UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Using clinical data for breast cancer risk prediction and follow-up

---

Author:
Sergio Hernández Antón

Supervisor:
Dr. Oliver Díaz Montesdeoca

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

January 17, 2025

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Using clinical data for breast cancer risk prediction and follow-up

by Sergio Hernández Antón

Breast cancer remains one of the leading causes of cancer-related morbidity and mortality worldwide, requiring robust methodologies for early risk prediction, recurrence forecasting, and survival analysis. This thesis defines a comprehensive pipeline for breast cancer risk prediction, emphasizing both technical precision and clinical relevance. The proposed framework integrates multiple components: data acquisition, preprocessing, feature extraction, model selection, interpretability, and explainability, in order to ensure accurate, transparent, and actionable outcomes.

Overall, this thesis aims to advance the field of breast cancer prediction by delivering a robust, interpretable, and clinically relevant pipeline, aligning with the important goal of improving patient outcomes through early and precise detection.

Additionally, in an attempt to make this thesis more reachable, we add a feature dictionary for both used datasets in Appendix A. On top of that, we also share the project in the shape of a *GitHub* repository[1], so that people can take profit of this research if at all possible. We also include a guide on its structure in Appendix B.

[1] https://github.com/SergioHernandezAnton/Final_Thesis_DataScience.git

Acknowledgements

First of all, I would like to thank my project advisor, Dr. Oliver Díaz, for continuously supporting me during this thesis. I am truly sorry for bothering you with constant emails, specially during Christmas holidays. I appreciate the time you invested in helping me with your ideas and suggestions. Because of this, I really regret not being able to bring this project to the point you envisioned.

Next, to all people working for making *Universitat de Barcelona* a better place. It is thanks to your efforts I am able to conclude my studies in which has been my second home for the last six years. I really appreciate the work of every single one of you, though almost none of you will read this.

To my friends, in particular to the group formed last year in class: *ML Quizzes*. Thank you for being there, I enjoyed going to lessons for which I had no interest whatsoever because of your company.

To all of my relatives, for whom the more I grow up, the less often I am able to pay a visit. I really appreciate our family gatherings, specially if some festivity is involved.

To my closest family, my mother and brother. I love you so much, and I am truly grateful for all you have done for me in the past, specially for these last six years. I hope we continue to look after each other as we always do.

And last but certainly not least, to all of you reading this. I really appreciate you found this thesis suited your interests, and I honestly hope you enjoy reading it.

Contents

Chapter 1

Introduction

1.1 Background on breast cancer risk prediction

Breast cancer is the second major cause of women's death after lung cancer [1], representing about 12% of all new cancer cases and 25% of all cancers in women. As shown in Figure 1.1, last year it was the most common diagnosed type of cancer among US citizens (similar statistics for different cancer types can be found here[1]). Hence, early detection and precise risk prediction are crucial for improving outcomes, enabling timely intervention and reducing unnecessary procedures.

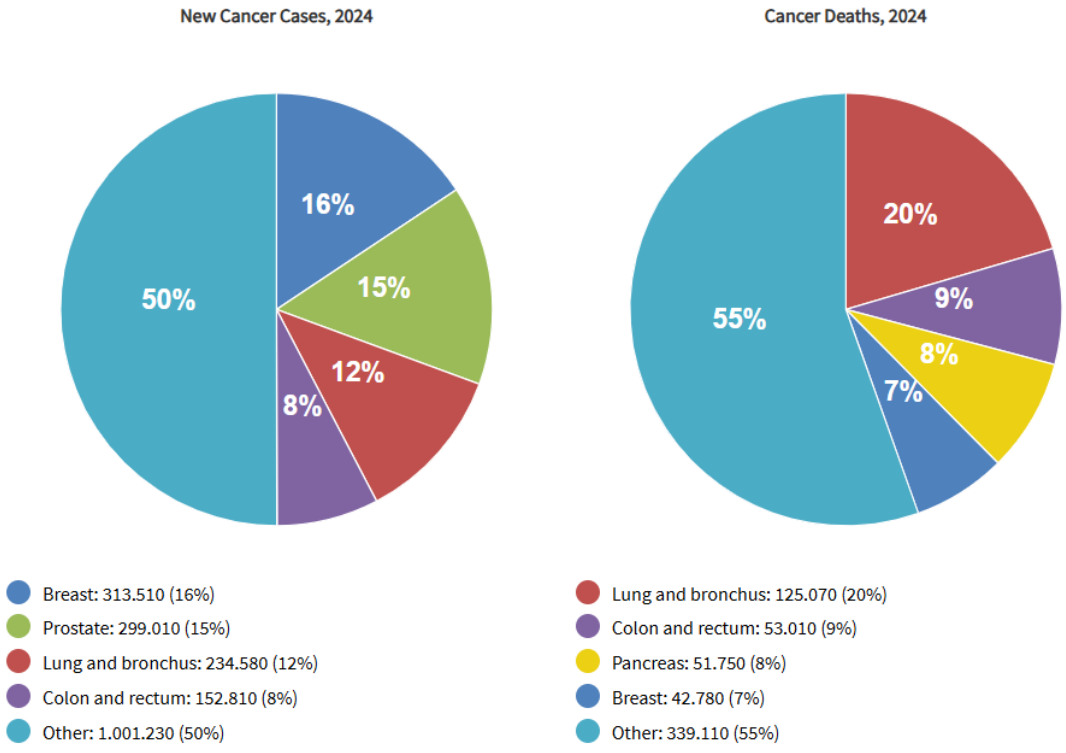

FIGURE 1.1: Statistics on new diagnosis and reported deceases for most common cancer types in 2024 for US citizens.

Risk prediction is accomplished by identifying characteristics associated with a high or low risk of developing a disease (traditional risk factors are listed in Section 1.2), and then combining those characteristics in a statistical model to produce a probability estimate of developing the disease over a certain period.

[1] https://cancercontrol.cancer.gov/ocs/statistics

Historically, demographic and clinical risk factors have been used in risk prediction models; more recently, genetic makeup has been added to certain models. Cancer risk prediction models have been used to estimate the costs of the population burden of cancer, plan intervention trials, create benefit–risk indices and design prevention strategies for at-risk populations [2].

There is an increasing interest in using risk prediction models, since they help individual patients to estimate their personal chance of being diagnosed with breast cancer. "What is my risk of getting cancer?" is a question that clinicians frequently encounter in their everyday practice. It is no wonder why the 2004 Institute of Medicine report on breast cancer screening identified individual risk assessment as essential to improve early detection of breast cancer [3].

Breast cancer risk prediction models are commonly assessed in two ways: by measuring their performance at the population level and at the level of the individual woman. In [4] it was assessed each model's performance at the population level. They compared the number of women the model estimated [E] would develop breast cancer to the ones actually diagnosed with breast cancer (observed [O]). The Italian and Gail models estimated that 186 and 180 women, respectively, would develop breast cancer. Therefore, the overall E/O ratios for the Italian and Gail models were similar (0.96, 95% confidence interval [CI] = 0.84 to 1.11; and 0.93, 95% CI = 0.81 to 1.08, respectively).

Naturally, the issue is a matter of concern for researchers worldwide, and new risk prediction Machine Learning (ML) and Artificial Intelligence (AI) approaches are being developed nowadays for most types of cancer, as we can observe in Figure 1.2. These methods include classical ML models such as Support Vector Machine (SVM) and Logistic Regression (LR). For breast cancer specifically, we can also consider traditional approaches such as the Gail and Tyrer-Cuzick models, which rely on clinical, genetic and lifestyle factors.

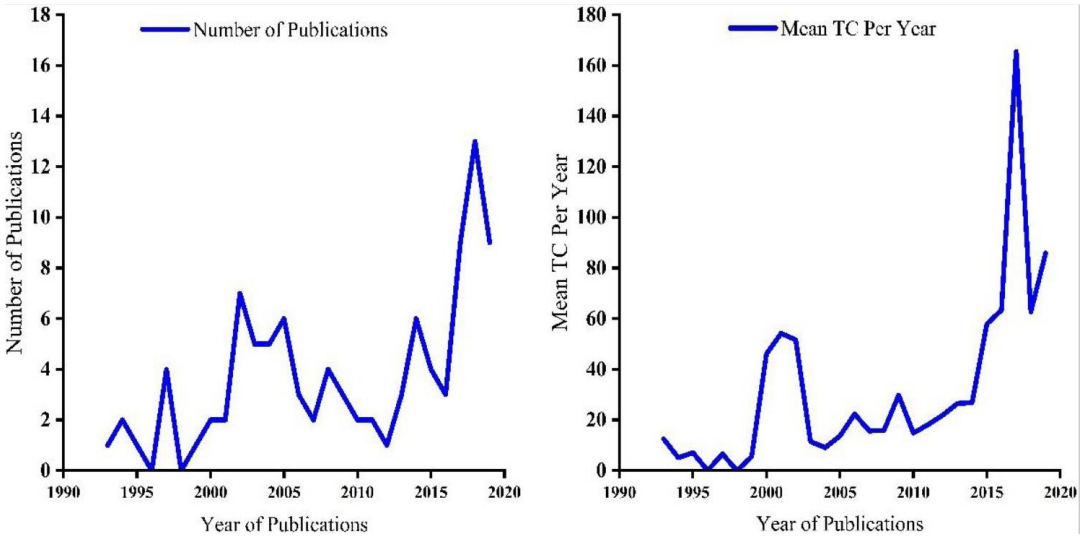

FIGURE 1.2: Annual growth of publications and mean of Total Citation Per Year (Mean TC Per Year) on LA and AI in cancer. As it was originally mentioned in [5], the results are for the top 100 cited articles indexed in Scopus database[2].

[2] https://www.elsevier.com/products/scopus

However, despite the success of the aforementioned methods, they still face significant limitations. In particular, they are trained with a single source of data, limiting their predicting power; they also suffer from inflexibility, failing to adapt to a patient's changes over time; and on top of that, they can introduce population-specific biases, reducing their applicability.

In order to mitigate these limitations, we consider the integration of human microbiome data, which has garnered substantial attention by both researchers and the media. The human microbiome refers to the collective genome of all bacteria, archaea, fungi, protists and viruses residing in and on the human body. For a more comprehensive review on microbiome studies, see [6].

Not only does human microbiome capture environmental and lifestyle influences, it also gives valuable insight into hormonal regulation and immune system modulation. Moreover, it also serves as a dynamic and modifiable biomarker, enabling personalized risk prediction while also bridging the gap for diverse populations. On top of that, by integrating microbiome data with clinical, genetic and imaging factors, risk prediction models can better capture the complex interplay of factors contributing to cancer, improving accuracy at prediction. For a more comprehensive and deeper explanation, one can check both [7] and [8].

We will come back to explain it more deeply in Section 1.4, but the main purpose of this research is to define a pipeline for breast cancer. Our intention is to show the full picture of a data-oriented project, so we go all the way from data acquisition to model explainability, including both recurrence and mortality on it. Afterwards, the original idea was to integrate microbiome data to the pipeline in order to explore and discover new risk factors through its usage. Unfortunately, we were unable to access such datasets, as we mention later in Section 5.1.

1.2 Traditional risk factors

In order to get used to the features we work with, in this section we list multiple breast cancer risk factors identified over the years, as discussed in [9].

- **Reproductive and hormonal risk factors:** Older age, older age at first live birth and at menopause, younger age at menarche, and nulliparity are associated with elevated breast cancer risk, all of which are related to prolonged exposure to endogenous estrogen. In addition, use of postmenopausal hormone therapy is a risk factor that is dependent on type and duration of use. Reproductive and hormonal factors are considered to be modest risk factors (with risk ratios ranging between 1.0 and 1.5) but, when multiple, have additive effects [10].
- **Breast density:** Dense breast tissue is an independent risk factor for breast cancer, with many studies demonstrating an odds ratio of 4.0 or greater when comparing the most dense to least dense categories [11]. Although increased breast density confers lower risk than some risk factors, it is more common among women and thus may account for a considerable proportion of population risk [12]. The addition of breast density as a risk factor improves calibration and discrimination of various risk prediction models [13].
- **Radiation exposure:** Radiation exposure between the ages of 10 and 30 years (ie, in survivors of Hodgkin lymphoma) is a known risk factor [14].

