chromosome parts. Some most cancers cells have double the ordinary variety of chromosomes.
- rely upon exceptional varieties of nutrients than ordinary cells. In addition, a few most cancers cells make power from vitamins in a exceptional manner than most ordinary cells. This we could most cancers cells develop extra quickly.
- Many times, most cancers cells depend so closely on those strange behaviors that they can't live on with out them. Researchers have taken gain of this fact, growing remedies that concentrate on the strange functions of most cancers cells. For example, a few most cancers remedies save you blood vessels from developing closer to tumors, basically starving the tumor of wanted nutrients.[29]

1.5 Benign vs malignant

After detecting the nodule a new problem arises. Radiologists need to determine whether a nodule is benign or malignant. To determine if a nodule is benign or malignant (the vast majority is benign) certain features of the nodule are checked.

1.6 types of cancer

There are in excess of 100 kinds of malignant growth. Sorts of disease are typically named for the organs or tissues where the malignancies structure. For instance, cellular breakdown in the lungs begins in the lung, and mind disease begins in the cerebrum. Malignant growths likewise might be portrayed by the kind of cell that framed them, like an epithelial cell or a squamous cell. [9]

1.6.1 Carcinoma

Carcinomas are the maximum famous kind of malignancy. They are framed through epithelial cells, that are the cells that cowl inside and outdoor surfaces of the frame. There are several styles of epithelial cells, which frequently have a segment like form whilst visible below a magnifying instrument.

Carcinomas that begin in diverse epithelial cell types have specific names:

Adenocarcinoma is a malignancy that systems in epithelial cells that produce drinks or physical fluid. Tissues with this form of epithelial cell are occasionally known as glandular tissues. Most malignancies of the bosom, colon, and prostate are adenocarcinomas.

Basal cell carcinoma is a malignancy that begins offevolved withinside the decrease or basal (base) layer of the epidermis, that's an individual's outside layer of pores and skin.

Squamous cell carcinoma is a malignant boom that systems in squamous cells, that are epithelial cells that falsehood simply beneathneath the outside floor of the pores and skin. Squamous cells likewise line severa one of a kind organs, along with the stomach, digestion tracts, lungs, bladder, and kidneys. Squamous cells appearance level, just like fish scales, whilst visible below a magnifying lens. Squamous cell carcinomas are right here and there known as epidermoid carcinomas.

Momentary cell carcinoma is a malignancy that systems in a form of epithelial tissue known as transient epithelium, or urothelium. This tissue, that's made from severa layers of epithelial cells that could get extra and extra modest, is observed withinside the linings of the bladder, ureters, and a part of the kidneys (renal pelvis), and a pair of various organs. A few tumors of the bladder, ureters, and kidneys are transient cell carcinomas.

1.6.2 Sarcomas

Sarcomas are illnesses that shape in bone and sensitive tissues, along with muscle, fat, veins, lymph vessels, and stringy tissue (like ligaments and tendons).

1.6.3 Osteosarcoma

Osteosarcoma is the maximum extensively diagnosed malignancy of bone. The maximum extensively diagnosed styles of sensitive tissue sarcoma are leiomyosarcoma, Kaposi sarcoma, dangerous stringy histiocytoma, liposarcoma, and dermatofibrosarcoma protuberans.

1.6.4 Leukemia

Tumors that begin withinside the blood-shaping tissue of the bone marrow are known as leukemias. These malignant growths do not body robust tumors. All matters being equal, large portions of weird white platelets (leukemia cells and leukemic effect cells) increase withinside the blood and bone marrow, swarming out common platelets. The low diploma of everyday platelets could make it tougher for the frame to get oxygen to its tissues, manage dying, or struggle illnesses.

There are 4 fundamental styles of leukemia, that are amassed depending on how hastily the contamination deteriorates (severe or constant) and at the

form of platelet the malignant boom starts offevolved in (lymphoblastic or myeloid). Intense styles of leukemia increase hastily and continual systems increase all of the extra steadily.

1.6.5 Lymphoma

Lymphoma is malignant boom that begins offevolved in lymphocytes (T cells or B cells). These are contamination scuffling with white platelets which are critical for the secure framework. In lymphoma, odd lymphocytes increase in lymph hubs and lymph vessels, simply as in one of a kind organs of the frame.

1.6.6 There are essential styles of lymphoma

Hodgkin lymphoma – People with this infection have uncommon lymphocytes which are known as Reed-Sternberg cells. These cells in most cases shape from B cells.

Non-Hodgkin lymphoma – This is a large accumulating of malignancies that starting in lymphocytes. The malignancies can increase hastily or steadily and may body from B cells or T cells.

1.6.7 Various Myeloma

Various myeloma is malignancy that begins offevolved in plasma cells, some other kind of secure cell. The uncommon plasma cells, known as myeloma cells, increase withinside the bone marrow and shape tumors in bones for the duration of the frame. Different myeloma is moreover known as plasma cell myeloma and Kahler sickness.

1.6.8 Melanoma

Melanoma is malignant boom that begins off evolved in cells that end up melanocytes, that are precise cells that make melanin (the coloration that offers pores and skin its tone). Most melanomas shape at the pores and skin, but melanomas can likewise form in different pigmented tissues, just like the eye.

1.6.9 Mind and Spinal Cord Tumors

There are diverse styles of cerebrum and spinal string tumors. These tumors are named depending on the form of cell wherein they framed and wherein the tumor initially formed withinside the focal sensory system. For instance, an astrocytic tumor begins offevolved in star-fashioned synapses known as astrocytes, which assist maintain nerve cells solid. Cerebrum tumors may be amiable (now no longer disease) or dangerous (malignant boom).[29]

1.7 Cancer Diagnosis

In most circumstances, a biopsy is required to detect cancer. A biopsy is a technique in which a sample of tissue is removed by the doctor. A pathologist examines the tissue under a microscope and does additional tests to determine whether it is cancerous.

A pathologist's findings are described in a pathology report, which includes information regarding your diagnosis. Pathology reports are crucial in identifying cancer and determining therapy options. Find out more about pathology reports and the types of data they contain.

1.8 Cancer stages and deadliness

Staging is crucial because it informs your treatment team about the treatments you require.

If your cancer is only in one location, your doctor may propose a local treatment such as surgery or radiotherapy. This could be enough to entirely eradicate the malignancy. A local treatment focuses on a specific part of the body.

If your cancer has spread, though, you may require medication that circulates throughout your entire body.

1.9 Importance of proper Diagnosis

Despite growing focus to EOL care quality, cancer patients continue to undergo aggressive treatment approaching death and have poor EOL results. Although 80 percent of patients prefer to die at home, the majority die in hospitals without having had a prior talk about their goals of care, and roughly a quarter of cancer patients receive chemotherapy within two weeks of death. Multiple factors influence outcomes.

Poor physician prognostic awareness and optimism bias are two examples. Patients and practitioners may focus on the explanation of advanced cancer as treatable but not curable as therapeutics advance, but they may overlook the more difficult sessions of how declining functional status and the biological limits of late-line therapies will inevitably limit the value of additional treatments.[3]

1.10 Early diagnosis of Cancer stage

One of the most important indicators of the efficiency of cancer services is cancer survival. Survival rates reflect both the system's ability to diagnose disease and whether people have quick access to adequate treatment. There is currently a substantial disparity in cancer survival rates due to a variety of patient-level, therapeutic, and biological factors.

Diagnosis of cancer at an earlier stage in its progression is linked to better outcomes and chances of survival. Reduced cancer waiting times, such as the time individuals wait to see a specialist after an urgent referral from a GP, or the time people wait for diagnostic testing, can help improve early diagnosis.

It can also be enhanced by public health initiatives such as screening programs and public awareness campaigns. For eight common malignancies, we look at the proportion of tumours diagnosed at an early stage in this indication.

The survival rates of cancer patients varies by country. This could be related to disparities in access to care and diagnostic and treatment delays, but it could also be due to population-level causes. We compared the UK's five-year survival rates for breast, cervical, and colon cancer with those of 17 other nations using data from the Organisation for Economic Co-operation and Development (OECD).

1.11 The 5 year relative survival

Cancer survival can be utilized as a population-level predictor of cancer prognosis as well as the efficacy of treatments.

Relative survival refers to the likelihood of remaining alive for a specific period of time following diagnosis when compared to the overall population's experience. The metric '5-year relative survival at diagnosis' (hereinafter referred to as '5-year survival') answers the question "what is the likelihood that an individual would survive their cancer for 5 years after a cancer diagnosis?"

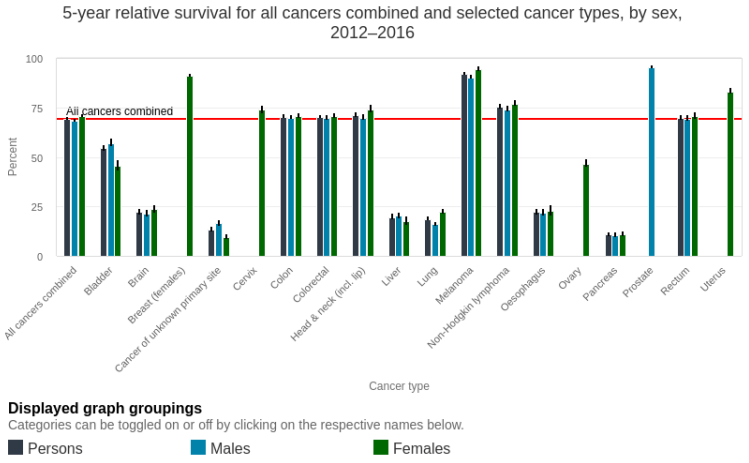

FIGURE 1.2: 5 year relative survival for all cancers combined in Australia

1.12 Cancer deaths in Data and numbers in UK as an example

The most recent data presented here for cancer stage at diagnosis are from December 2020, but the latest data for cancer survival rates are from before the coronavirus (Covid-19) pandemic. The pandemic caused considerable disruption to cancer services, which is likely to affect cancer survival, but estimating the impact is difficult.

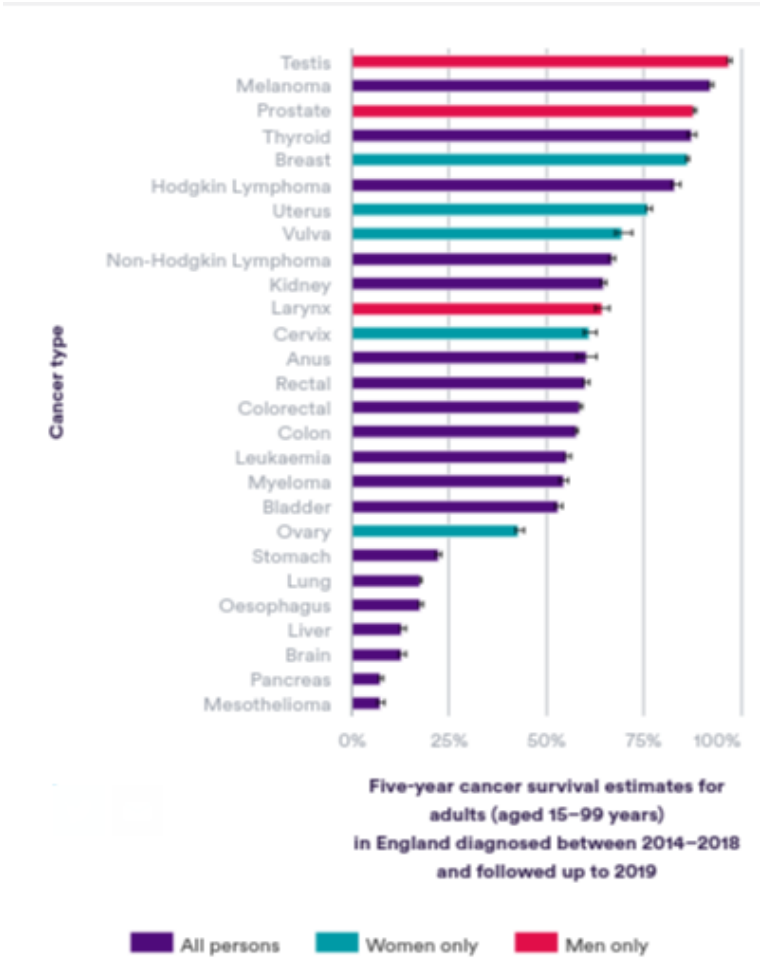

FIGURE 1.3: 5 years survival estimation for adults (aged 15-99) years in England diagnosed between 2014-2018 followed up to 2019

The five-year net survival estimates for adults (aged 15-99 years) in England diagnosed with one of the 27 most frequent malignancies between 2014 and 2018, and tracked up to 2019, are presented here.