- **Genetic factors:** Family history (in particular, an affected mother, sister, or male relative, early onset disease, and bilateral disease) is an established risk factor [9]. Inheritance of high-risk genetic mutations, such as BReast CAncer gene 1 (BRCA1) and BRCA2, account for some but not all of this risk [15]. Common risk variants, mostly single-nucleotide ones (formerly known as single-nucleotide polymorphisms), can explain up to 18% of the familial risk of breast cancer and, when aggregated, can be incorporated into risk prediction models as a polygenic risk score [16].
- **Benign breast disease and prior biopsy:** Proliferative disease with atypia is a known risk factor. Specifically, there exists a 6- to 10-fold increased risk of breast cancer in women with lobular carcinoma in situ and a 4- to 5-fold increased risk in women with atypical ductal hyperplasia [17]. In addition, prior breast biopsy alone is a modest risk factor for breast cancer, with relative risk associated with histologic findings (ie, proliferative disease with atypia is of higher risk than proliferative disease without atypia, which is of higher risk than nonproliferative disease) [18].
- **Lifestyle factors:** Obesity is associated with elevated breast cancer risk in postmenopausal women, though it is believed to have a protective effect in premenopausal ones [9]. In postmenopausal obese women, the aromatase enzyme in adipose tissue converts androgens to estrogen, thus increasing breast cancer risk [19]. Premenopausal obese, however, have lower levels of serum estradiol [20]. Additionally, physical activity decreases breast cancer risk in a dose-dependent manner [21]. High levels of alcohol intake are associated to elevated breast cancer risk [9].

1.3 Integration of recurrence and mortality

In our context, incorporating recurrence and mortality prediction models is essential for building a comprehensive framework for patient care and outcomes. These predictions serve as vital components in understanding the broader implications of breast cancer management, as they address not only the risk of developing breast cancer, but also its progression and ultimate prognosis, as done in [22].

By recurrence, we refer to the return of cancer after initial treatment. This can manifest locally (at the original site or in nearby lymph nodes), or distant metastasis. We refer to these as local and distant recurrence, respectively. By predicting it, clinicians are able to anticipate disease progression, personalize treatment and guide follow-ups, all of which are also crucial for survival.

On the other hand, mortality prediction focuses on estimating the risk of death, whether from cancer or other causes. It is an essential factor for assessing overall survival, informing palliative care and public health planning. Unfortunately, even though decease rates for different types of cancer have only diminished over the last years, there is still a long way to go, as it is reflected in Figure 1.3.

Moving on to what it piqued our interest for this research specifically, recurrence and death are closely tied to breast cancer risk prediction, as they share important features such as age, genetic predisposition and receptor statuses. Moreover, while risk prediction emphasizes prevention and early detection, recurrence and death models address outcomes post-diagnosis, forming a continuum from prevention to survivorship, which is precisely what we pursue with this thesis.

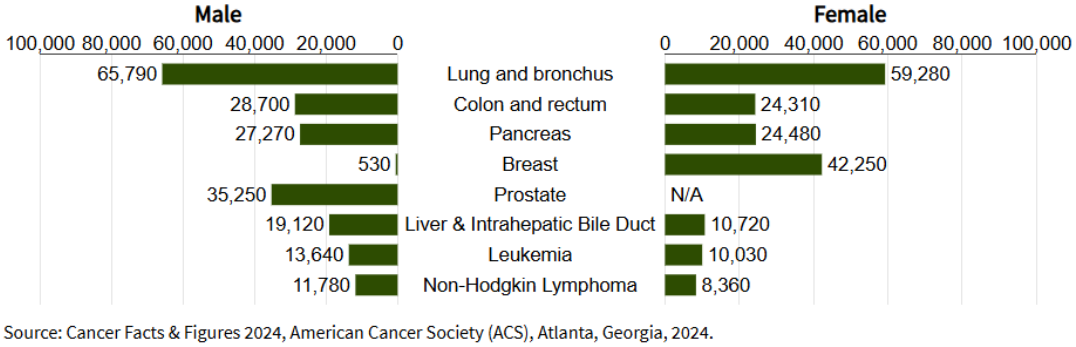

FIGURE 1.3: Count of reported deceases for the eight deadliest cancer sites in 2024 for US citizens.

Hence, the inclusion of both recurrence and mortality in a breast cancer risk prediction pipeline enriches the model's utility by addressing the full spectrum of patient outcomes. These predictions provide actionable insights, influencing clinical decision-making, optimize treatment plans, and improve patient quality of life.

1.4 Aims of the research

As it is already mentioned in Section 1.1, our main goal is to define a pipeline for breast cancer using clinical data. We made it the focus of our research because, as we will see in Section 2.2, current research does not focus on constructing robust, interpretable and fair models, all of which are crucial in breast cancer risk prediction. Thus, not only should this pipeline work for healthy and infected subjects alike, but also has to be able to explain and justify its decision-making.

We will design and test the pipeline at the same time, so we require clinical data and get results supporting our findings. The main idea is to extract the most important features for our predictions and see if these features are consistent by justifying their appearance. In the great scheme of things, the pipeline looks as Figure 1.4, though there are some details we omit to enhance simplicity. Nonetheless, we are going to properly explain the whole procedure in Chapter 3.

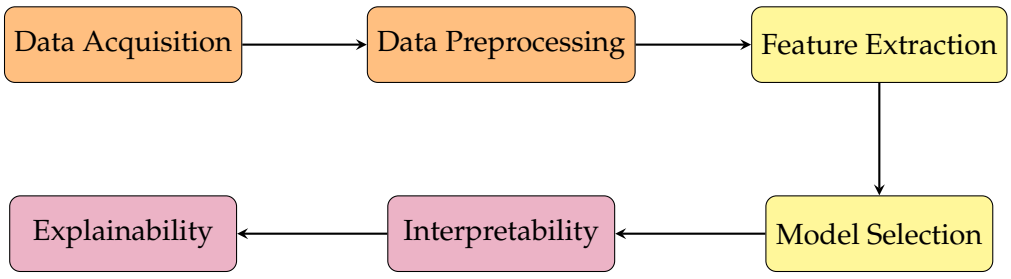

FIGURE 1.4: Pipeline for breast cancer risk prediction, which also can be used for recurrence and mortality. Blocks are partitioned by colors.

Additionally, we want the pipeline to be as robust as possible, so we intend to test it with as many data and models as we can. For this reason, in our experiments we do not stick to a particular model, but rather select the best performing one for each case. Furthermore, several metrics are defined in order to control the performance of models from different perspectives. By implementing these measures, we expect to obtain coherent and conclusive results.

Related to robustness, our next goal, and perhaps the most important one from an ethical point of view, is to select models treating all protected classes[3] equally. Figures 1.5 and 1.6 reflect our concern, since black patients have a lower rate of diagnosis but a higher rate of decease than white ones. We are fully aware there might be some other factors causing this outcome, but the point is we want our pipeline to avoid that.

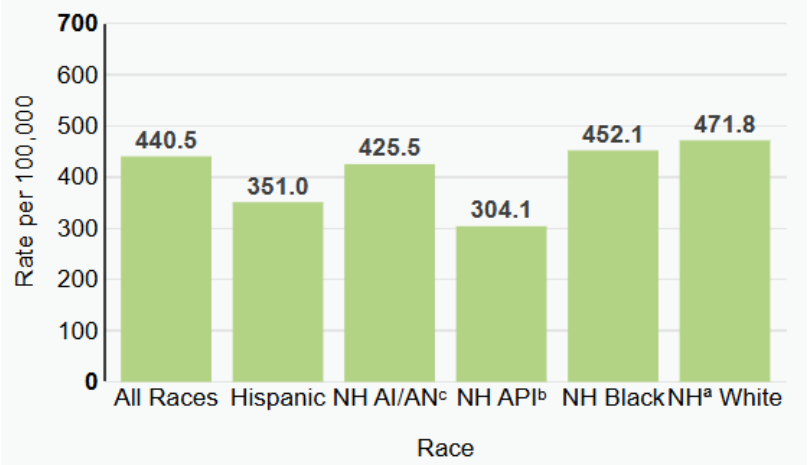

FIGURE 1.5: New cancer cases in US during 2017-2021 by race/ethnicity. [a]Non-Hispanic, [b]Asian/Pacific Islander, [c]American Indian/Alaska Native.

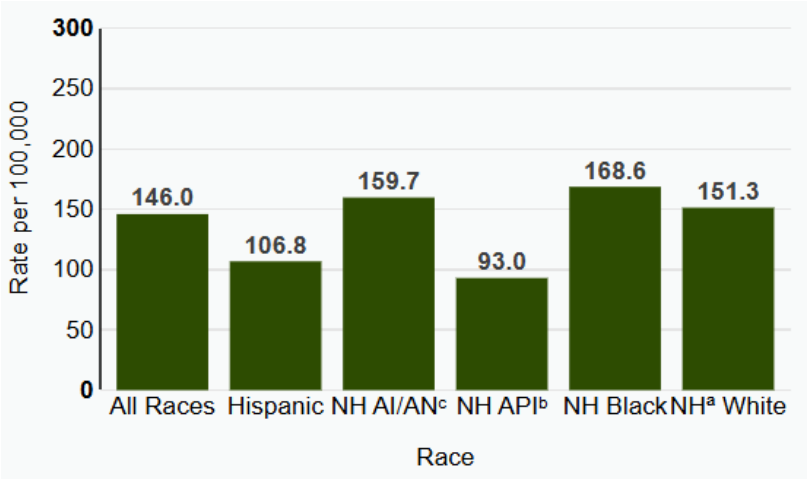

FIGURE 1.6: Reported deceases for cancer in US during 2018-2022 by race/ethnicity. [a]Non-Hispanic, [b]Asian/Pacific Islander, [c]American Indian/Alaska Native.

There are also groups which are discriminated due to data availability. As shown in Figure 1.5, women between 55-74 years old cover more than 50% of all new diagnoses. Thus, it is highly likely models will show a tendency on predicting correctly for these age ranges, while for other age groups they will not perform nearly as well. This is the reason why in Chapter 3 we perform a data exploration, using the feature dictionary in Appendix A, to manually define which variables are sensitive to become protected attributes.

[3] https://www.senate.ca.gov/protected-classes

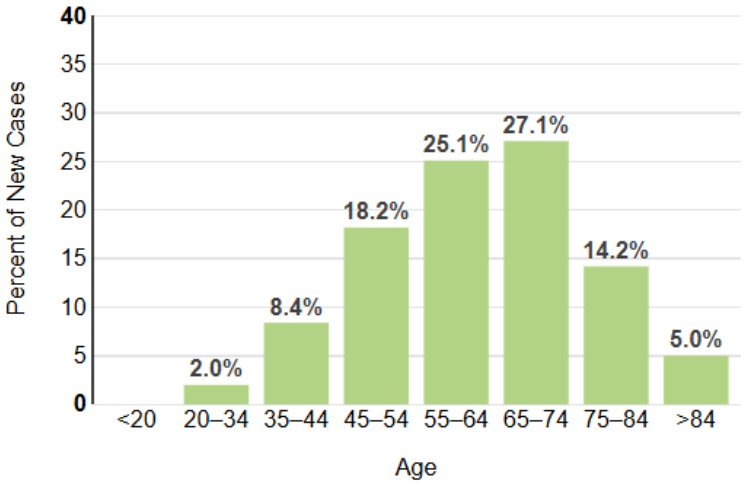

FIGURE 1.7: New confirmed breast cancer cases during 2017-2021.