Mesothelioma (7.2 percent), pancreatic cancer (7.3 percent), and brain cancer had the lowest five-year survival estimates (12.8 percent). Patients with testicular cancer (97 percent), cutaneous melanoma (92.3 percent), and prostate cancer have the best five-year survival rates (88 percent).

1.13 Cancer globally as Pandemic

In every country, cancer is a primary cause of death and a significant impediment to extending life expectancy. According to World Health Organization (WHO) figures, cancer is the first or second major cause of death before the age of 70 in 112 of 183 nations, and it ranks third or fourth in another 23. [32] The rise of cancer as a leading cause of death is partly due to significant improvements in stroke and coronary heart disease mortality rates in several nations when compared to cancer.

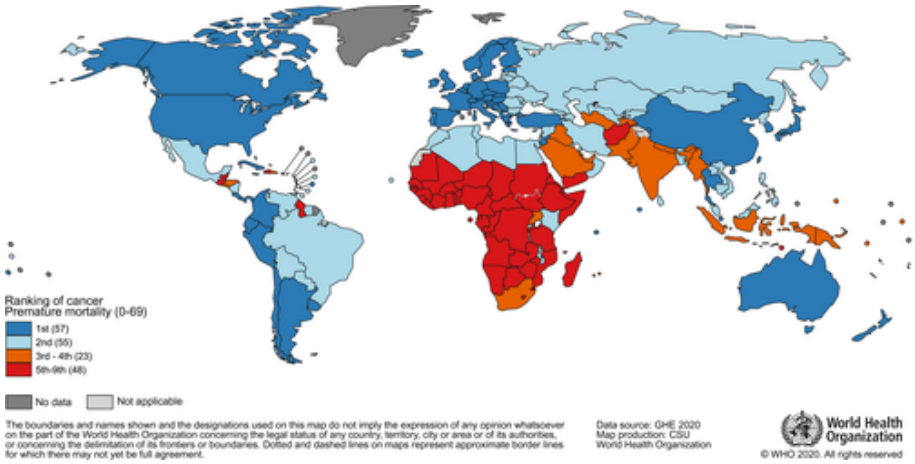

FIGURE 1.4: National Ranking of Cancer as a Cause of Death at Ages <70 Years in 2019. The numbers of countries represented in each ranking group are included in the legend. Source: World Health Organization.

Overall, the global burden of cancer incidence and death is quickly increasing, reflecting both population aging and growth, as well as changes in the prevalence and distribution of the key cancer risk factors, many of which are linked to socioeconomic development[26] [13]. Comparing the maps in Figures 1.6 and 1.4, the latter exhibiting the 4-tier Human Development Index (HDI) based on the United Nations' 2019 Human Development Report, to illustrate how much cancer's position as a cause of premature death correlates national levels of social and economic development [12].

1.14 Global and Local Cancer Patterns

The most commonly diagnosed cancers and leading causes of cancer death, respectively, by sex at the national level. The maps reveal substantial global diversity in leading cancer types, particularly for incidence in men (8 different cancer types) and for mortality in both men (8 types) and women (7 types). In men, prostate cancer is the most frequently diagnosed cancer in 112 countries, followed by lung cancer in 36 countries, and colorectal cancer and liver cancer each in 11 countries (Fig. 5A). With regard to mortality (Fig. 6A), lung cancer is the leading cause of cancer death in men in 93 countries, in

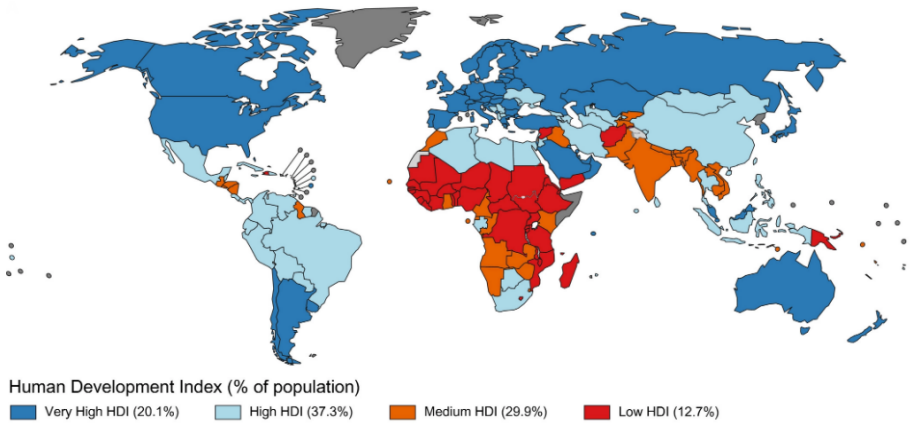

FIGURE 1.5: human development index 2019

part because of its high fatality rate,21 followed by prostate cancer (48 countries) and liver cancer (23 countries). In contrast to men, the most commonly diagnosed cancer in women is dominated by 2 cancer sites: breast cancer (159 countries) and cervical cancer (23 of 26 remaining countries) (Fig. 5B). The mortality profile in women is more heterogeneous (Fig. 6B), with breast and cervical cancer the leading causes of cancer death in 110 and 36 countries, respectively, followed by lung cancer in 25 countries.

1.15 Cancer in Algeria

According to the world health organisation (WHO), in 2020 Algeria have recorded 58.418 new case. 12 536 (21.5%) happened to be Breast Cancer which happens to be the leading cancer death cause for females in Algeria[33].

Also according to WHO, Algeria have recorded 32.802 Cancer death (the number is just an estimate). **the WHO claims there's no data About Cancer Mortality in Algeria** and 32.802 is a calculated estimate using data from neighboring countries.

Secondly we observe that the most frequent types of cancers for males are : Lung, Prostate, Colorectum, Bladder ,Stomach..

and for women : Breast, Colorectum, Thyroid, Cervix uteri, Ovary..

Globally for men we find that Prostate is more common unlike what we have here nationally and I personally think that's something to look into.

Incidence Data was collected from: Tumour Registry of Algiers, Annaba Cancer Registry, Cancer Registry of the Wilaya of Batna, Sétif Cancer Registry, Cancer Registry of Sidi-BelAbbès, Tizi-Ouzou Cancer Registry, Tlemcen Cancer Registry.

As for Mortality data they claim there isn't any.

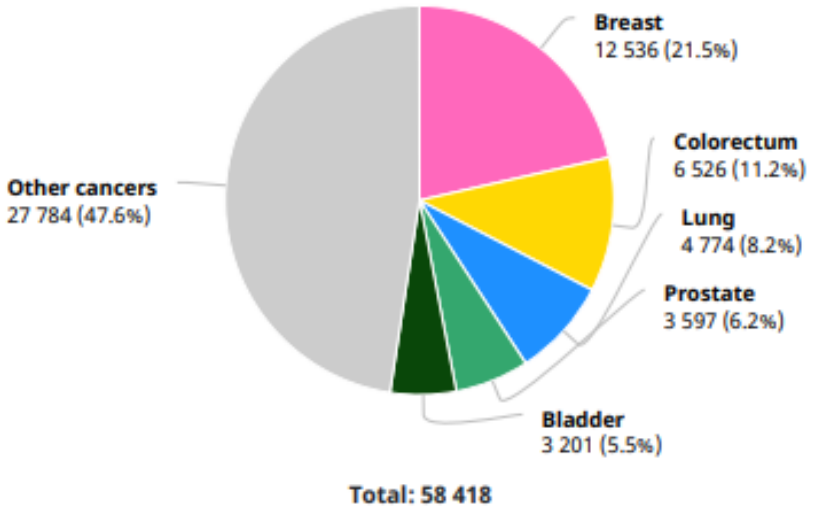

FIGURE 1.6: Number of new cases in 2020, both sexes, all ages

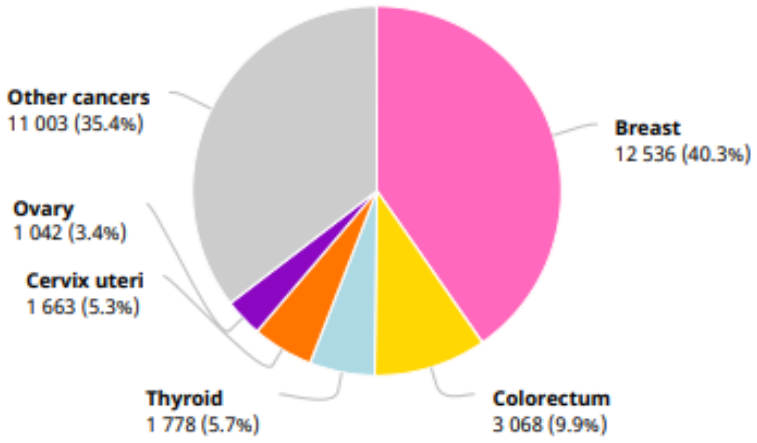

FIGURE 1.7: Number of new cases in 2020, female, all ages

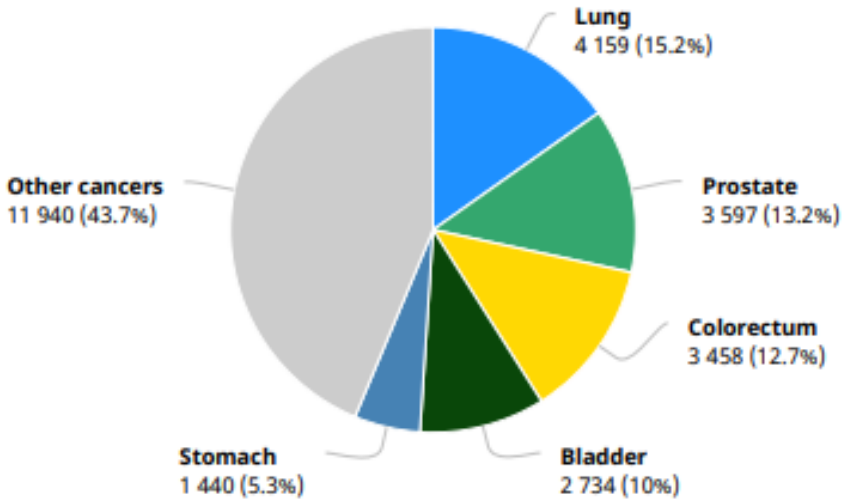

FIGURE 1.8: Number of new cases in 2020, male, all ages

Chapter 2

State of the Art

2.1 Introduction

Data mining is a scientific method of figuring out and coming across hidden styles and records in a massive dataset. Data evaluation is a subset of information mining, which includes studying and visualizing information to derive conclusions approximately beyond occasions and use those insights to optimize destiny outcomes.