Overall, we summarize the objectives of this research in the following:

- Define a pipeline starting from data acquisition and going all the way to model interpretability and explainability.
- Such pipeline should work not only for breast cancer risk prediction, but also for recurrence and mortality prediction. Thus, covering all possible outcomes.
- Test it with datasets as described above (the bigger, the better). If possible, these tests should include new or unexplored data, such as microbiome.
- Increase its robustness through the definition of several metrics, fairness assessment and the usage of different kinds of models.

Chapter 2

State of the art

ML has revolutionized healthcare by enabling data-driven solutions for disease prediction, prognosis, and management. In breast cancer, ML models have been widely adopted to enhance early detection, risk stratification, recurrence forecasting and mortality prediction. This chapter provides an exploration of state-of-the-art applications of ML in these domains, while paying special attention to Generative Adversarial Imitation Learning (GAIL) models and prognosis prediction models. Additionally, a prior review remaking models should include interpretability, explicability and fairness strategies is also examined.

2.1 ML in breast cancer

Breast cancer risk prediction models aim to stratify individuals based on their likelihood of developing the disease. Traditionally, models such as the Gail and Tyrer-Cuzick models utilized epidemiological and genetic data. However, ML has introduced more sophisticated algorithms capable of handling complex and high-dimensional datasets.

2.1.1 Traditional models

One of the most well-known models for predicting a woman's chance of being diagnosed with breast cancer is the Gail model. It includes age, race and first degree relatives with breast cancer amongst another risk factors, making it not only holistic but also robust. This model predicts the likelihood of having a breast cancer diagnose within the following 5 years and within her lifetime (up to 90 years old). Additionally, the Gail model and similar risk prediction models are readily available to clinicians and patients around the world through the Internet (see [23]). Moreover, a version of the model, available on the *National Cancer Institute* webpage[1], is accessed 20,000 to 30,000 times each month.

The aforementioned breast cancer risk prediction model, commonly known as the Gail model, was developed in 1989 and originally introduced in [24]. The model has undergone several modifications over the years, and is now available (in the link) as the Breast Cancer Risk Assessment Tool (BCRAT), see [25]. The most recent version of the BCRAT returns estimates for 5-year and lifetime invasive cancer risk for women who are at least 35 years old, as presented in [26]. In particular, they use the risk calculator to identify individuals with a 5-year risk greater than 1.67.

[1] https://bcrisktool.cancer.gov/

The Tyrer-Cuzick model, or I/O Buffer Information Specification (IBIS) model, is also among the most widely used tools for the task at hand. It was firstly presented in [27], based on data from the IBIS conducted in UK. It combines a genetic segregation model for familiar risk and a regression model for other risk factors. This genetic segregation assumes a two-locus genetic model, with one locus for BRCA1 or BRCA2 and the other locus for an unknown, low penetrance gene. This model considers as risk factors age at menarche, parity and height, among many others. The latest additions are breast density and polygenic risk scores, introduced in [28]. Additionally, a UK study demonstrated these scores based on a large number of single-nucleotide variants lead to improved risk stratification when combined with Tyrer-Cuzick risk and breast density, as shown in [29].

The computer program for the Tyrer-Cuzick model displays a chart which shows a woman's breast cancer risk until 85 years of age, in addition to 10-year and lifetime risks [27]. This program also calculates the likelihood of having BRCA mutation of a hypothetical autosomal dominant gene mutation, assumed to have a low penetrance but a high frequency in the population. Furthermore, the risk estimates can be used to identify women who would benefit from chemoprevention [30] and/or Magnetic Resonance Imaging (MRI) screening.

The model includes a multitude of genetic and nongenetic risk factors, can be used in women younger than 35, and requires use of a specific computer program. It demonstrates good calibration and discrimination when used in high-risk populations. Moreover, recent evidence suggests that models with multigenerational family history, such as Tyrer-Cuzick, estimate risk better, even for women with below-average or average breast cancer risk, which is studied in [31].

2.1.2 GAIL models

GAIL models represent a cutting-edge approach to learning complex distributions and imitating expert strategies. Thus, they have emerged as a promising framework in healthcare, including breast cancer risk prediction. By leveraging the strengths of Generative Adversarial Networks (GANs), these models can simulate complex data distributions and address the challenges of imbalanced datasets, often found in breast cancer research.

For instance, GAIL models can synthesize realistic patient profiles to augment training datasets, thereby improving the performance of risk prediction models. Such models are particularly beneficial in scenarios where data scarcity or privacy concerns restrict access to comprehensive datasets. Nevertheless, they still present some limitations, for instance their risk of overfitting to synthetic data and its high computational cost.

In the original paper [32], GAIL models are presented as an approach to address the challenge of learning policies directly from expert demonstrations without requiring explicit reward signals, a common limitation in Reinforcement Learning (RL). They draw inspiration from GANs, where a generator and discriminator compete in a zero-sum game.

The generator (here, the policy), learns to produce actions indistinguishable from the expert's behavior, while the discriminator identifies whether these actions are taken by the agent or the expert. This problem is defined as minimizing a divergence between the state-action distribution of the learned policy and that of the expert.

Mathematically, the GAIL framework optimizes the objective function:

$$\min_{\pi} \max_{D} \mathbb{E}_{\pi} \left[\log \left(D(s,a) \right) \right] + \mathbb{E}_{\pi_E} \left[\log \left(1 - D(s,a) \right) \right] - \lambda H(\pi),$$

where:

- π represents the policy being learned.
- D denotes the discriminator function.
- π_E is the expert policy.
- $H(\pi)$ signifies the entropy of the policy, promoting exploration.
- λ is a weighting factor balancing imitation and exploration.

This formulation allows the policy to learn behaviors statistically similar to the expert demonstrations, by minimizing the divergence between the agent's and expert's state-action distributions.

GAIL models benefit from great scalability, the unnecessity of defining an explicit reward function and data efficiency. As shown in [32], these models bridge the gap between imitation learning and RL, providing a robust method to learn complex behaviors from expert data without requiring hand-crafted rewards. Thus, they lay the groundwork for future research in imitation learning and generative modeling for policy learning, which applies to disease detection and, in particular, to breast cancer risk prediction.

2.1.3 Prognostic prediction models

Despite the continuous advancements in treatment, predicting the likelihood of recurrence and overall survival for breast cancer patients remains a critical challenge in clinical oncology. Traditional prognostic models often rely on well-established clinical and pathological factors such as tumor size, nodal status and hormone receptor expression. While these variables are valuable, they fail to fully capture the molecular complexity and diversity of breast cancer. This limitation has driven a growing interest in integrating genomic data with clinical features in order to develop more precise and personalized prognostic models.

The study in [33] represents a major step forward in this domain. The authors constructed models leveraging the wealth of information available from multigenomic datasets alongside traditional clinical data. Their objective was to improve the accuracy of recurrence and survival predictions, providing clinicians with tools to better tailor treatments and interventions for individual patients.

To develop their predictive models, the authors employed a robust and methodical approach. They began by using feature selection techniques to identify the most informative genomic and clinical variables. Methods such as Least Absolute Shrinkage and Selection Operator (LASSO) regression were utilized to reduce the dimensionality of the data, while retaining its predictive power. This was important given the high-dimensional nature of genomic datasets, which lead to overfitting if not managed carefully. Once the key features were identified, the authors used the Cox proportional hazards model to estimate time-to-event outcomes, including recurrence-free survival and overall survival. Ensemble ML techniques were also explored to enhance predictive accuracy by combining the strength of multiple models.

A significant strength of the study lies in its rigorous validation strategy. The authors conducted both internal validation using cross-validation techniques and external validation with independent datasets. This dual validation ensured the models were not only accurate, but also generalizable across diverse patient populations.

The findings were striking, models integrating genomic and clinical data significantly outperformed those based solely on clinical variables. The integrated models demonstrated higher metrics, reflecting their superior ability to rank patients by their risk of recurrence and decease. Furthermore, the models provided individualized risk scores, enabling a more nuanced assessment of each patient's prognosis.

The study also underscores the practical implications of integrating genomic data into clinical workflows. By combining traditional and molecular insights, the proposed models represent a major advancement in personalized oncology. Nevertheless, they also presented several limitations, such as bias introduction and uncertainty of utility in real-world clinical settings. Additionally, the computational complexity of handling high-dimensional genomic data presents practical challenges, requiring sophisticated infrastructure and expertise.

Overall, the study represents a significant contribution to the field, providing a foundation for future work aimed at integrating molecular and clinical data for personalized risk prediction. This aligns with broader trends in oncology, where the focus is increasingly shifting towards understanding the unique molecular landscape of each patient's disease to guide their treatment and improve outcomes.

2.2 Further interests for our research

The focus of current literature is on developing AI-based risk analyses for breast cancer. Nonetheless, [34] delves into the development of a comprehensive pipeline for breast cancer risk prediction, by reviewing ML approaches to the problem and highlighting their shortcomings. It also remarks the importance of integrating diverse features and ensuring models are both interpretable and fair.

The authors systematically examined 20 studies out of an initial pool of 600, focusing on those which employed ML techniques for breast cancer risk assessment. They categorized the variables used into imaging and non-imaging types. In particular, non-imaging features encompassed clinical factors, such as age, family history, and genetic information, providing a holistic view of an individual's risk profile.

In terms of model selection, the review identified various ML algorithms utiilized across the studies, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN) and ensemble methods. However, the authors also pointed out the challenges associated with model selection, such as the risk of overfitting and the need for large, diverse datasets to train robust models.

Feature extraction was another critical topic of discussion, since the authors emphasized the importance of selecting relevant variables to improve model accuracy and generalizability. Techniques like Sequential Forward Floating Selection (SFFS[2]) and Principal Component Analysis (PCA) were highlighted as effective methods for dimensionality reduction and feature selection, ensuring models focus on the most informative attributes.

[2] https://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

Furthermore, the review underscored the necessity of evaluating models for potential biases, particularly those arising from imbalanced datasets. Ensuring models are trained on diverse populations is vital to develop equitable risk prediction tools generalizable across different demographic groups. Thus, fairness assessment in ML models is crucial, particularly in medical applications where biased predictions can lead to disparities in healthcare outcomes.

On top of that, interpretability and explainability of ML models were also focal points of the review. The authors discussed the application of Explainable Artificial Intelligence (XAI[3]) techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME[4]), which help in elucidating the decision-making process of complex models. These methods enhance transparency, allowing clinicians to understand and trust the predictions made by the models, thereby facilitating their integration into clinical practice.

In conclusion, the review offers a comprehensive overview of the shortcomings of current ML-based breast cancer risk prediction models. The insights provided serve as a valuable guide for researchers and clinicians aiming to develop and implement effective, fair, and transparent breast cancer risk prediction tools.

[3] https://www.ibm.com/think/topics/explainable-ai

[4] https://christophm.github.io/interpretable-ml-book/lime.html

Chapter 3

Materials and methods

The most difficult part of Data Science is arguably gathering qualitative and quantitative data. This is also the case for healthcare domain, where one has to take further considerations in order to access data. We discuss in Chapter 5 the particular inconveniences we had with this data acquisition, while in this one we detail the data used for our experiments and the followed methodology.

3.1 Data acquisition

After an extensive period of looking for datasets related to the matter at hand, we then had to choose which ones suited our interests best.

Unfortunately, the public datasets we found did not have control subjects (that is, patients not infected with breast cancer). Thus, we decided to use the one in [35], which we found in the literature, for breast cancer risk prediction. We were fully aware of its limitations before using it, but we decided to start our experiments with it regardless and move to a better dataset whenever it arrived. However, due to our lack of responses we ended up sticking to it.

For our other line of research (recurrence and mortality risk prediction) it was easier, we just reviewed all the datasets we have found and selected the one with more subjects and qualitative variables. This dataset is included in the *Cancer Imaging Archive* collection[1]. Because we focus on prediction using clinical data (no use of medical imaging), we only took the `.xlsx` file with title *Clinical and Other Features*.

3.2 Data summary

We will refer to the datasets used in [35] and from *Cancer Imaging Archive* as Lifestyle and Duke datasets, respectively. As a reference when looking at our experiments, a feature dictionary for both datasets is included in Appendix A.