Data analysis is described as a process of cleaning, transforming, and modeling data to discover beneficial information for commercial enterprise selection-making. The motive of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

In this Chapter we are going to explore state of the art, and strongly recent work done with different data mining and data analysis techniques in medical statistical data and some other overlapping fields.

2.2 Data sources

2.3 Data mining

2.3.1 What is Data mining?

Data mining is defined as a process used to extract usable data from a larger set of raw data. doctors can learn more about their patients and more about diseases and develop more effective strategies related to various business functions, allowing them to leverage resources more optimally and wisely. This helps us get closer to their goals and make better decisions. Data mining involves efficient data collection, data warehousing, and computer processing. To segment data and assess the probability of future events, data mining uses sophisticated mathematical algorithms.

2.3.2 Data mining process the Medical field

as we can see in figure 2.1 the Data mining process starts with

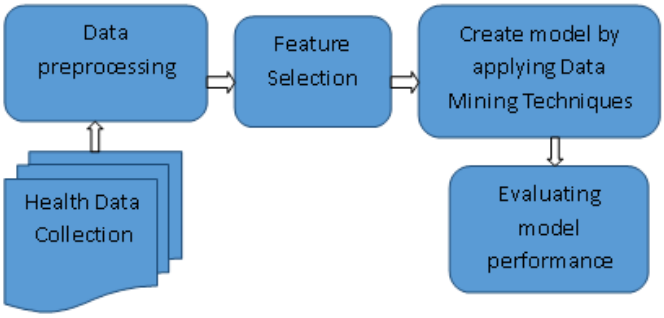

FIGURE 2.1: Visual Representation of the Data Mining process in Medical Field [28]

2.3.2.1 Health Data Collection

Data in the healthcare and the medical field in collected in different ways and methods, the aim and goal is to collect not only as much data as possible, but to also to have that data as accurate and as relevant and up to date as possible, with fewer entry errors, some of the sources for data in health are :

Household surveys

Household surveys are questionnaires distributed to a sample of a population's households. Their main benefit is that they provide the interviewer a lot of leeway when it comes to the information they need from responders. The information provided by the respondent is frequently erroneous (response error), and in many situations, the desired information is not provided at all.

Another option is to employ stratified samples, which are often based on tax returns and oversample the wealthy. According to studies, the wealthy had far higher rates of response inaccuracy and nonresponse than the middle class. Furthermore, 'weighting' the sample to represent the actual population distribution is problematic. Extensive use of household surveys has been made.

Medical Records Review

When you go to a doctor, a lot of data is collected, stored, processed, analyzed, or disseminated. A big chunk of that data may involve numbers related to your health, such as your heart rate during a particular visit. But there is potentially more to it than that.

In the study conducted by researchers at Monash Health, Allied Health Research Unit in Melbourne Victoria, Australia [30], the goal of the study was to determine if administrative data from electronic patient management programs are an effective data collection method for key hospital outcome measures when compared with alternative hospital data collection methods.

Only 376 (69%) of the 542 inpatient episodes they obtained from the hospital administrative computerized patient management program were manually

collected from ward-based sources. For both length of stay (93.4%) and discharge destination (91%) data, administrative data from the computerized patient management software exhibited the highest levels of agreement with inpatient medical record review.

2.3.2.2 Data Preparation (Data Preprocessing)

In the data mining process, data pretreatment is critical. In data mining and machine learning initiatives, the adage "garbage in, trash out" is especially pertinent. Data collection methods are frequently uncontrolled, resulting in out-of-range values (for example, Income: 100), impossible data combinations (for example, Sex: Male, Pregnant: Yes), and missing values, among other things.

Knowledge discovery during the training phase is more challenging if there is a lot of irrelevant and duplicated information or noisy and inaccurate data. The stages of data preparation and filtering can take a long time to complete. Cleaning, instance selection, normalization, transformation, feature extraction and selection, and so on are all examples of data preparation.

2.3.2.3 Feature Selection

Machine learning relies heavily on feature selection. The process of minimizing the number of inputs for processing and analysis, or selecting the most significant inputs, is referred to as feature selection. The practice of collecting relevant information or characteristics from existing data is referred to as feature engineering (or feature extraction).

Feature Selection's importance

For a variety of reasons, feature selection is crucial to the development of a good model for a variety of reasons, feature selection is crucial to the development of a good model. it necessitates some degree of cardinality[1] reduction, i.e., a limit on the number of characteristics that can be considered while constructing a model. Data nearly usually contains either too much or the wrong kind of information for the model to be built.

2.3.2.4 Create model by applying Data Mining Techniques

A mining model is created by applying an algorithm to data, but it is more than an algorithm or a metadata container: it is a collection of data, statistics, and patterns that can be used to forecast and infer relationships from fresh data.

[1]The cardinality of a set is a measure of a set's size, meaning the number of elements in the set. The cardinality of a set is a measure of a set's size, meaning the number of elements in the set. For instance, the set A = 1 , 2 , 4 A = 1,2,4 A=1,2,4 has a cardinality of 3 for the three elements that are in it.

Mining Model Architecture

A data mining model obtains information from a mining structure and analyzes it using a data mining algorithm. The mining model and the mining structure are two independent objects. The data source is defined by the information stored in the mining structure. A mining model saves information produced from statistical data processing, such as patterns discovered through analysis.

Until the data produced by the mining structure has been processed and examined, a mining model is empty. After a mining model has been processed, it is returned to the mining structure with metadata, results, and bindings.

2.3.2.5 Evaluating Model performance + validation

We have designed and created the data mining model. After that comes the evaluation and validation phase where we measure the performance of the model we created on the given problem:

2.3.2.6 The train/test/validation split

The most important thing you can do to ensure that your model is appropriately evaluated is to avoid training it on the full dataset. Normally, 70 percent of the data is used for training and 30 percent is used for testing.

2.3.2.7 Classification metrics (Accuracy, Precision, and Recall)

Accuracy is the quintessential classification metric. It is pretty easy to understand. And easily suited for binary as well as a multiclass classification problem.

Accuracy

The most basic categorization metric is accuracy. It's really simple to comprehend. And it's well-suited to both binary and multiclass classification problems.

Accuracy is the proportion of true results among the total number of cases examined.

For classification problems that are well balanced and not skewed or have no class imbalance, accuracy is a suitable choice of evaluation.

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

Precision

Precision, essentially is answering the following question: "what proportion of predicted Positives is truly Positive?"

$$Precision = \frac{(TP)}{TP + FP}$$

Recall

Another very useful measure is recall, which answers a different question: what proportion of actual Positives is correctly classified?

$$Recall = \frac{(TP)}{(TP + FN)}$$

Bias vs Variance

Machine learning algorithms employ mathematical or statistical models with two types of inherent errors: reducible and irreducible. Natural variability within a system causes irreducible error, or inherent uncertainty. Reducible error, on the other hand, is more controllable and should be minimized to ensure greater accuracy. Reduced error is made up of two components: bias and variance. To reduce errors, you'll need to choose models with the right level of complexity and flexibility, as well as the right training data. To reduce error and build accurate models, we must fully comprehend the differences between bias and variance.

K-Fold Cross-Validation

Cross-validation[5] is a resampling technique for evaluating machine learning models on a small sample of data.

The procedure has only one parameter, k, which specifies the number of groups into which a given data sample should be divided. As a result, the procedure is frequently referred to as k-fold cross-validation. When a specific value for k is chosen, it can be substituted for k in the model's reference, for example, k=10 for 10-fold cross-validation.

Cross-validation is a technique used in applied machine learning to estimate a machine learning model's skill on unknown data. That is, to use a small sample to estimate how the model will perform in general when used to make predictions on data that was not used during the model's training.

Performance metrics for imbalanced datasets

There are more meaningful performance measures for imbalanced datasets than Accuracy, in the case of imbalanced datasets **we pay closer attention to**

Sensitivity or Recall, Precision, F1-Score (F-Measure), and Balanced accuracy, all of which can be derived from the Confusion Matrix, which is generated while testing the model. The following are the different combinations of recall and precision for a given class:

- High Precision High Recall: The model handled the classification task properly
- High Precision Low Recall: The model cannot classify the data points of the particular class properly, but is highly reliable when it does so
- Low Precision High Recall: The model classifies the data points of the particular class well, but misclassifies high number of data points from other classes as the class in consideration
- Low Precision Low Recall: The model handled the classification task poorly

2.3.3 Approaches to dealing with imbalanced data

Imbalanced datasets can be found in a variety of domains where machine learning is used, such as business, finance, and biomedical science. In imbalanced datasets, one (or more) class(es) has an extremely low number of occurrences compared to the others. When traditional machine learning models are trained on such datasets, biased models with increased FP and TN rates result.

The use of an oversampling procedure to generate synthetic instances of the minority class is a popular solution to this problem. SMOTE [8] is a well-known oversampling method. It chooses an arbitrary minority class data point and its k closest minority class neighbors.SMOTE then creates synthetic minority class data points along line segments that connect the k closest neighbors. SMOTE, on the other hand, has a number of flaws, including the fact that it ignores the distribution of minority classes and latent noise in a data set.

During the classification process, the distribution of data among the classes is crucial. Traditional classification techniques (TCT) function well with data that is balanced or has similar class sizes. The balanced data sets are favored by their internal design. TCT is incapable of detecting classes given an unbalanced data set, because the algorithm's intrinsic architecture skews the results towards the larger class TCT may treat the smaller class as noise in some instances.

Exceptional cases, such as credit card fraud, tumor detection, fraudulent telephone calls, Shuttle system failure, text classification, oil spill detection, web-spam detection, risk management, information retrieval, intrusion detection, earthquake and nuclear explosion detection, helicopter gear-box fault monitoring, and so on[40][25],It is occasionally required to detect in real-world

situations. Because TCT is unable to recognize smaller classes, it is ineffective in these cases. The Class Imbalance Problem is the name given to this problem (CIP). Researchers have taken a variety of approaches to this problem, resulting in the formation of a new field of study called as Class Imbalance Learning, which is rapidly expanding. In several articles, the acronyms IDS (Imbalanced data sets), exceptional cases, skewed data sets, and skewed distributions are used. The work in this topic has been written about by a number of researchers. In 2015[2], Ali and his colleagues conducted a survey on the issue of class disparity. He highlighted a variety of topics, including class imbalance difficulties, and conducted a review of classifiers attempting to address class imbalance issues up till 2015. A recent survey, Haixiang et al[14]. discussed the technical as well as the practical aspects of the class imbalance problem with some future directions.That study[14] delves into the data-level approaches to the problem of class imbalance through 2021. Undersampling, Oversampling, and Hybrid-data-level techniques are the three main kinds of Data Level Approaches. as we see all of the approaches in figure 2.2.

FIGURE 2.2: all the algorithms and approaches to dealing with imbalance Data [35]

2.3.4 LoRAS and critics of the widely used SMOTE method

LoRAS[4] relies on locally approximating the manifold by generating a random convex combination of noisy minority class data points. LoRAS generates Gaussian noise in small neighbourhoods around the minority class samples and creates the final synthetic data with convex combinations of multiple noisy data points (shadows samples) as opposed to SMOTE-based

strategies that consider a combination of only two minority class data points. Adding these shadow samples allows LoRAS to better estimate the local mean of the latent minority class data distribution.

2.3.5 Review of The case of the island of Crete, Greece

(Varlamis, Iraklis and Apostolakis, Ioannis and Sifaki-Pistolla, Dimitra and Dey, Nilanjan and Georgoulias, Vassilios and Lionis, Christos in 2017) [36] done an intensive study on Cancer data from the local Cancer registry data, in which they applied various Data mining and data analysis:

Data preparation

- The original data had several data entry faults, incomplete or out of range values and several categorical data which overlapped.
- The patient identification step aimed in giving each patient a unique ID based on his personal details . The next step was that they removed all the redundant entries of patients' visits.
- First, from the 22 thousand incidents, they filtered out those that do not relate to Lassithi or Rethimno, which resulted in approximately 6100 incidents for patients from the two prefectures.
- Then, they removed duplicate incident entries, taking care to distinguish between duplicate recording of the same visit and consecutive visits of the patient.
- The next step was to condense all the information concerning a patient to a single row in our dataset. Since they're not interested in the temporal aspect of a patient status, and the only thing that potentially changes between consecutive visits was the diagnosis, they decided to merge the consecutive visits of a patient into a single record. This resulted in 1670 patient records for Rethimno and 2093 for Lassithi.
- they fixed these entries manually so that each code corresponds to a different entity. they fixed some out of range and missing values in the date of birth field, using patient age information and the date of the incidence, where available. Finally, they replaced the diagnosis code to "Unknown" for a couple of contradicting entries .

Feature selection

- Some features were be removed, either because they had many missing values, or because they were not related to the prediction class and other features because they did not have significant variation among patients.
- From the 41 features that they recorded for the patients, they removed features that were irrelevant to the survivability rule and features that had many missing values , thus leaving them work with 16 features.

- Gain was calculated, A high information gain value for a feature means that the feature can better assist the predictions.
- However, from the results, it was obvious that the following features were the most informative in the study, the features "Information Source", "City of Residence", and "City of Birth" demonstrated a higher information gain.
- the place of residence and birth are two features that can contribute to the survival prediction model. Another aspect of this result is that in their study, there is a difference in the survival rate of incidents between two counties, or even between the different cities among counties, **which was an interesting finding that we must work on in this thesis (on Annaba data) as well.**

Classification and survival prediction

- Although the accuracy of the survival prediction models was around 80%, one must be careful when using the model to predict the outcome of a future incident.
- When the number of cases is limited, either because the focus is on a limited time period or in specific regions or age groups, then it would **makes sense to use all types of cancer in the mortality or morbidity analysis** as it is done in or in the earlier work of Yancik & Ries's[39] study on cancer and older people and so will we do in this thesis.
- classification algorithms and methods used: C4.5 80.31% with sensitivity 0.936% which is a great result comparing to Naïve Bayes which did perform well relatively to other methods, Random Forests acc: 80.69% and sensitivity 0.915%, Logistic Regression, a standard SVM were used with similar results.

2.3.6 Review of more recent works done mortality prediction (approaches and methods)

(Veith Steele, 2018)[37] have Applied multiple ML algorithms to a large dataset of over 58,000 ICU admissions from the MIMIC III database that was used in the development, training and evaluation of the patient mortality predictive models. The evaluation results were presented, demonstrating favorable performance when compared to previous research work done (in their time), they have found that **LazyKStar** was the best model and was ranked number 1 according to the ROC[2] curve value compared to other models. altho the performace was great but I don't personally think we can attribute to the Weka model (LazyKstar) because it turned out to be a more or less a version of a simple KNN. the data was of good quality and quantity.

[2]An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

(Cordell Brettle Robert Steele, 2021)[7] researchers at Computer Science Lab in Capitol Technology University (United Kingdom) have studied the mortality of Elective hospital admitted patients, Elective hospital admissions are those that are planned ahead of time by a physician in consultation with the patient to address a non-urgent health problem. As a result, there would be no anticipation of the patient dying. However, data shows that there is a chance of death, which can be as high as 1% of elective admissions,

The capacity to predict mortality ahead of time, say at the time of admission, could be a life-saving technique. Using machine learning approaches, they were able to predict elective admission mortality from a **large dataset of around 600,000 Maryland hospital patient admissions from 2018**.

The best performing model achieved an AUC[3] of 0.95 and accuracy of 82.8% exhibiting great predictive performance. This demonstrates the significant practical usefulness of the techniques used **(Naive Bayes, Bagging Naive Bayes, Adaboost Naive Bayes, Bayes Net, Bagging Bayes Net, KNN and Random Tree)**, given that 1% of elective admissions equates to tens of thousands of elective patient in-hospital mortality in the United States each year.

KNN showed the worse results and most of the Naive Bayes performed well.

(Christian Steinmeyer, Dr. Lena Wiese†) [31] Research Group Bioinformatics in Fraunhofer Institute for Toxicology and Experimental Medicine Germany, also worked on mortality prediction, Predicting mortality is an important aspect of the decision-making process in the medical field. with the amount of data they had they thought using **(Artificial) Neural Networks would give them the most promising good results, they used NNs to test the reproducibility of a published mortality prediction method.**

(Sumit Tripathi, Sunidhi Batra Shivam Pandey) [34] Department of CSE Jaypee University of Engineering Technology Guna in India, worked on mortality prediction as well but more specifically dealing with unbalanced data using machine learning, they have been provided with a dataset that contains 342 features. The training data consists of 80,000 entry (patient's records) which are highly biased as 90 percent of the labels belong to the class '0'. Various oversampling techniques were tried due to the skewed nature of the data. Using machine learning techniques, researchers were able to predict a person's risks of dying during their hospital stay based on their medical records. On the testing data, an accuracy of 76.68% was reached (20,000 records).

Summary of the reviewed state of the art approaches

[3]The AUC provides an aggregate measure of performance for all possible classification thresholds. The AUC can be interpreted as a measure of the probability that the model will classify a random positive example over a random negative example.

Year	Study	Best performing algorithms	Field	Results
2018	(Veith & Steele)	LazyKStar (Weka)	mortality predictive model	90 % accuracy
2021	[Cordell Brettle & Robert Steele)	Naive Bayes, Bagging Naive Bayes, Adaboost Naive Bayes	predict mortality	82.8% accuracy

TABLE 2.1: Comparison between State of the art mortality prediction approaches

According to the discussed and reviewed Data mining, Imbalanced Data and mortality prediction approaches, the comparison of the proposed articles is demonstrated in Table 3 which shows the methods used and research done

2.4 Data Analysis

The process of analyzing, cleansing, manipulating, and modeling data with the objective of identifying usable information, informing conclusions, and assisting decision-making is known as data analysis. Data analysis has several dimensions and approaches, including a wide range of methodologies under several titles and being applied in a variety of business, science, and social science sectors. In today's world of business, Data analysis aids in making more scientific decisions and assisting firms in operating more efficiently.

Because of the growing availability of cancer registry data, statistical methods for descriptive analysis of population-based cancer data have advanced significantly. When practical and suitable, these methodological innovations should be used routinely to descriptive epidemiology and contemporary cancer data reporting.

2.5 Forecasting

Forecasting is the practice of creating forecasts based on historical and current data, most typically through trend analysis. An everyday example would be the estimation of a variable of interest at a future date. Prediction is a related but broader phrase. Both terms could be used to describe formal statistical procedures that use time series data.

data from different time periods or longitudinal data, or less formal judgement procedures. The terms "forecast" and "forecasting" are sometimes reserved in hydrology for predictions of values at specific future periods, whereas the term "prediction" is used for more general estimations, such as the number of times floods will occur during a given time period.

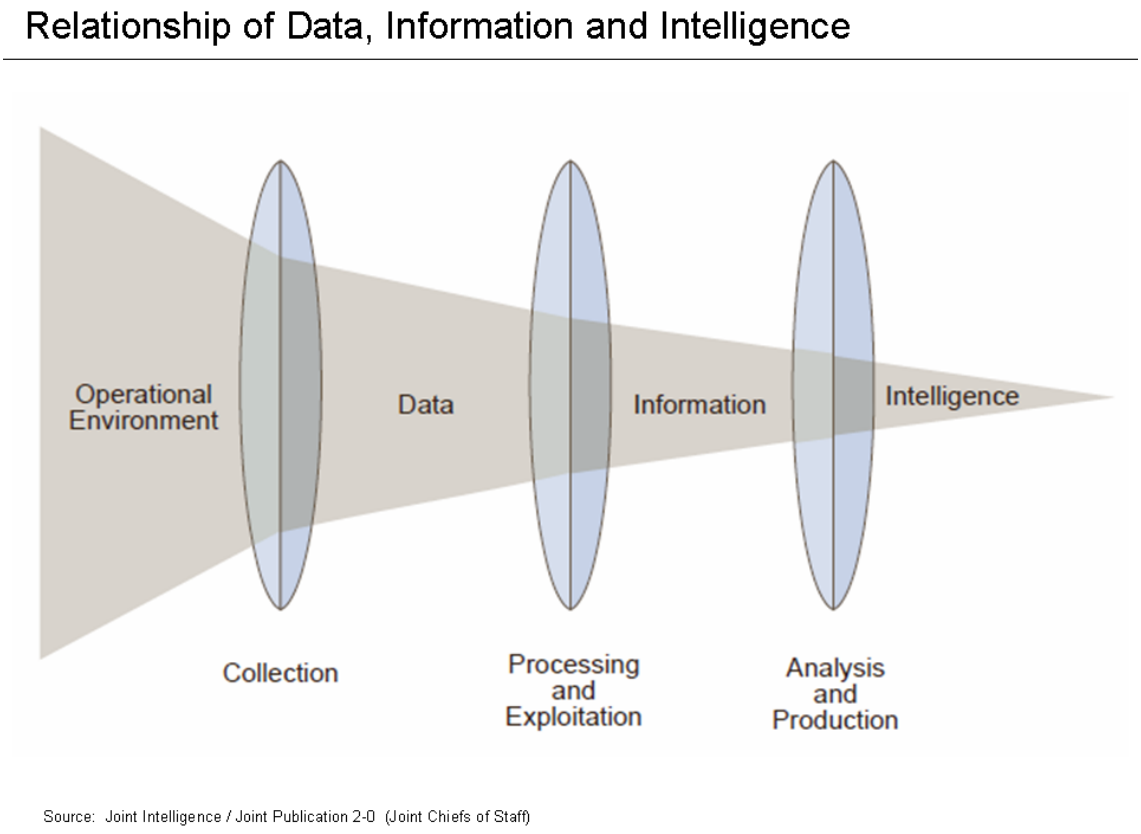

FIGURE 2.3: Statistical Data Analysis

Chapter 3

Conception

3.1 Introduction

We will discuss the general process of our work as well as the implementation of the system proposed in this thesis in this section. We will go over the entire data mining technique that was used for the Annaba Region Cancer Registry (and east Algeria). Our first goal was to develop a better prediction model for better data analysis and cancer reporting in the region.