3.2.1 Lifestyle dataset

As explained in [35], this dataset contains the demographic information, moderate physical activity, lesion volume and story memory recall data for each subject included in the analysis (a total of 58). Moreover, it also includes disease and treatment characteristics of breast cancer survivors (30 subjects).

[1] https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/

A great aspect about this dataset is the fact it is already clean. This obviously means no need of preprocessing it, but more importantly, there is no imputation of values in order to use some particular models, which could have a negative impact in our results. Nonetheless, some variables were removed due to being correlated to the target column `Group`. Thus, we end up using 16 out of all 22 variables.

3.2.2 Duke dataset

In contrast to the first, based on results recorded in a study, this dataset is a single-institutional, retrospective collection of 922 biopsy-confirmed invasive breast cancer patients over a decade. Not only does it contain demographic, clinical, pathology, treatment, outcomes and genomic data, but also data from screening tests.

Since we had a total of 98 features, we removed all imaging-related data alongside unnecessary features such as `Patient_ID` (not appearing in Appendix A). Furthermore, similarly to the Lifestyle dataset, we have a set of features correlated to the targets `Recurrence` and `Dead`, so we included them in another dataset. After these two removals, we start our experiments using the 69 variables that are left.

3.3 Followed methodology

We now focus on following the steps we made along the project and justify our decision-making. We divide this section in model selection, data preprocessing, metrics definition, baseline assessment and experiments.

3.3.1 Model selection

The first thing discussed after obtaining our data was which models we should use. Even though neural networks would most likely obtain the best results, we discarded them to consider models with a higher level of interpretability. Since we are defining a pipeline for potential patients, and the construction of a model heavily depends on the particular problem, we take some of the most popular ML classifiers and see which of them works best in our scenarios.

These models were not selected arbitrary, we particularly wanted to include Logistic Regression (LR) and k-Nearest Neighbors (k = 5 by default) k-NN, but also probabilistic and tree models are used. After discussing it again and again, we end up using the aforementioned ones, in addition to Decision Tree (DT), Random Forest (RF), XGBoost (XGB) and SVM. With the exception of K-NN, all these models accept a `random_state` parameter (initialized at 42), which enables reproducibility.

3.3.2 Data preprocessing

It is already mentioned in Section 3.2.1 Lifestyle dataset is already clean, so we only have to focus on the Duke dataset. Some of the models we use (LR, for instance) do not allow `NaN` values, which are caused due to no present and no conclusive (ambiguous) instances. Thus, we must resort to some missing data strategy like imputation. Nonetheless, in exchange of widening the range of models we can use, it may introduce noise in our data. For this reason, we remove all features which have more than 600 `NaN` values out of all 922 instances. This way we can reduce the risk and retain only those variables which will benefit from the imputation, as we will see later.

After changing the feature `Age` from days to years, we use the mean value of each feature to fill missing data. Although this is a typical approach, we will see it indeed improves the performance of models with the capacity of dealing with `NaN`, such as XGB. However, we only fill the columns where it makes sense to use the mean. After completing `Nottingham_grade` by definition, we use an iterative imputer to fill the remaining missing values (limiting to 10 its iterations). Additionally, we use a label encoder to transform categorical string data into numeric types.

Due to the nature of our data, we only have columns specifying the days until our target features. Thus, we have to manually add columns which will serve us as labels for our classifications. Moreover, we combine both `Days_to_local_recurrence` and `Days_to_distant_recurrence` into a single column `Days_to_recurrence` because there is not enough data to treat each variable separately. Finally, we fill all `NaN` values of our labels dataset with zeros to include them in our control instances.

3.3.3 Metrics definition

There are quite a few ways to measure the performance of a model, but we prefer to continue with a more classical machine learning approach. This does not mean we selected metrics arbitrary. After a first gathering of the most popular and used metrics in Health domain, we chose the five most relevants for our study:

- **Accuracy:** Even though accuracy can be misleading in imbalanced datasets (like ours), it still gives us an overall sense of model performance.
- **Precision:** Defined as the proportion between true positives among all predicted positives, it measures the reliability of a positive prediction. In our particular problem, false positives can lead to unnecessary interventions or anxiety, so it is crucial to assess how trustworthy our positive predictions are.
- **Recall:** This is the proportion of true positives correctly identified. It is critical when missing a true positive has severe consequences, such as undiagnosed breast cancer.
- **F1-Score:** It is defined as the harmonic mean of precision and recall. It measures the trade-off between identifying positives and avoiding false positives, making it useful while working with imbalanced datasets.
- **Matthews Correlation Coefficient (MCC):** Defined as a correlation coefficient considering all four confusion matrix outcomes, it measures the overall quality of a classification. Moreover, not only does it provide a balanced evaluation, but it is also robust to imbalanced datasets.

Test	Accuracy	Precision	Recall	F1-Score	MCC
Recurrence (NaN)	0.874 ± 0.009	0.139 ± 0.086	0.067 ± 0.054	0.088 ± 0.064	0.035 ± 0.061
Recurrence	0.890 ± 0.010	0.140 ± 0.196	0.022 ± 0.027	0.037 ± 0.046	0.021 ± 0.072
Dead (NaN)	0.924 ± 0.018	0.450 ± 0.210	0.164 ± 0.107	0.219 ± 0.124	0.225 ± 0.132
Dead	0.926 ± 0.011	0.467 ± 0.287	0.097 ± 0.035	0.150 ± 0.039	0.173 ± 0.069

FIGURE 3.1: Metrics before and after data imputation. Predictions for our target variables `Recurrence` and `Dead` using XGB.

Since we both want to check these metrics and whether our preprocessing is successful, a little test was performed. In Figure 3.1 we show the metrics while predicting both target variables (`Recurrence` and `Dead`) using data before and after imputation (features with excessive `NaN` values already removed) using XGB. Notice that metrics are shown with some unconfidence, we will explain it in the experiments section.

3.3.4 Baseline assessment

Before starting with our experiments and in order to see if the preprocessing and the models are successful, we it would be best to first compute metrics for a simple model with unprocessed data. In the case of Duke dataset specifically, this means we retain features with more than 600 `NaN`.

For the Lifestyle dataset, we chose 1-NN, since it is the simplest of all the considered models. On the other hand, for Duke we only had XGB as an option, due to being able to natively deal with `NaN`, contrary to the rest of classifiers.

3.3.5 Experiments

As mentioned in Section 1.4, our goal is to assess both risk prediction for breast cancer and, additionally, to its recurrence and mortality. For the latter ones, we also try to predict the time before the events occur, though it obviously is a much more difficult problem. For its performance to be acceptable, it requires a quality and quantity of data we do not have. We will come back to it in Chapter 5, but the limits of our data will be reflected in the performance of the models, so please take into account we focus on giving a pipeline for breast cancer prevention.

For this reason, we really put emphasis on selecting the best model for our purposes, assessing a trade-off between performance and fairness. Once we choose the model, we try to interpret it by showing the most relevant features in the classification.

In case you check the code provided in the *GitHub* repository, you will notice notebooks only contain the essential. This is because all hand-crafted methods are gathered in the `utils.py` module to both improve the reading and the running performance as much as possible.

Risk prediction

Our goal is to compare all selected models, in addition to the baseline, using the metrics defined before. Since we want to perform a 80-20 split, we take the opportunity to also include a 5-fold stratified cross-validation (which maintains the proportion of the split for both classes) to increase robustness. This is the reason why in Figure 3.1 and the ones we will see in Chapter 4, metrics are represented with a certain degree of confidence. The reason for that is we are showing the mean and standard deviation from all 5 computations, which we consider it a natural way to express it.

Additionally, we scale data to better capture its patterns. This is controled by a boolean parameter passed to the methods, which we set to `False` while computing the baseline. Furthermore, another boolean was added, this one controls whether or not we balance the training set. This is a typical approach while working with imbalanced datasets, and the most popular strategy is *SMOTE*, originally proposed in [36]. Thus, we test this technique with a view to improve our results.

Fairness assessment

In order to measure fairness, we first select some of our features as protected attributes (variables for which if we split data depending on its value the model should behave similarly). After carefully considering our features (see Appendix A), we chose `Age` for both datasets, in addition to `Mol_Subtype` and `Race_and_Ethnicity` for the Duke dataset specifically.

Unfortunately, Lifestyle dataset does not benefit from an abundancy of columns, and the ones it has are not elegible for this test. On the other hand, we wanted to also include `Tumor_Location` for Duke dataset, but changed our minds after noticing instances for both classes were equally distributed and, on top of that, we did not have any clue on which should be the privileged class.

The next step is to split data in privileged and unprivileged classes. In the case of `Mol_Subtype` and `Race_and_Ethnicity`, we count each of the values and notice there is a predominant value, which we consider as the privileged class (the more instances we have from one class, the more the model will adapt to its particular characteristics). The actual values of these features are 0 for the former (associated to `luminal-like`) and 1 for the latter (label for `white`).

We want to follow a similar approach for `Age` even though it is not a categorical feature. Hence, by defining a threshold we create a new feature (denoted as `Age_Group`) splitting data into two classes. Based on the literature, we set the threshold to 50 years, when women start gaining access to routine breast screenings (increasing the rate of detection).

We still have to see how to measure the change of behaviour between classes. To solve this issue, we select the best models from our risk prediction test and compute the same metrics as before by classes. That is, we are going to split data depending on the class of each protected attribute and assess performance for all divisions. After the test, we will be able to select our top model as the one with the best trade-off between performance and fairness.

Interpretability

Once we have selected a model based on the previous two steps (we only take one because for some models the computations are quite expensive), we focus on understanding its predictions. The most popular approach is arguably the use of SHapley Additive exPlanations (SHAP) values[2], which is a game theoretic approach to explain the output of any machine learning model. By taking the mean of the absolute value of each SHAP value, one can rank features by its importance in a prediction. Since we have too many features to represent, we only take the five most relevant.

Survival time prediction

In the cases of recurrence and mortality specifically, we predict the time before these events occur using both the target features (`Recurrence`, `Dead`) and duration ones (`Days_to_recurrence`, `Days_to_death`) using a Cox's proportional hazard model[3]. Although the results are not promising (see Chapter 4), we consider it an essential part of the pipeline and include it regardless.

[2] https://shap.readthedocs.io/en/latest/index.html

[3] https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html

Chapter 4

Results and discussion

After explaining our experiments in the previous chapter, we are going to analyze and extract conclusions from them. Although the order while working in the project was first the Duke dataset, for the pipeline it makes much more sense to start with Lifestyle dataset, so we swap their places instead. We encourage you to check out the *GitHub* repository if you face any trouble.

4.1 Breast cancer risk prediction

Following the process detailed in Chapter 3, we first compare all models using the metrics we defined before. In Figure 4.1 you can see the performance for the models without balancing the dataset through *SMOTE*, while in Figure 4.2 we show the same visualization using this strategy. One can notice we maintained the same baseline in both comparisons, since it does not make sense to augmentate data while assessing a baseline.

Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.536 ± 0.106	0.533 ± 0.133	0.533 ± 0.194	0.526 ± 0.170	0.062 ± 0.216
Logistic Regression	0.673 ± 0.110	0.676 ± 0.107	0.667 ± 0.183	0.667 ± 0.148	0.343 ± 0.219
Decision Tree	0.447 ± 0.091	0.467 ± 0.081	0.500 ± 0.105	0.482 ± 0.091	-0.115 ± 0.190
Random Forest	0.503 ± 0.144	0.439 ± 0.238	0.567 ± 0.309	0.495 ± 0.268	-0.023 ± 0.323
XGBoost	0.433 ± 0.126	0.441 ± 0.146	0.433 ± 0.170	0.434 ± 0.155	-0.145 ± 0.253
SVM	0.656 ± 0.108	0.626 ± 0.072	0.800 ± 0.163	0.701 ± 0.106	0.331 ± 0.244
K-Nearest Neighbors	0.482 ± 0.094	0.500 ± 0.069	0.667 ± 0.105	0.570 ± 0.082	-0.061 ± 0.208

FIGURE 4.1: Comparison between models while predicting the risk of breast cancer (without *SMOTE*).