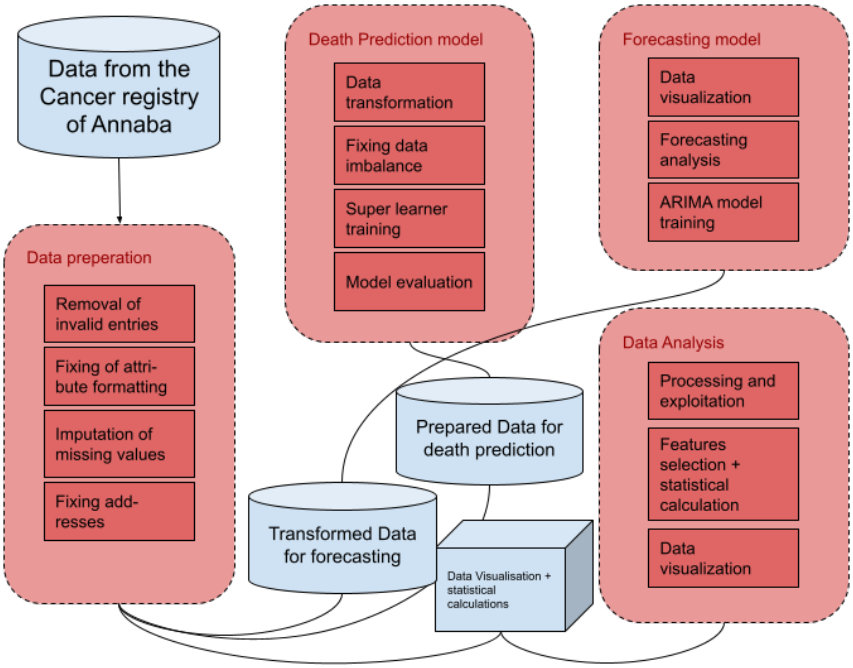

FIGURE 3.1: System Architecture

as we have mentioned before in this thesis, main problematic was to better understand the cancer data and analyse from the basis of this problematic four points :

3.1.1 Statistical Data Analysis

the core access of this work statistical analysis on cancer registry, this one serves the other points of research our contribution in this work, can be summarized into 3 main points: improvement of the mortality prediction model and the statistical analysis of the data from the cancer registry.

(stat and analysis source and aim)

3.1.2 Mortality Prediction Model

the objective is to design, validate, and evaluate a machine learning algorithms that use structured electronic health record data to predict mortality among cancer patients before a clinic visit.

3.1.3 Forecasting

In the third point our contribution is considered generating over time of statistic from the cancer registry: Extrapolation is a term used in the classical statistical treatment of time series data to make predictions about the future.

Time series forecasting is a term used in more recent domains to describe the topic.

Forecasting includes adapting models to historical data and then utilizing those models to forecast future events.

ARIMA models offer a different perspective on time series forecasting. The two most generally used approaches to time series forecasting are exponential smoothing and ARIMA models, which provide complimentary approaches to the problem. While exponential smoothing models try to explain the data's trend and seasonality, ARIMA models strive to characterize the data's autocorrelations.

3.2 Conception

3.2.1 System Architecture

The research and contributions we did in this thesis focused on 4 main axes, firstly we needed to work on the data mining part and improve the results of the mortality prediction model and more importantly justify the results found, secondly we needed the statistical analysis of the data whom which we'll need in the other three points of the contribution since having a statistical visualization of the data will paint a clearer vision as to what needs to be addressed (for example: the information gain calculation will aid you in the prediction model conception).

third point was to work on predicting the prevalence and distribution of the cancer data in future periods of times using state of the art techniques such as ARIMA, FB prophet, etc ...

forth point, was to study the factors of prevalence for the disease in the region,

3.2.2 Data Source

The Annaba Cancer Registry, in partnership with the World Health Organization, gathered data on cancer fatalities and newly diagnosed cases (WHO). the information provided for our research encompasses the years 2013 to 2017. It is made up of 5875 instances, each with 162 unique characteristics. Because the goal is to create a model that can predict a patient's survival, The latter is utilized as the class output because the goal is to create a model to forecast a patient's survivability.

3.2.3 Data preparation

3.2.3.1 Feature selection

When creating a predictive model, the technique of feature selection is used to reduce the number of input variables.

Reduce the amount of input variables to save modeling costs and, in some situations, increase model performance.

The relationship between each input variable and the goal variable is evaluated using statistics, and the input variables with the strongest link with the target variable are selected. Although the choice of statistical measures depends on the data type of both the input and output variables, these methods can be quick and successful.

In order to predict the target variable, feature selection approaches are used to reduce the number of input variables to those that are thought to be most relevant to a model.

"Feature selection is primarily focused on removing non-informative or redundant predictors from the model." [19]

A large number of variables can slow the development and training of models and require a lot of system memory, which is one of the problems with predictive modeling. Additionally, some models' performance can suffer when input variables that are unrelated to the target variable are included.

Many models will estimate parameters for each term in the model, especially those based on regression slopes and intercepts. As a result, the presence of non-informative variables can increase uncertainty in predictions and reduce the model's overall effectiveness. [19]

To do our feature selection, we are going to use **Information gain**. By evaluating the information gain for each variable and selecting the variable that maximizes the information gain, which in turn minimizes the entropy and best splits the dataset into groups for effective classification, and to achieve better modeling results.

By splitting a dataset according to a given value of a random variable, Information Gain, or IG for short, measures the reduction in entropy or surprise. A higher information gain imply a lower entropy group or groups of samples, and thus a lower level of surprise.

"information gain, is simply the expected reduction in entropy caused by partitioning the examples according to this attribute." [24]

3.2.3.2 Missing values imputation

According to statistical guideline articles, bias is likely in studies with more than 10% missing data, and if more than 40% of data is missing in significant variables, the results should only be used to generate hypotheses [11][16].

and so before doing anything with the data we made sure the features we selected did not contain more than 40% missing values.

Missing values are common in real-world data. Missing values can occur for a variety of reasons, including observations that were not recorded or data corruption. Many machine learning algorithms do not support data with missing values, so handling missing data is critical. In fact, as the size of the data set grows, the likelihood of at least one missing data point grows.[18]

Imputing refers to using a model to replace missing values. Data that is missing can be imputed. we can estimate the values of other predictors by using information from the training set predictors.[20]

When there is missing data, not all algorithms fail. There are algorithms that can be made to be robust to missing data, such as the k-Nearest Neighbors algorithm, which can ignore a column from a distance measure when a value is missing. When making a prediction, Naive Bayes can also take into account missing values. **One of the best features of Naive Bayes is that missing values are not an issue.** [38] but a few predictive models, particularly tree-based approaches, can account for missing data specifically [21].

3.2.3.3 Dealing with imbalanced Data

As we talked about it in Chapter 2 imbalanced classes in classification models, of which each class has a disproportionate number of observations, are a common problem in machine learning classification (such as our case). Medical diagnosis, spam filtering, and fraud detection are all examples of areas where there is a class imbalance.

Extension	Description
Borderline1/2 SMOTE (Han et al. 2005)	Identifies borderline samples and applies SMOTE on them
ADASYN (Haibo et al. 2008)	Adaptively changes the weights of diferent minority samples
SVM-SMOTE (Suh et al. 2017)	Generates new minority samples near borderlines with SVM
Safe-level-SMOTE (Bunkhumpornpat et al. 2009)	Generates data in areas that are completely safe
Safe-level-SMOTE (Bunkhumpornpat et al. 2009)	Generates data in areas that are completely safe
MWMOTE (Barua et al. 2014)	Identifes and weighs ambiguous minority class samples

TABLE 3.1: Popular algorithms built on SMOTE

Borderline1/2 SMOTE (Han et al. 2005) Identifes borderline samples and applies SMOTE on them ADASYN (Haibo et al. 2008) Adaptively changes the weights of diferent minority samples SVM-SMOTE (Suh et al. 2017) Generates new minority samples near borderlines with SVM Safe-level-SMOTE (Bunkhumpornpat et al. 2009) Generates data in areas that are completely safe MWMOTE (Barua et al. 2014) Identifes and weighs ambiguous minority class samples

SMOTE [8] (Synthetic Minority Oversampling Technique) is a frequently used technique for data processing. datasets that are unbalanced SMOTE is well-known for over-generalizing the minority class, resulting in misclassifications for the majority. (Narek Davtyan and al.) [4] as mentioned in Chapter 2 presented Localized Randomized Affine Shadowsampling (LoRAS), which produces better ML models for imbalanced datasets, compared to state-of-the art oversampling techniques such as SMOTE and several of its extensions. Drawing samples from a locally approximated data manifold of the minority class can produce balanced classification ML models, as demonstrated by computational investigations and a mathematical proof. They tested the method on 12 publicly available imbalanced datasets, comparing the results of three state-of-the-art convex-combination based oversampling strategies to **LoRAS** performance.

SMOTE does not take into account neighboring examples from other classes when generating synthetic examples. This could result in more class overlap and noise. For high-dimensional data, SMOTE isn't very useful.

3.2.3.4 Transforming data for the forecasting model

A time series dataset is different. Time series adds an explicit order dependence between observations: a time dimension. This additional dimension is both a constraint and a structure that provides a source of additional information. A time series is a sequence of observations taken sequentially in time.[6]

And so we took the data we had, and turned into a Time series dataset where our index was the date(first by day and then by month) and our variable was the number of diagnostics and the number of deaths on that date, as well as the ASMR (age standardized morbidity for that month). to feed it into our time series forecasting model.

3.2.4 Mortality Prediction Model

the work we did focused on creating a mortality prediction model with superior accuracy and a more validated result from an AI theoretical perspective.

For a predictive modeling challenge, there are hundreds of models to select from, making it difficult to identify which is appropriate for specific jobs.

Then, once you've decided on a model, you need to decide which model parameters again work best for your problem.

For this very reason we decided to use The super learner.[22]

3.2.4.1 The super learner

The super learner is an ensemble machine learning technique that takes all of the models and model configurations you might look into for a predictive modeling problem and uses them to generate a prediction that is as good as or better than any single model you looked into.

This problem is addressed by **superlearning**, which combines numerous typical machine learning algorithms into an ensemble that discovers the best combination of different learning algorithms.

It's a strategy in which we aggregate all of the models and configurations you could look into for a predictive modeling challenge and use them to generate a prediction that's as good as or better than any single model you could have looked into [23].

The concept behind the SuperLearner, is that it's an algorithm of estimating the performance of multiple machine learning models, or the same model with different settings, using cross-validation. Using the test data performance, it then creates an optimal weighted average of those models, dubbed an "ensemble." This method has been shown to be as accurate as **the best possible prediction algorithm**.

Individual models would be trained on k-fold data split, and then a final meta-model would be trained on their output, also known as an out-of-fold prediction from each model, in a super-learner that would be a form of stacking or k-fold cross-validation, as we seen in figure 3.2

3.2.4.2 Classification algorithms used in the meta model

A classifier is a machine learning model that is used to discriminate different objects based on certain features. With trials and errors, some experimentation and based on the state of the art work done we've decided to go with these algorithms:

3.2.4.3 Logistic Regression

Logistic regression is another statistical tool that machine learning has taken. It's the method of choice usually for binary classification issues (problems

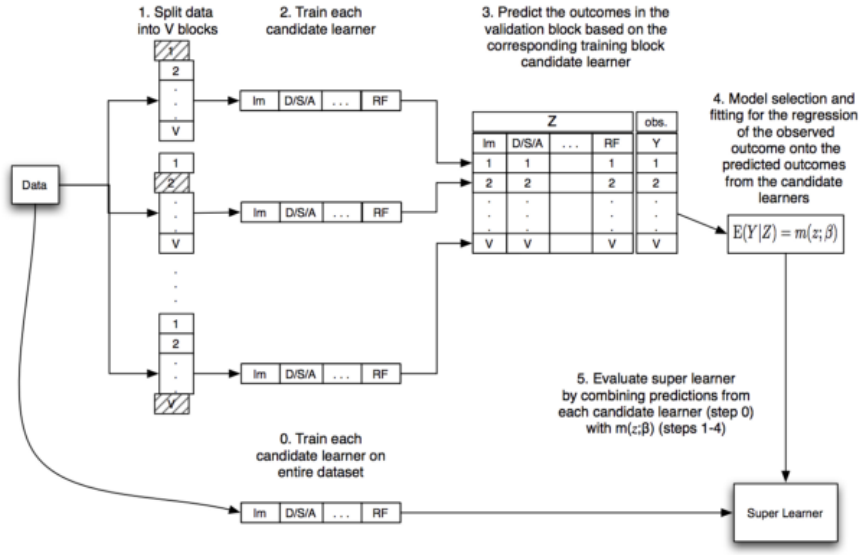

FIGURE 3.2: Super learner Architecture

with two class values) just like our death prediction problematic (death or survival of the patient).

Logistic Function

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Logistic Regression Predicts Probabilities

Logistic regression models the probability of the default class (e.g. the first class).

For example, if we are modeling people's sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height, or more formally:

In the work done by (Sandip S.Panesar and Al) [27] in which they made a comparative analysis between Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database of Raw input consisted of 21 variables (then with 12 variable) and achieved performance of accuracy/area (C.I.) under the curve of 70.7%/0.70 (49.9–88.5) for ANN, 68%/0.72 (53.4–90.4) for SVM, 66.7%/0.64 (43.6–85.0) for LR, and 65%/0.70 (51.6–89.5) for DT. Feature selected input consisted of 14 variables and achieved performance of 73.4%/0.75 (62.9–87.9) for ANN, 73.3%/0.74 (62.1–87.4) for SVM, 69.3%/0.73 (60.0–85.8) for **LR**, and 65.2%/0.63 (49.1–76.9) for DT.

The result from their work in my opinion is a strong argument for using **Logistic Regression**.

3.2.4.4 Decision Tree Classifier

For classification and regression, decision trees (DTs) are a non-parametric supervised learning method. The goal is to learn simple decision rules using data attributes to build a model that predicts the value of a target variable. A piecewise constant approximation can be visualized as a tree.

3.2.4.5 SVC

a Linear SVC's goal is to fit the data you provide and provide a "best fit" hyperplane that divides or categorizes your data. Following that, you may input some features to your classifier to check what the "predicted" class is after you've obtained the hyperplane. This makes this algorithm particularly ideal for our purpose, however it can be used in a variety of circumstances.

3.2.4.6 Naive Bayes

All Naïve Bayes algorithms are based on this principle: A probabilistic machine learning model called a Naive Bayes classifier is utilized to perform classification tasks. The Bayes theorem[15] lies at the heart of the classifier and all of it's variations:

$$P(y|\mathbf{x}) = P(y)\frac{P(\mathbf{x}|y)}{P(\mathbf{x})} \tag{3.1}$$

Types of Naive Bayes Classifier:

- **Multinomial Naive Bayes** This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.
- **Bernoulli Naive Bayes** This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.
- **Gaussian Naive Bayes** When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

Sentiment analysis, spam filtering, recommendation systems, and other applications use naive Bayes algorithms. They are quick and simple to implement, but the necessity that predictors be independent is their major drawback. The predictors are usually dependant in real-life situations, which limits the classifier's effectiveness.

3.2.4.7 AdaBoost Classifier

AdaBoost[17] was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting.

Boosting is an ensemble technique that aims to create a strong classifier from a collection of weak ones. Weak models are added one by one, and they are trained with weighted training data. The process is repeated until a predetermined number of weak learners has been created (a user parameter) or the training dataset can no longer be improved. After you've finished, you'll have a group of weak students, each with a stage value.

3.2.4.8 Random Forest Classifier

Random forest is a machine learning algorithm that uses a group of algorithms to learn anything. Given its outstanding or great performance across a broad range of classification and regression predictive modeling problems, it is the most popular and commonly used machine learning method. It's also simple to use, thanks to a small number of critical hyperparameters and logical heuristics for tuning them. The model is built on the basis of randomness. This means that the method will yield a slightly different model each time it is run on the same data. It is best practice to evaluate ml algorithms with a stochastic learning algorithm by averaging their performance across numerous runs or cross-validation repeats. When fitting a final model, it may be preferable to either increase the number of trees or fit numerous final models and average their predictions until the model's variance is minimized throughout successive evaluations.

3.2.4.9 Extra Trees Classifier

3.2.4.10 GaussianProcessClassifier

3.2.4.11 DecisionTreeClassifier

3.2.4.12 SVC

3.2.4.13 KNeighborsClassifier

3.2.4.14 AdaBoostClassifier

3.2.4.15 BaggingClassifier

3.2.4.16 RandomForestClassifier

3.2.4.17 ExtraTreesClassifier

3.2.5 Data Analysis

3.2.6 Forecasting

Time series forecasting is a crucial aspect of machine learning that is sometimes overlooked. Because there are so many prediction issues with a temporal component, it's critical. These issues are overlooked since it is the time component of time series problems that makes them more difficult to solve.

When making predictions for new data, keep in mind that the actual outcome may not be known until a later date. Although the future is predicted, all previous observations are almost always given equal weight. Perhaps with some minor temporal dynamics to avoid the concept of "concept drift," such as using only the most recent year of observations rather than all available data.

as discussed earlier in this chapter **??**A time series dataset is different. Time series adds an explicit order dependence between observations: a time dimension. This additional dimension is both a constraint and a structure that provides a source of additional information. A time series is a sequence of observations taken sequentially in time.[6]

Chapter 4

Implementation and results

4.1 Introduction

Our contribution in this work, can be summarized into 3 main points: improvement of the mortality prediction model, the statistical analysis of the data from the cancer registry and create a mortality and morbidity time series model.

The implementation is mainly useful in the aim of achieving a validated prediction mortality prediction model and doing a statistical data analysis and cancer reporting in the region of Annaba. In addition, we made several useful applications of data manipulation such as data cleaning and visualization and processing in a Time series dataset. The implemented work have been done in the programming language Python 3.6 version. Using the anaconda environment for development mostly with IDE Spyder, as for the death prediction meta model we had to use Google Colab because it relatively took more execution time and computational power.

4.2 Materials

4.2.1 Colab

Colaboratory, often abbreviated to "Colab", is a product of Google Research. Colab allows anyone to write and run Python code of their choice through the

browser. It is an environment particularly suited for machine learning, data analysis and education. In more technical terms, Colab is a hosted Jupyter notebook service that requires no configuration and provides free access to computing resources, including GPUs.

4.2.2 Spyder

Spyder is a cross-platform, open-source integrated development environment (IDE) for scientific Python programming. Spyder works with a variety of popular Python packages, including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy, and Cython, as well as other open-source applications. It's made available under the MIT license.

4.2.3 Sklearn

Scikit-learn (commonly known as sklearn) is a Python machine learning library that is available for free. It includes support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy. Sklearn simply was the go to machine learning library for building our death prediction meta model for:

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

4.2.4 pyLoras

4.2.5 Pandas

pandas is a data manipulation and analysis software package for the Python programming language. It includes data structures and methods for manipulating numerical tables, dataframes and time series, It's open-source software with a three-clause BSD license.

4.3 Methods and results

4.3.1 Mortality prediction model

4.3.2 dealing with imbalanced dataset

As we mentioned before in chapter 2 for imbalanced datasets Sensitivity or Recall, Precision, and F1-Score (F-Measure), as well as Balanced accuracy, may all be calculated from the Confusion Matrix, which is generated while testing the model. For a given class, the different combinations of recall and precision have the following meanings:

- High Precision & High Recall: The model handled the classification task properly.
- High Precision & Low Recall: The model cannot classify the data points of the particular class properly, but is highly reliable when it does so.
- Low Precision & High Recall: The model classifies the data points of the particular class well, but misclassifies high number of data points from other classes as the class in consideration.
- Low Precision & Low Recall: The model handled the classification task poorly.
- The F1-Score is the harmonic mean of precision and recall, which balances a model in terms of precision and recall. (Abd Elrahman et al.)[1] has thoroughly defined and discussed these measures. Balanced accuracy is the mean of the individual class accuracies and in this context, it is more informative than the usual accuracy score. High Balanced accuracy ensures that the ML algorithm learns adequately for each individual class.

	precision	recall	f1-score	support
1	0.89	0.97	0.93	1722
2	0.58	0.28	0.38	278
accuracy			0.87	2000
macro avg	0.73	0.63	0.65	2000
weighted avg	0.85	0.87	0.85	2000

TABLE 4.1: classification report for the super learner ensemble (with randomly sampled test data)

	precision	recall	f1-score	support
1	0.69	0.98	0.81	200
2	0.97	0.56	0.72	200
accuracy			0.78	400
macro avg	0.83	0.77	0.76	400
weighted avg	0.83	0.78	0.76	400

TABLE 4.2: classification report for the super learner ensemble (with 200-200 test data)

Model validation refers to the process of confirming that the model actually achieves its intended purpose. In most situations, this will involve confirmation that the model is predictive under the conditions of its intended use.

4.3.3 Forecasting model

4.4 Results and discussion

4.4.1 Mortality Prediction model

4.4.2 Statistical data analysis of cancer data

4.5 Conception

4.5.1 axes of research

The research and contributions we did in this thesis focused on 4 main axes, firstly we needed to work on the data mining part and improve the results of the mortality prediction model and more importantly justify the results found, secondly we needed the statistical analysis of the data whom which we'll need in the other three points of the contribution since having a statistical visualization of the data will paint a clearer vision as to what needs to be addressed (for example: the information gain calculation will aid you in the prediction model conception).

third point was to work on predicting the prevalence and distribution of the cancer data in future periods of times using state of the art techniques such as ARIMA, FB prophet, etc ...

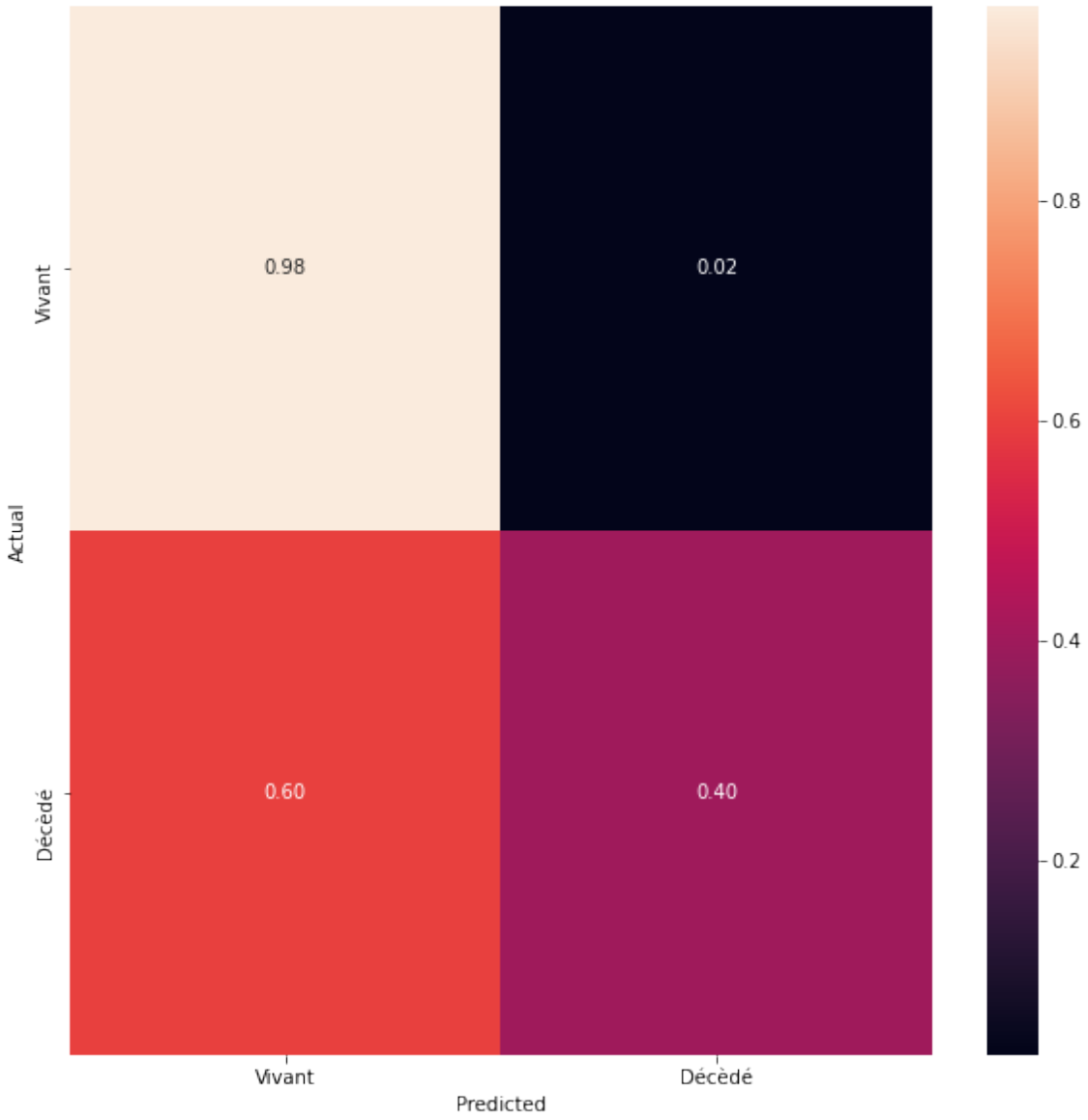

FIGURE 4.1: Confusion Matrix

forth point, was to study the factors of prevalence for the disease in the region,

4.5.2 design

From the research axes we set, we designed a system ..