Half of the models slightly benefit from the data augmentation, specially LR. As we can see in both figures, an increase in precision (which means higher rate while predicting positives) led to a better overall performance. Consequently, it seems the use of *SMOTE* really helps in this particular case. However, as we are going to exemplify next, if we want to improve the overall performance of all models, an abundance of data is required (in order to really capture data relationships for the minority class).

Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.536 ± 0.106	0.533 ± 0.133	0.533 ± 0.194	0.526 ± 0.170	0.062 ± 0.216
Logistic Regression	0.708 ± 0.145	0.733 ± 0.170	0.667 ± 0.183	0.695 ± 0.175	0.416 ± 0.294
Decision Tree	0.483 ± 0.119	0.525 ± 0.131	0.533 ± 0.067	0.521 ± 0.074	-0.037 ± 0.261
Random Forest	0.488 ± 0.173	0.469 ± 0.271	0.467 ± 0.245	0.463 ± 0.251	-0.045 ± 0.374
XGBoost	0.450 ± 0.117	0.458 ± 0.127	0.433 ± 0.170	0.438 ± 0.149	-0.106 ± 0.239
SVM	0.570 ± 0.093	0.589 ± 0.110	0.633 ± 0.067	0.606 ± 0.074	0.131 ± 0.206
K-Nearest Neighbors	0.398 ± 0.140	0.414 ± 0.146	0.467 ± 0.194	0.437 ± 0.166	-0.216 ± 0.277

FIGURE 4.2: Comparison between models while predicting the risk of breast cancer (with *SMOTE*).

For both RF and SVM some metrics improve while other diminish, but any of these changes is of big magnitude. Nonetheless, 5-NN experiences a decreasing in all metrics, reflecting the fact that *SMOTE* does indeed not work for all models in this particular scenario.

Taking a deeper look to Figure 4.1, the best performing models in this first phase are without a doubt LR and SVM, which move on to the fairness test. Remember that for the Lifestyle dataset we only considered `Age_Group` as a protected attribute, so the comparison will be much easier than latter.

For LR, we show in Figure 4.3 which are the metrics when splitting data in classes depending on the value of the protected attribute. Similarly, in Figure 4.4 we show the same kind of results but for SVM instead.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.667 ± 0.380	0.333 ± 0.422	0.400 ± 0.490	0.360 ± 0.445	0.275 ± 0.497
Age Group	Privileged	0.668 ± 0.123	0.703 ± 0.200	0.697 ± 0.127	0.694 ± 0.151	0.326 ± 0.254

FIGURE 4.3: Metrics for LR when splitting data depending on the value of the protected attribute `Age_Group` (privileged class above 50 years) for breast cancer risk prediction.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.450 ± 0.332	0.180 ± 0.223	0.400 ± 0.490	0.248 ± 0.305	-0.052 ± 0.290
Age Group	Privileged	0.756 ± 0.168	0.756 ± 0.160	0.840 ± 0.196	0.783 ± 0.153	0.539 ± 0.333

FIGURE 4.4: Metrics for SVM when splitting data depending on the value of the protected attribute `Age_Group` (privileged class above 50 years) for breast cancer risk prediction.

Even though SVM gets better metrics for the privileged class, the disparity of performance is clearly greater than for LR. In fact, the latter has almost the same overall performance (accuracy) for both classes, though the rest of metrics suffer from reductions of around 50% in most cases. In spite of this, it seems reasonable to pick LR as the best model when considering the trade-off between performance and classes disparity.

Finally, we show which features are the most relevant for prediction when using LR, which can be seen in Figure 4.5. The most important feature is `Story_Memory_Recall`, which might be surprising at first glance, but it is actually an indirect effect of cancer. In fact, in [37] they show a cognitive decline among cancer survivors, most likely due to the nature of its aggressive treatments. Hence, it seems the model regarded it as an important factor to assess breast cancer. Moreover, cognitive decline is further enhanced by aging, so it is no wonder why it happens to be another of these features.

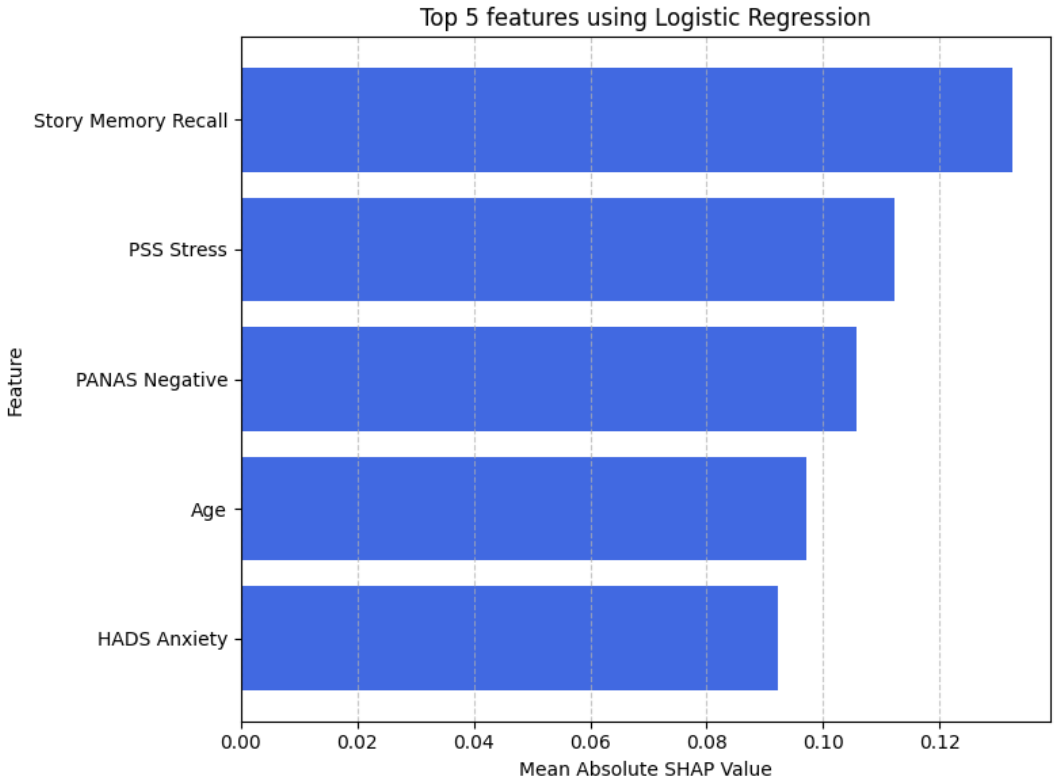

FIGURE 4.5: Plot of the 5 most relevant features for breast cancer risk prediction when using LR (using SHAP values).

On the other hand, the rest of the features (expect `Age`) reflect psychosocial stressors, which are intricately linked to physical health. In particular, chronic stress and negative affect are thought to exacerbate cancer risk through immune suppression, chronic inflammation and unhealthy lifestyle behaviours. One can also notice all of these may be aggravated by aging; thus, it is indeed consistent the appearance of `Age` as a relevant feature.

4.2 Recurrence and mortality prediction

The goal of this section is to follow the same procedure as in the previous one. However, due to the nature of our experiments and in order to maintain the structure by not mixing explanations, we will focus first on recurrence and, afterwards, on mortality. Notice almost all visualizations will be of the same kind as the ones from the previous section (though more models are considered for fairness assessment), with the single exception of the survival time prediction plot at the end.

4.2.1 Target 1. Recurrence

As before, the first step is to compare the performance of all models with our processed data, seen in Figure 4.6, and using *SMOTE* to balance the dataset, shown in Figure 4.7, which hopefully will improve the results.

Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.889 ± 0.011	0.367 ± 0.323	0.056 ± 0.000	0.090 ± 0.009	0.091 ± 0.071
Logistic Regression	0.897 ± 0.011	0.200 ± 0.400	0.022 ± 0.044	0.040 ± 0.080	0.042 ± 0.139
Decision Tree	0.801 ± 0.018	0.097 ± 0.018	0.122 ± 0.022	0.108 ± 0.018	-0.002 ± 0.024
Random Forest	0.900 ± 0.003	0.200 ± 0.245	0.022 ± 0.027	0.040 ± 0.049	0.047 ± 0.078
XGBoost	0.890 ± 0.010	0.140 ± 0.196	0.022 ± 0.027	0.037 ± 0.046	0.021 ± 0.072
SVM	0.901 ± 0.002	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.005 ± 0.010
K-Nearest Neighbors	0.899 ± 0.008	0.200 ± 0.400	0.011 ± 0.022	0.021 ± 0.042	0.032 ± 0.098

FIGURE 4.6: Comparison between models while predicting the risk of recurrence (without *SMOTE*).

Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.889 ± 0.011	0.367 ± 0.323	0.056 ± 0.000	0.090 ± 0.009	0.091 ± 0.071
Logistic Regression	0.674 ± 0.022	0.144 ± 0.023	0.478 ± 0.097	0.221 ± 0.038	0.109 ± 0.057
Decision Tree	0.280 ± 0.122	0.101 ± 0.020	0.778 ± 0.099	0.177 ± 0.031	-0.010 ± 0.112
Random Forest	0.378 ± 0.109	0.105 ± 0.012	0.700 ± 0.083	0.182 ± 0.019	0.018 ± 0.075
XGBoost	0.171 ± 0.047	0.103 ± 0.005	0.967 ± 0.044	0.186 ± 0.008	0.058 ± 0.041
SVM	0.830 ± 0.041	0.164 ± 0.061	0.156 ± 0.054	0.156 ± 0.056	0.065 ± 0.071
K-Nearest Neighbors	0.644 ± 0.041	0.116 ± 0.029	0.389 ± 0.070	0.178 ± 0.042	0.040 ± 0.067

FIGURE 4.7: Comparison between models while predicting the risk of recurrence (with *SMOTE*).

We first notice that the use of *SMOTE* does not seem to particularly help in any case. It certainly improves recall in all instances and precision for the worst values, but it is not worth when considering how much the overall performance gets diminished (with the exception of SVM).

When selecting which models should move on to the fairness step, we should immediatly discard SVM. It is indeed the model with the best accuracy, but the rest of the metrics make it inelegible. Next, even though DT gets the highest values for recall and F1-score, the disparity between it and models with better overall performance (almost by 10%) for precision is similar to the one for the metrics which DT is superior, so we also discard it.

Due to the fact that for Duke dataset we consider three protected attributes, we should discard at least one more model. Since the rest of them are better than the others in some metric, we ruled LR out. The cause is it almost gets the same values as RF, but in all cases with higher standard deviation.