4.5.2.1 Data mining

the work we did focused on creating a mortality prediction model with superior accuracy and a more validated result from an AI theoretical perspective.

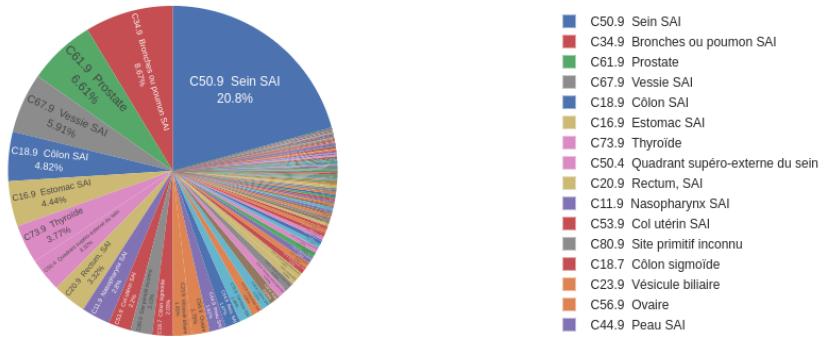

FIGURE 4.2: Cancer by type in the wilaya of Annaba

4.6 Obstacles and challenges

4.6.1 to rewrite

Several cancer variations were found among the communes, per gender and age groups. The ASIRs in the regions were higher than the international and European ones. This is a finding that truly deserved further discussion since it is unknown to what extend certain methodological considerations (the short duration of the observational period, the analysis of two out of the four counties of the island, a manually-driven collection methodology) may had an impact on the study results. However, it should be interpreted with caution. More evidence is needed to support that cancer morbidity has been increased especially in an island like Crete, where cancer epidemiology was among the lowest in Greece and Europe.

4.6.2 Cancer morbidity and mortality data

4.6.3 Statistical data analysis

4.7 Material and methods

4.8 Results and discussion

4.9 Obstacles and challenges

Residence			Age Group, N of incidents (N of deaths)								Total
			=<14	15-24	25-34	35-44	45-54	55-64	65-74	>=75	
EL-BOUNI	Gender	Male	1(1)	5(1)	1(0)	3(2)	22(6)	31(8)	30(30)	18(18)	106(30)
		Female	5(2)	1(0)	8(0)	38(3)	49(4)	34(34)	19(19)	11(11)	165(15)
	Total		6(3)	6(1)	9(0)	41(5)	71(10)	71(10)	65(11)	49(49)	271(45)
ANNABA	Gender	Male	4(2)	1(1)	8(2)	18(2)	51(11)	93(20)	86(86)	84(84)	336(72)
		Female	14(7)	3(1)	13(2)	88(5)	124(8)	94(94)	59(59)	46(46)	439(49)
	Total		18(9)	4(2)	21(4)	106(7)	175(19)	175(19)	187(32)	145(145)	775(121)
SIDI AMAR	Gender	Male	0(0)	2(0)	4(0)	2(0)	7(2)	20(3)	23(23)	12(12)	67(10)
		Female	4(3)	0(0)	4(0)	24(3)	25(1)	20(20)	10(10)	6(6)	90(12)
	Total		4(3)	2(0)	8(0)	26(3)	32(3)	32(3)	40(6)	33(33)	157(22)
EL-HADJAR	Gender	Male	1(0)	0(0)	0(0)	2(0)	2(1)	17(1)	9(9)	5(5)	36(3)
		Female	0(0)	0(0)	6(1)	13(1)	19(1)	11(11)	7(7)	6(6)	62(6)
	Total		1(0)	0(0)	6(1)	15(1)	21(2)	21(2)	28(3)	16(16)	98(9)
Com.inconnu	Gender	Male	4(0)	10(0)	10(0)	43(1)	115(9)	153(7)	172(172)	153(153)	646(33)
		Female	3(0)	10(0)	52(0)	129(5)	157(3)	187(187)	113(113)	70(70)	713(20)
	Total		7(0)	20(0)	62(0)	172(6)	272(12)	272(12)	340(13)	285(285)	1359(53)
CHEURFA	Gender	Male	0(0)	0(0)	0(0)	0(0)	1(0)	3(0)	0(0)	0(0)	4(0)
		Female	0(0)	0(0)	1(0)	3(1)	0(0)	1(1)	1(1)	0(0)	6(1)
	Total		0(0)	0(0)	1(0)	3(1)	1(0)	1(0)	4(0)	1(1)	10(1)
AIN-BERDA	Gender	Male	1(1)	0(0)	0(0)	2(0)	4(0)	6(0)	6(6)	6(6)	25(4)
		Female	1(0)	1(0)	1(0)	9(0)	4(0)	4(4)	7(7)	5(5)	28(2)
	Total		2(1)	1(0)	1(0)	11(0)	8(0)	8(0)	10(1)	13(13)	53(6)
OUED-ANEB	Gender	Male	0(0)	0(0)	0(0)	2(0)	1(0)	5(0)	3(3)	1(1)	12(0)
		Female	1(0)	0(0)	4(1)	2(0)	4(0)	5(5)	1(1)	1(1)	18(1)
	Total		1(0)	0(0)	4(1)	4(0)	5(0)	5(0)	10(0)	4(4)	30(1)
BERRAHAL	Gender	Male	2(0)	1(0)	3(0)	1(0)	4(0)	6(0)	7(7)	5(5)	29(5)
		Female	1(0)	2(0)	4(1)	5(0)	8(1)	8(8)	1(1)	3(3)	32(2)
	Total		3(0)	3(0)	7(1)	6(0)	12(1)	12(1)	14(0)	8(8)	61(7)
CHETAIBI	Gender	Male	1(1)	0(0)	1(1)	1(0)	1(0)	4(1)	3(3)	6(6)	17(6)
		Female	0(0)	0(0)	0(0)	2(0)	0(0)	1(1)	0(0)	0(0)	3(0)
	Total		1(1)	0(0)	1(1)	3(0)	1(0)	1(0)	5(1)	3(3)	20(6)
SERAIDI	Gender	Male	0(0)	0(0)	0(0)	0(0)	1(0)	2(1)	2(2)	4(4)	9(2)
		Female	1(0)	0(0)	0(0)	0(0)	2(0)	1(1)	0(0)	1(1)	5(0)
	Total		1(0)	0(0)	0(0)	0(0)	3(0)	3(0)	3(1)	2(2)	14(2)
TREAT	Gender	Male	0(0)	0(0)	0(0)	0(0)	1(0)	0(0)	1(1)	1(1)	2(0)
		Female	0(0)	0(0)	0(0)	0(0)	1(0)	0(0)	1(1)	2(2)	3(0)
	Total		0(0)	0(0)	0(0)	0(0)	2(0)	2(0)	0(0)	2(2)	5(0)
EULMA	Gender	Male	0(0)	0(0)	0(0)	0(0)	1(0)	0(0)	0(0)	0(0)	1(0)
		Female	0(0)	0(0)	2(0)	3(0)	0(0)	1(1)	0(0)	1(1)	7(0)
	Total		0(0)	0(0)	2(0)	3(0)	1(0)	1(0)	1(0)	0(0)	8(0)

TABLE 4.3: Number of cases and deaths reported per region, gender and age group, from 2013 to 2017