The next step is fairness assessment, where we are going to show the disparity of metrics when splitting data depending on the value of the protected attributes selected in the previous chapter. One can see these tables in Figures 4.8, 4.9 and 4.10 for RF, XGB and 5-NN, respectively.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.888 ± 0.027	0.100 ± 0.200	0.020 ± 0.040	0.033 ± 0.067	0.031 ± 0.076
Age Group	Privileged	0.912 ± 0.021	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.011 ± 0.014
Mol Subtype	Unprivileged	0.868 ± 0.022	0.100 ± 0.200	0.022 ± 0.044	0.036 ± 0.073	0.030 ± 0.080
Mol Subtype	Privileged	0.916 ± 0.015	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.011 ± 0.014
Race and Ethnicity	Unprivileged	0.860 ± 0.010	0.067 ± 0.133	0.033 ± 0.067	0.044 ± 0.089	0.023 ± 0.077
Race and Ethnicity	Privileged	0.916 ± 0.009	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.005 ± 0.010

FIGURE 4.8: Metrics for RF while predicting recurrence. Data was splitted depending on the value of the protected attributes `Age_Group`, `Mol_Subtype` and `Race_and_Ethnicity`.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.880 ± 0.034	0.100 ± 0.200	0.020 ± 0.040	0.033 ± 0.067	0.009 ± 0.089
Age Group	Privileged	0.904 ± 0.013	0.200 ± 0.400	0.013 ± 0.027	0.025 ± 0.050	0.023 ± 0.111
Mol Subtype	Unprivileged	0.865 ± 0.026	0.067 ± 0.133	0.022 ± 0.044	0.033 ± 0.067	0.016 ± 0.057
Mol Subtype	Privileged	0.904 ± 0.018	0.100 ± 0.200	0.017 ± 0.033	0.029 ± 0.057	0.007 ± 0.085
Race and Ethnicity	Unprivileged	0.857 ± 0.017	0.100 ± 0.200	0.033 ± 0.067	0.050 ± 0.100	0.022 ± 0.114
Race and Ethnicity	Privileged	0.905 ± 0.019	0.200 ± 0.400	0.015 ± 0.031	0.029 ± 0.057	0.026 ± 0.121

FIGURE 4.9: Metrics for XGB while predicting recurrence. Data was splitted depending on the value of the protected attributes `Age_Group`, `Mol_Subtype` and `Race_and_Ethnicity`.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.883 ± 0.037	0.200 ± 0.400	0.017 ± 0.033	0.031 ± 0.062	0.031 ± 0.122
Age Group	Privileged	0.914 ± 0.021	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.005 ± 0.010
Mol Subtype	Unprivileged	0.862 ± 0.041	0.200 ± 0.400	0.020 ± 0.040	0.036 ± 0.073	0.037 ± 0.134
Mol Subtype	Privileged	0.916 ± 0.012	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.010 ± 0.012
Race and Ethnicity	Unprivileged	0.859 ± 0.033	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.019 ± 0.038
Race and Ethnicity	Privileged	0.914 ± 0.009	0.100 ± 0.200	0.018 ± 0.036	0.031 ± 0.062	0.027 ± 0.081

FIGURE 4.10: Metrics for 5-NN while predicting recurrence. Data was splitted depending on the value of the protected attributes `Age_Group`, `Mol_Subtype` and `Race_and_Ethnicity`.

It can be noticed the overall performance (accuracy) is similar in all three cases. However, for RF and 5-NN, one of the classes gets horrible metrics. One could argue it is due to the fact that there are very few cases in these classes, and the test set does not contain any. Nonetheless, we know the splits are the same across models (same `random_state`) and it does not happen for XGB, so we discard both RF and 5-NN.

We now proceed to the interpretability step, where we show the most important features for our predictions. Figure 4.11 represents the top 5 variables while using XGB. The most relevant feature in this context is `Days_to_Surgery`, due to the fact that longer delay between diagnosis and surgery can allow the tumor to grow and potentially metastasize. Tumor cells might also spread locally or systemically during this period, increasing the risk of recurrence.

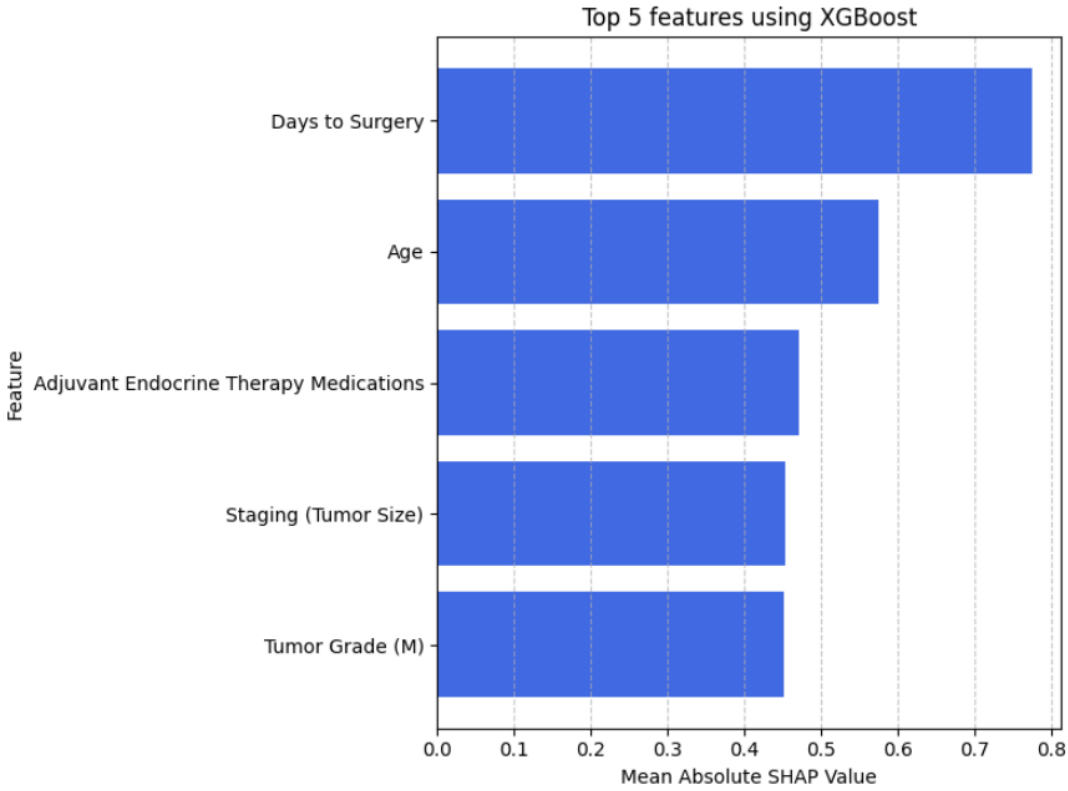

FIGURE 4.11: Plot of the 5 most relevant features for recurrence prediction when using XGB (using SHAP values).

Moving on, adjuvant endocrine therapy is used to reduce recurrence in hormone receptor-positive breast cancers. It is critical because adherence to therapy significantly lowers recurrence risk, while its type and duration influence the degree of risk reduction. This is worsened by aging, due to the incapability of receiving aggresive treatment regimens.

Moreover, larger tumors are associated with higher recurrence risk due to an increased likelihood of residual disease post-surgery and greater probability of micrometastasis. In addition, tumor size is a critical component of the TNM staging system and directly correlates with disease severity and recurrence likelihood, and larger tumors often require more aggresive adjuvant treatments.

Finally, tumor grade reflects the histological severity of the cancer, describing how abnormal the cells look. Higher grades often correspond to increased proliferation rates, genomic instability and resistance to therapies, all of which increase the likelihood of recurrence.

Most, if not all, of the aforementioned features are further enhanced by `Age`, so its appearance is consistent in both contracting cancer and relapsing. Hence, it seems reasonable to expect it is also a critical factor for mortality prediction.

The last experiment performed was a survival time prediction, which can be seen in Figure 4.12. For recurrence specifically, it means the time before relapsing. Notice that the tendency of the model is to overpredict when the actual time is less than 850 days, and to underpredict otherwise.

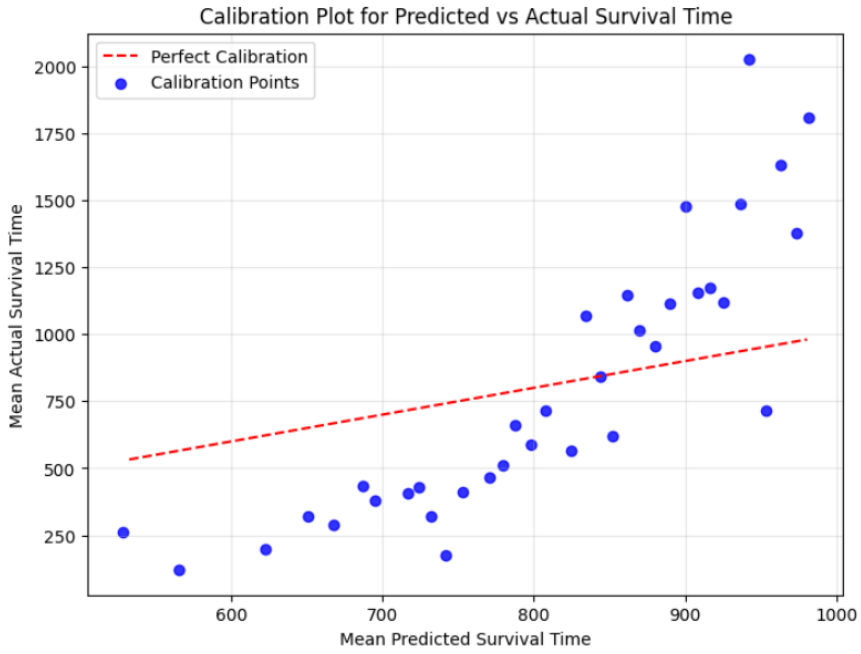

FIGURE 4.12: Plot of time estimation before `Recurrence` occurs.

We already foreshadowed in Chapter 3 the performance would not be great, the reason behind it being these types of models require more quantity of data than the one at our disposal. Specially, since it is common in this domain to have datasets of more than eigthy thousand instances, while Duke contains a couple of hundreds.

4.2.2 Target 2. Mortality

Similarly to the previous two experiments, we first compare in Figure 4.13 the models without balancing the dataset and, afterwards, their performance when using *SMOTE*, as ilustrated in Figure 4.14.

Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.932 ± 0.011	0.557 ± 0.259	0.128 ± 0.037	0.202 ± 0.059	0.238 ± 0.096
Logistic Regression	0.935 ± 0.009	0.660 ± 0.280	0.163 ± 0.055	0.250 ± 0.084	0.292 ± 0.100
Decision Tree	0.905 ± 0.026	0.331 ± 0.151	0.308 ± 0.066	0.312 ± 0.098	0.266 ± 0.113
Random Forest	0.938 ± 0.003	0.733 ± 0.389	0.096 ± 0.060	0.166 ± 0.098	0.248 ± 0.130
XGBoost	0.926 ± 0.011	0.467 ± 0.287	0.097 ± 0.035	0.150 ± 0.039	0.173 ± 0.069
SVM	0.933 ± 0.002	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
K-Nearest Neighbors	0.936 ± 0.009	0.500 ± 0.447	0.082 ± 0.075	0.141 ± 0.128	0.190 ± 0.182

FIGURE 4.13: Comparison between models while predicting the risk of death (without *SMOTE*).

Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.932 ± 0.011	0.557 ± 0.259	0.128 ± 0.037	0.202 ± 0.059	0.238 ± 0.096
Logistic Regression	0.761 ± 0.017	0.160 ± 0.029	0.595 ± 0.097	0.252 ± 0.044	0.213 ± 0.062
Decision Tree	0.357 ± 0.210	0.070 ± 0.016	0.660 ± 0.178	0.125 ± 0.027	-0.016 ± 0.092
Random Forest	0.595 ± 0.192	0.106 ± 0.018	0.614 ± 0.233	0.174 ± 0.025	0.116 ± 0.033
XGBoost	0.239 ± 0.125	0.074 ± 0.008	0.888 ± 0.109	0.137 ± 0.013	0.053 ± 0.057
SVM	0.899 ± 0.025	0.248 ± 0.106	0.192 ± 0.078	0.208 ± 0.075	0.162 ± 0.086
K-Nearest Neighbors	0.764 ± 0.014	0.149 ± 0.010	0.532 ± 0.031	0.233 ± 0.014	0.183 ± 0.020

FIGURE 4.14: Comparison between models while predicting the risk of death (with *SMOTE*).

Here we observe a similar behaviour than before: balancing of the dataset improves all instances of recall, while it fails for the rest of the metrics. It certainly increases almost all values in the case of SVM, for which it obtains quite interesting results. However, for the other models it diminishes so much the overall performance we consider it detrimental rather than helpful.

Next, we must select the models which are going to advance to the second step. Since *SMOTE* is not used in further computations, discarding SVM is an easy call. Moreover, although 5-NN almost ties in accuracy with LR, it is inferior to the latter for all other metrics, so we also rule it out.

Following the same logic as for recurrence, we should discard at least another model. Notice that LR and RF lead both accuracy and precision, while DT is on top for recall and F1-score. Thus, we remove XGB from our selection.

We now move on to the next step, fairness assessment. In a similar way as we did in the previous two experiments, we are going to show the disparity of performance when splitting data depending on the values of our predefined protected attributes. This is represented in Figures 4.15, 4.16 and 4.17, each of them showing metrics using LR, DT and RF, respectively. One can observe it is a much more difficult situation that the previous ones, where we only had one protected attribute or could discard a model at first glance.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.948 ± 0.020	0.467 ± 0.400	0.233 ± 0.200	0.311 ± 0.267	0.319 ± 0.277
Age Group	Privileged	0.924 ± 0.010	0.500 ± 0.447	0.126 ± 0.112	0.189 ± 0.159	0.213 ± 0.207
Mol Subtype	Unprivileged	0.898 ± 0.025	0.633 ± 0.306	0.186 ± 0.046	0.268 ± 0.056	0.286 ± 0.088
Mol Subtype	Privileged	0.955 ± 0.006	0.500 ± 0.447	0.119 ± 0.109	0.189 ± 0.170	0.228 ± 0.210
Race and Ethnicity	Unprivileged	0.912 ± 0.031	0.500 ± 0.447	0.187 ± 0.165	0.256 ± 0.215	0.279 ± 0.234
Race and Ethnicity	Privileged	0.945 ± 0.013	0.547 ± 0.394	0.175 ± 0.121	0.245 ± 0.155	0.274 ± 0.187

FIGURE 4.15: Metrics for LR while predicting mortality. Data was splitted depending on the value of the protected attributes `Age_Group`, `Mol_Subtype` and `Race_and_Ethnicity`.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.912 ± 0.015	0.205 ± 0.121	0.267 ± 0.200	0.219 ± 0.140	0.183 ± 0.146
Age Group	Privileged	0.877 ± 0.036	0.281 ± 0.133	0.278 ± 0.043	0.257 ± 0.057	0.206 ± 0.075
Mol Subtype	Unprivileged	0.870 ± 0.051	0.373 ± 0.131	0.330 ± 0.175	0.330 ± 0.112	0.272 ± 0.132
Mol Subtype	Privileged	0.905 ± 0.028	0.154 ± 0.121	0.159 ± 0.093	0.150 ± 0.102	0.104 ± 0.109
Race and Ethnicity	Unprivileged	0.875 ± 0.033	0.267 ± 0.154	0.306 ± 0.207	0.283 ± 0.175	0.219 ± 0.172
Race and Ethnicity	Privileged	0.901 ± 0.034	0.214 ± 0.148	0.222 ± 0.150	0.217 ± 0.148	0.165 ± 0.164

FIGURE 4.16: Metrics for DT while predicting mortality. Data was splitted depending on the value of the protected attributes `Age_Group`, `Mol_Subtype` and `Race_and_Ethnicity`.

Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.946 ± 0.015	0.400 ± 0.490	0.067 ± 0.082	0.114 ± 0.140	0.158 ± 0.194
Age Group	Privileged	0.930 ± 0.013	0.400 ± 0.490	0.077 ± 0.111	0.125 ± 0.174	0.159 ± 0.218
Mol Subtype	Unprivileged	0.908 ± 0.021	0.600 ± 0.490	0.087 ± 0.077	0.151 ± 0.131	0.216 ± 0.181
Mol Subtype	Privileged	0.953 ± 0.006	0.400 ± 0.490	0.057 ± 0.070	0.100 ± 0.122	0.144 ± 0.183
Race and Ethnicity	Unprivileged	0.912 ± 0.030	0.400 ± 0.490	0.100 ± 0.133	0.157 ± 0.204	0.191 ± 0.241
Race and Ethnicity	Privileged	0.948 ± 0.006	0.500 ± 0.447	0.084 ± 0.071	0.140 ± 0.116	0.193 ± 0.160

FIGURE 4.17: Metrics for RF while predicting mortality. Data was splitted depending on the value of the protected attributes `Age_Group`, `Mol_Subtype` and `Race_and_Ethnicity`.

Firstly, we notice all three models perform similarly, so we rely on particular observations to select the best. For example, DT seems to be consistent for both `Age_Group` and `Race_and_Ethnicity`. However, for `Mol_Subtype` the disparity in metrics is greater than for the other two models (for precision, recall and F1-score values are doubled from one class to the other). Thus, we discarded it.

To rule out one from the two remaining models, we focus on precision, recall and F1-score, since accuracy is almost the same for all instances. For `Race_and_Ethnicity`, LR slightly outperforms RF, while the latter beats the former for `Age_Group`. Finally, in `Mol_Subtype` we can see bigger difference in values, with LR getting better metrics.

In spite of the fact that RF slightly underperforms the aforementioned model, it is certainly the most robust option. Taking a deeper look to the values, one can notice standard deviation is more consistent for both classes in this case. Notice that this supports the idea that not only does the model should perform similarly for both classes to be fair, but also has to return the same level of uncertainty amongst all subgroups.

After selecting RF as our model, we now proceed to interpretate its predictions. Figure 4.18 shows the top 5 relevant features while predicting mortality, following the same procedure as in both previous experiments. Then, we justify their appearance. Notice that, as we were expecting, `Age` happens to also be an important fatality factor.

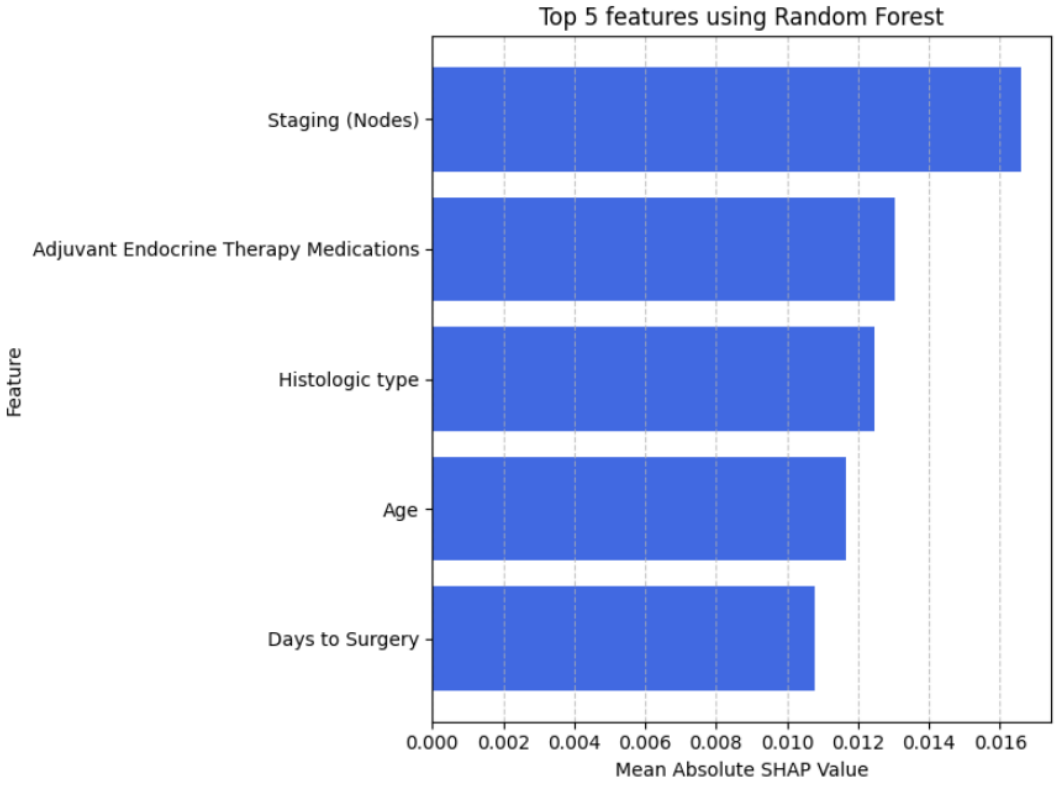

FIGURE 4.18: Plot of the 5 most relevant features for mortality prediction when using RF (using SHAP values).

First of all, staging of lymph nodes reflects the extent of cancer spread within the lymphatic system. A higher number of affected nodes is associated with advanced disease and higher mortality risk, while also preceding systemic spread. Hence, it is one of the strongest predictors of breast cancer survival (which we already know for being part of the TNM staging system).

Next, as we already discussed for the recurrence case, adjuvant endocrine therapy is highly effective for hormone receptor-positive breast cancer. Its relevance also extends to mortality, since inconsistent or incomplete therapy can increase the risk of disease progression and death.

Similarly to the previous experiment, delayed surgery can allow the cancer to grow and potentially metastasize. Moreover, patients with advanced disease may require more preoperative evaluation, contributing to longer delays and, of course, with worse survival outcomes.

Moving on to the last of these features, `Histologic_type` describes the microscopic structure and characteristics of cancer cells, with certain types having better or worse prognoses. It heavily influences treatment decisions, response to therapy, and likelihood of metastasis, all of which directly affect mortality risk.

Finally, we perform a survival time prediction, which is shown in Figure 4.19. Its behaviour presents the same tendencies as for recurrence: there are two periods for which the model overpredicts in the first one, while mostly underperforming in the other. One can also notice that the threshold is now around 800 days, which seems to reflect that severe patients generally stay alive less time than cancer survivors remain healthy before relapsing.

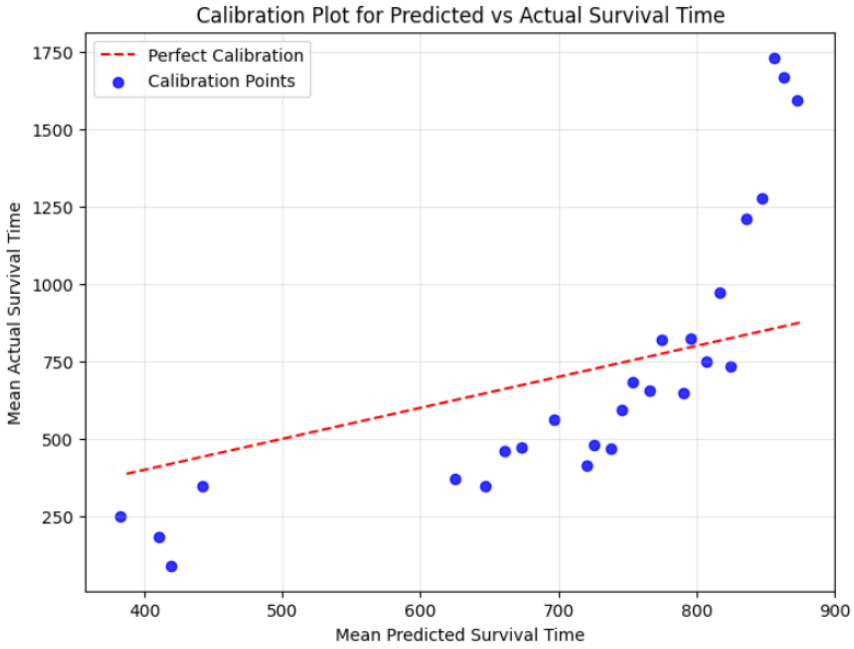

FIGURE 4.19: Plot of time estimation before the event `Dead` occurs.

4.3 Further discussion

As we have already explained in previous chapters, our goal in this thesis is not to obtain the best model for each particular scenario. Rather, we want to test the robustness of our pipeline for these specific cases. This is also one of the reasons why we selected RF for mortality prediction, since we wanted to show the performance of the whole process with as many models as possible.

Summing up our findings, LR was the top performer for breast cancer risk prediction, XGB for recurrence prediction and RF for mortality prediction. We are fully aware the obtained metrics are not acceptable in healthcare domain (specially in the first experiment), since we are dealing with human lives and the requirements for a model to be elegible are quite high. However, we are trying to solve a particularly difficult problem, and we were able to reflect through our metrics that models are indeed learning from data.