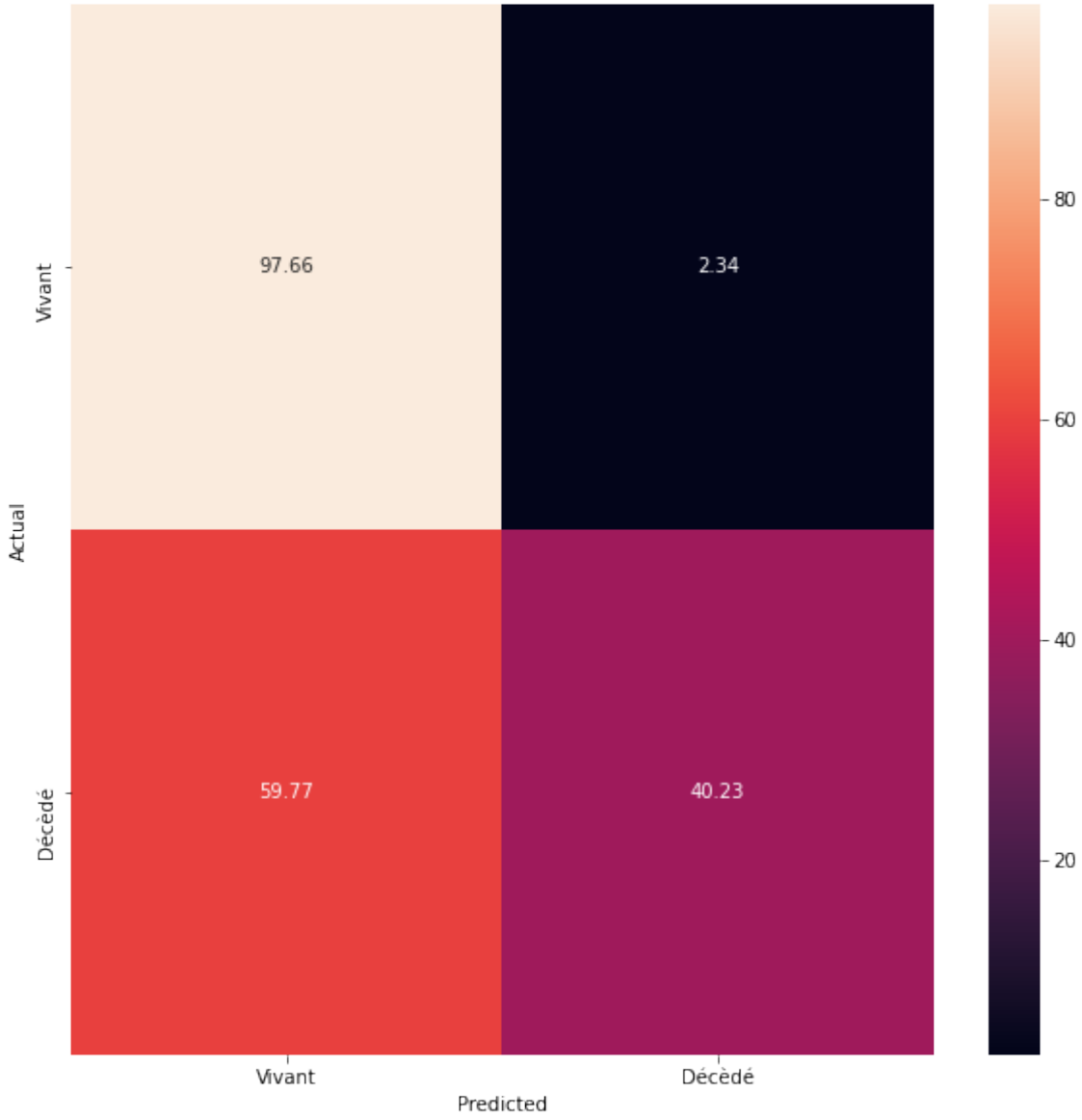

FIGURE 4.3: Confusion Matrix

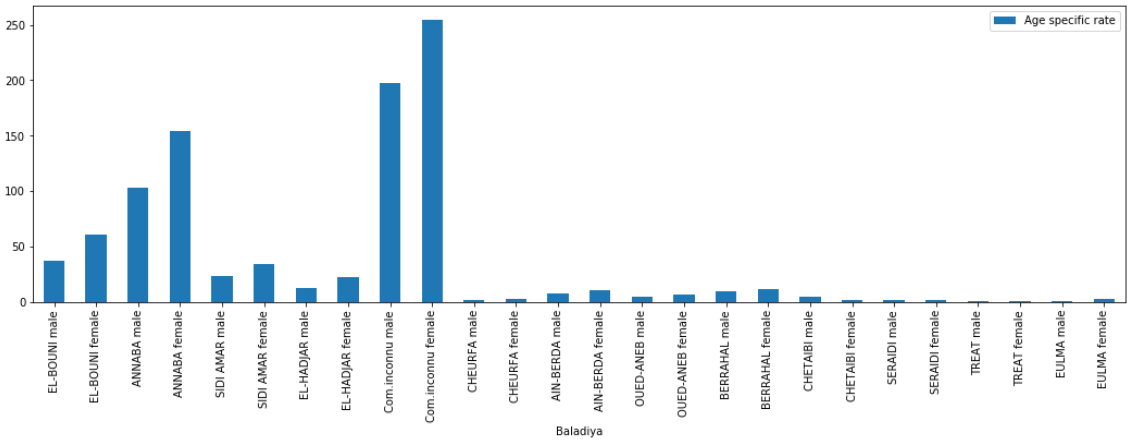

FIGURE 4.4: Age standardized Morbidity for the Wilaya of Annaba by Provinces

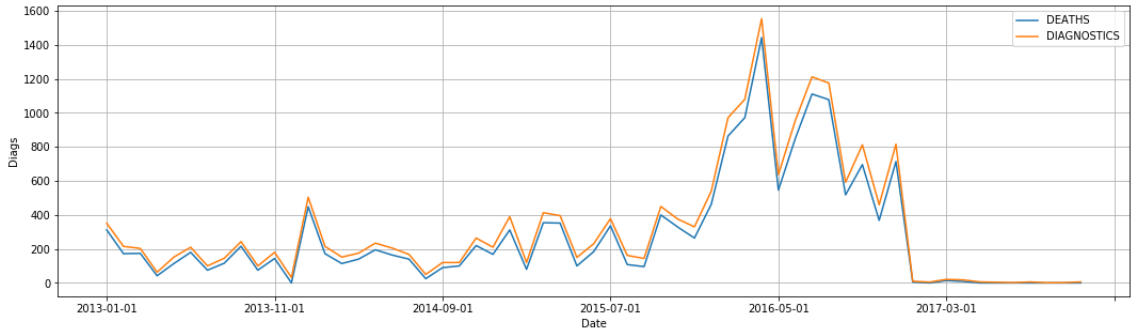

FIGURE 4.5: Age standardized Morbidity for the Wilaya of Annaba by Provinces by months

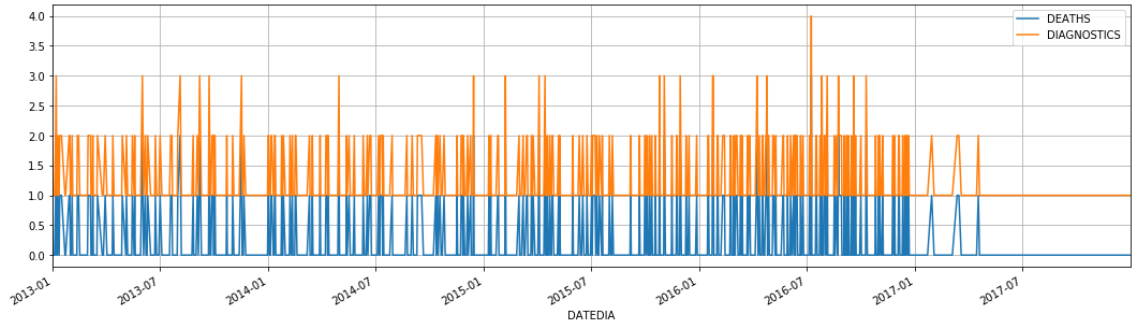

FIGURE 4.6: Diagnostics and Deaths on function of time in the Wilaya of Annaba

Bibliography

[1] Shaza M Abd Elrahman and Ajith Abraham. "A review of class imbalance problem". In: *Journal of Network and Innovative Computing* 1.2013 (2013), pp. 332–340.

[2] Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. "Classification with class imbalance problem: A Review". In: *Int. J. Advance Soft Compu. Appl* 7.3 (2015).

[3] Laura Van Metre Baum and Debra Friedman. "The Uncertain Science of Predicting Death". In: *JAMA Network Open* 3.4 (Apr. 2020), e201736–e201736. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.1736. eprint: https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2763659/baum_2020_ic_200015.pdf. URL: https://doi.org/10.1001/jamanetworkopen.2020.1736.

[4] Saptarshi Bej et al. "LoRAS: An oversampling approach for imbalanced datasets". In: *Machine Learning* 110.2 (2021), pp. 279–301.

[5] Daniel Berrar. "Cross-Validation". In: Jan. 2018. ISBN: 9780128096338. DOI: 10.1016/B978-0-12-809633-8.20349-X.

[6] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[7] Cordell Brettle and Robert Steele. "Advance Prediction of Maryland Elective Admission Fatalities Using Machine Learning". In: Mar. 2021. DOI: 10.1109/ICIM52229.2021.9417047.

[8] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[9] *Comprehensive Cancer Information*. URL: https://www.cancer.gov/.

[10] Tanvi Desai et al. "Identification of lipid-phosphatidylserine (PS) as the target of unbiasedly selected cancer specific peptide-peptoid hybrid PPS1". In: *Oncotarget* 7 (Apr. 2016). DOI: 10.18632/oncotarget.8929.

[11] Yiran Dong and Chao-Ying Joanne Peng. "Principled missing data methods for researchers". In: *SpringerPlus* 2.1 (2013), pp. 1–17.

[12] Programme des Nations Unies pour le Développement. *Human Development Report 2019*. 2019th ed. United Nations, 2019. URL: https://www.un-ilibrary.org/content/books/9789210044967.

[13] Omer Gersten and John R. Wilmoth. "The Cancer Transition in Japan since 1951". In: *Demographic Research* 7.5 (2002), pp. 271–306. DOI: 10.4054/DemRes.2002.7.5. eprint: https://www.demographic-research.org/volumes/vol7/5/7-5.pdf. URL: https://www.demographic-research.org/volumes/vol7/5/.

[14] Guo Haixiang et al. "Learning from class-imbalanced data: Review of methods and applications". In: *Expert Systems with Applications* 73 (2017), pp. 220–239.

[15] David J Hand and Keming Yu. "Idiot's Bayes—not so stupid after all?" In: *International statistical review* 69.3 (2001), pp. 385–398.

[16] Janus Christian Jakobsen et al. "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts". In: *BMC medical research methodology* 17.1 (2017), pp. 1–10.

[17] Balázs Kégl. "The return of AdaBoost. MH: multi-class Hamming trees". In: *arXiv preprint arXiv:1312.6086* (2013).

[18] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[19] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[20] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[21] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013, p. 42.

[22] Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007). DOI: doi:10.2202/1544-6115.1309. URL: https://doi.org/10.2202/1544-6115.1309.

[23] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[24] TM Mitchell. "Machine Learning McGraw-Hill International". In: (1997).

[25] R Mollineda, R Alejo, and J Sotoca. "The class imbalance problem in pattern classification and learning". In: *II Congreso Espanol de Informática (CEDI 2007). ISBN*. Citeseer. 2007, pp. 978–84.

[26] AR Omran. *The epidemiologic transition: a theory of the epidemiology of population change. The Milbank Memorial Fund Quarterly, 49, 509-538*. 1971.

[27] Sandip S. Panesar et al. "Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database". In: *World Neurosurgery: X* 2 (2019), p. 100012. ISSN: 2590-1397. DOI: https://doi.org/10.1016/j.wnsx.2019.100012. URL: https://www.sciencedirect.com/science/article/pii/S2590139719300432.

[28] Pallavi Reddy et al. "A Review on Data Mining Techniques and Challenges in Medical Field". In: *International Journal of Engineering Research and* V9 (Aug. 2020). DOI: 10.17577/IJERTV9IS080143.

[29] RW Ruddon, F Holland, et al. "What makes a cancer cell a cancer cell". In: *Hoolland-Frei Cancer Medicine* (2003).

[30] Mitchell Nicholas Sarkies et al. "Data collection methods in health services research: hospital length of stay and discharge destination". In: *Applied clinical informatics* 6.1 (2015), p. 96.

[31] Christian Steinmeyer and L. Wiese. "Sampling methods and feature selection for mortality prediction with neural networks". In: *Journal of biomedical informatics* (2020), p. 103580.

[32] Hyuna Sung et al. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249. DOI: https://doi.org/10.3322/caac.21660. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

[33] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[34] Sumit Tripathi, Sunidhi Batra, and Shivam Pandey. "Unbiased Mortality Prediction for Unbalanced Data Using Machine Learning". In: *2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE. 2019, pp. 1–5.

[35] Kamlesh Upadhyay, Prabhjot Kaur, and Svav Prasad. "State of the Art on Data level methods to address Class Imbalance Problem in Binary Classification". In: 8 (Mar. 2021), pp. 975–903.

[36] Iraklis Varlamis et al. "Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of Crete, Greece". In: *Computer Methods and Programs in Biomedicine* 145 (2017), pp. 73–83.

[37] Nick Veith and Robert Steele. "Machine learning-based prediction of ICU patient mortality at time of admission". In: *Proceedings of the 2nd International Conference on Information System and Data Mining*. 2018, pp. 34–38.

[38] Ian H Witten and Eibe Frank. "Data mining: practical machine learning tools and techniques with Java implementations". In: *Acm Sigmod Record* 31.1 (2002), pp. 76–77.

[39] Rosemary Yancik and Lynn AG Ries. "Aging and cancer in America: demographic and epidemiologic perspectives". In: *Hematology/oncology clinics of North America* 14.1 (2000), pp. 17–23.

[40] Yang Yong. "The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm". In: *Energy Procedia* 17 (2012), pp. 164–170.