Furthermore, the interpretability step was a success in all three experiments, where each feature was consistent with its relevance and `Age` was proved to be one of the most important risk factors for breast cancer.

Overall, we ended up satisfied with our results, while being optimistic on the performance and usefulness of the pipeline for different datasets. In spite of this fact, we consider there are some improvable aspects, which we will include in the next chapter alongside with how we think this research should continue.

Chapter 5

Conclusions

After discussing our work and the results extracted from it, we focus on connecting the dots left behind during this thesis. In particular, we are going to list the obstacles we faced, remark the limit of our findings and, perhaps most importantly, how this project can be improved and used in the future.

5.1 Limitations of our research

The aim of this project was initially to work with microbiome data in order to potentially discover new risk factors for breast cancer. We obviously needed first such data (which is becoming increasingly available, but requires access permission), so we contacted as many databanks as possible.

We started with the ones featuring microbiome data, but as time went by we decided it would be for the best to also get in touch with databanks which could provide us qualitative and quantitative data for breast cancer specifically. Our intention was, then, switched to just discovering new risk factors, not necessarily coming from microbiome data.

Unfortunately, very few people actually answered us, and the ones who did it took at least a week per message. Moreover, some of the datasets had access fees, which obviously required careful discussion. Adding these to the fact that we had to constantly fill documents for bureaucratic reasons (which needed to be signed by some faculty responsible we had to find), we realised we were running out of time.

Hence, we lowered our expectations again and searched for public access data which could be used for our interests. We discarded imaging data, since we wanted to retain the essence of discovering (or, at least, highlighting) risk factors. After an extensive period of searching, we selected the aforementioned Lifestyle and Duke datasets for our experiments. In particular, Duke was the most complete dataset among the ones containing breast cancer patients, while Lifestyle was the only one we found with control subjects (and, thus, the only one elegible for predicting breast cancer risk).

It was in this moment when we noticed we could present a pipeline for breast cancer, and started our experiments with the delusion of moving to bigger datasets when access to them was granted. By the time Christmas holidays arrived, the few people responding us ceased to do so. We then had no other option than to move on and stick to the data available for us, which we know it would affect to the overall performance of the pipeline.

The experiments themselves did not present any particular issue, apart from the usual learning process when working with libraries and data new to you. Nevertheless, it is really frustating for us that the overall success of the project is compromised due to not having bigger datasets at our disposal. Ironically, after vacation was over, we finally gained access to the PLCO dataset[1].

This happened exactly one week prior to the delivery, so we quickly explored the data to see if we could improve our results and support our findings with it. Sadly, the provided data contained a main dataset with more than 70,000 subjects and more than a hundred of features. Hence, we did not have enough time to understand and process this data in order to use it for our purposes.

5.2 Future work

The most noticeable shortcoming of our project is the lack of testing for bigger datasets, which are common in healthcare domain. For this reason, the first step would be to follow the same procedure as presented in this thesis with such data (for instance, we could use the PLCO dataset previously mentioned) and see whether our findings are consistent with it or not.

Next, we could try to improve the overall performance of the pipeline by adapting the models to the data. That is, given a particular data, finding the model which best adapts to it. For risk prediction specifically, one could aim to specialise predictions by selecting the best model for each class of the protected attributes. Not only would it increase performance, but also would be the best solution to the fairness assessment issue. However, it is currently a pipe dream, since it requires big datasets like PLCO for each class.

Another field of interest may be the integration of neural networks to the pipeline. We decided it would take us too much time for it to work, but with proper time and dedication one could construct a network adapted to some particular data. With further effort, outcomes could also be interpreted and explained (see [38]), which would make the network elegible for these type of scenarios.

As it was first planned and was originally the focus of this thesis, the pipeline could also be used to explore and discover new risk factors for breast cancer, but also for different types of cancer or, even, for other diseases (provided one gets data to do so). In particular, we would like to see how the pipeline performs when integrating microbiome data, and see which features are determined to be the most critical for predicting breast cancer risk specifically.

Outside research, once all the aforementioned proposals were implemented, we would like to share it with citizens. The final idea we have in mind is to develop an app which stores your personal health data (which some countries have already implemented) with an option to predict the risk of cancer based on your clinical history. We are fully aware it is an almost impossible task, but a huge number of people would be diagnosed in time if the project succeeded. Moreover, for terminal state patients, it may help them to better understand the situation and how to manage it.

Overall, we are confident in the potential of this project, and if time had allowed it, we would have liked to bring its progress as far as possible.

[1] https://cdas.cancer.gov/plco/

Appendix A

Feature dictionary

As mentioned in this thesis, for us it is crucial readers comprehend what we are referring to while naming a feature. Not only does this help them to better understand the interpretability part of our experiments, but also serves as a way to get a grasp on our data structure. Notice we include both training and target features, which are alphabetically ordered to facilitate the search as much as possible.

A.1 Lifestyle dataset

- **Adjuvant Treatment:** Additional cancer treatment given after the primary treatment to lower the risk that the cancer will come back. Takes the values `RadiationOnly`, `ChemotherapyOnly` and `RadChemo` for patients and `Control` for the rest of the subjects.
- **Age:** Time since birth, expressed as an integer.
- **BMI:** Body Mass Index.
- **Cancer:** Target variable, `True` if the subject suffers from breast cancer, `False` otherwise.
- **Education (Years):** Time following full-time studies.
- **Estrogen Receptor Positive:** Whether or not breas cancer has receptors for the hormone estrogen. For control subjects its value is `Not_applicable`.
- **Gray Matter Volume (mm):** Amount of gray matter.
- **Group:** Variable from which we constructed our target, it takes the values `NonCancer` and `BreastCancer`.
- **HADS Anxiety:** Hospital Anxiety and Depression Scale, it measures anxiety and depression in a general medical population of patients.
- **HADS Depression:** Same as above.
- **MMSE:** Mini-Mental State Examination, it is the best-known and the most often used short screening tool for providing an overall measure of cognitive impairment in clinical, research and community settings.
- **Moderate Physical Activity:** Level of physical exertion that falls between light and vigorous activity. It is commonly defined in terms of metabolic equivalent tasks (METs), perceived exertion, or specific examples of activities.

- **Months Since Treatment End:** Time since primary treatment finished. For control subjects it takes the value `Not_applicable`.
- **PANAS Negative:** Positive and Negative Affect Schedule, it is a self-report questionnaire that consists of two 10-item scales to measure both positive and negative affect.
- **PANAS Positive:** Same as above.
- **PSS Stress:** Perceived Stress Scale, a popular tool for measuring psychological stress. It is a self-reported questionnaire that was designed to measure the degree to which situations in one's life are appraised as stressful.
- **STAI State:** State-Trait Anxiety Inventory, it is a commonly used measure of trait and state anxiety.
- **STAI Trait**: Same as above.
- **Stage:** Current step of primary treatment, for control subjects takes the value `Not_applicable`.
- **Story Memory Recall:** A cognitive testing paradigm used to assess verbal episodic memory.
- **Surgery yes or no:** Whether or not patients received surgery, for control subjects takes the value `Not_applicable`.
- **White Matter Lesion Volume (%):** Measures the total volume of white matter hyperintensities or other lesions in the brain's white matter. These are often visible on magnetic resonance imaging scans and are quantified to assess their clinical or research significance.
- **White Matter Lesion Volume (mm):** Same as above.

A.2 Duke dataset

- **Adjuvant Anti-Her2 Neu Therapy:** Whether or not a subject follows the therapy, which consists of trastuzumab for the remainder of the one year of total anti-HER2 therapy.
- **Adjuvant Chemotherapy:** Additional cancer treatment given after the primary treatment to lower the risk that the cancer will come back.
- **Adjuvant Endocrine Therapy Medications:** Whether or not a subject takes medication for an adjuvant endocrine treatment.
- **Adjuvant Radiation Therapy:** Same as `Adjuvant_Chemotherapy`.
- **Age:** Time since birth, expressed as a floating point number.
- **Bilateral breast cancer?:** Whether or not a patient has cancer in both breasts.
- **Contralateral Breast Involvement:** If a tumor in the opposite breast was diagnosed more than 6 months following the detection of the first cancer.
- **Days known alive / to death:** If death is reported, days until decease. Otherwise, days since last time the patient was known to be alive.
- **Days to Surgery:** Time since surgery (negative values) or until surgery.

- **Days to death:** Time until confirmed decease, otherwise the value is missing.
- **Days to distant recurrence:** Time until a patient suffers from recurrence in a part of the body that is far away from where the original tumor first formed.
- **Days to last distant recurrence free assessment:** Time since last assessment that a patient is free from distant recurrence.
- **Days to last local recurrence free assessment:** Similar to above, but for local recurrence instead.
- **Days to local recurrence:** Time until a patient suffers from recurrence in a part of the body that is very close to where the original tumor first formed.
- **Days to recurrence:** Hand-crafted feature, time until recurrence of any type.
- **Dead:** Hand-crafted target feature, whether or not a patient is confirmed to be deceased.
- **ER:** Whether a patient has receptors for the hormone estrogen (ER positive) or not (ER negative).
- **HER2:** Whether a patient tests positive for a protein called Human Epidermal growth factor Receptor 2.
- **Histologic type:** Classification of the cancer based on the microscopic appearance of the cancer cells and tissues. It provides important information about the tumor's structure, behavior, and how it may respond to treatment.
- **Known Ovarian Status:** Whether or not the status of the ovaries of a patient is known.
- **Lymphadenopathy or Suspicious Nodes:** Presence of enlarged or abnormal lymph nodes that may indicate cancer. In breast cancer specifically, it is often a critical feature used to assess the extent of disease spread and help in determining prognosis and treatment plans.
- **Menopause (at diagnosis):** Whether a patient is already menopausal or not.
- **Metastatic at Presentation (Outside of Lymph Nodes):** Presence of cancer that has spread to parts of the body beyond the lymph nodes at the time of diagnosis or initial presentation.
- **Mol Subtype:** Classification system that categorizes tumors based on their gene expression profiles or the presence/absence of certain biomarkers. It helps to understand the tumor's biological behavior, response to treatment, and prognosis.
- **Multicentric/Multifocal:** Whether or not a breast cancer is multicentric (multiple tumors develop in different quadrants of the breast) and multifocal (more than one distinct tumor within the same quadrant).
- **Neoadjuvant Anti-Her2 Neu Therapy:** Use of anti-HER2 targeted treatments before the main treatment (usually surgery) for HER2-positive breast cancer.
- **Neoadjuvant Chemotherapy:** Whether or not a patient recieved the treatment before main treatment.
- **Neoadjuvant Endocrine Therapy Medications:** Whether or not a subject takes medication for a neoadjuvant endocrine treatment.

- **Neoadjuvant Radiation Therapy:** Same as `Neoadjuvant_Chemotherapy`.
- **Nottingham grade:** System used to evaluate the aggressiveness and prognosis of breast cancer. It is based on the microscopic appearance of the cancer cells, specifically assessing the tumor grade and the likelihood of it spreading. This grading system helps to classify breast cancer tumors and determine the potential behavior of the cancer, guiding treatment decisions.
- **PR:** Whether a patient has receptors for the hormone progesterone.
- **Pec/Chest Involvement:** Whether or not there is involvement of the pectoral muscles or the chest wall by a breast cancer tumor. Commonly used in clinical staging, particularly when describing the extent or spread of breast cancer during radiological imaging (such as mammography, ultrasound, or MRI) or during surgical examination.
- **Race and Ethnicity:** Race of patient, takes the values `white`, `black`, `asian`, `native`, `hispanic`, `multi`, `hawa` and `amer_indian`.
- **Received Neoadjuvant Therapy:** Whether of not a patient was administered a treatment before the main treatment (usually surgery).
- **Recurrence:** Hand-crafted target feature, whether or not the patient suffers from recurrence.
- **Recurrence event(s):** Same as above, but not used due to ambiguities with `Days_to_local_recurrence` and `Days_to_distant_recurrence`.
- **Skin/Nipple Involvement:** Whether or not the tumor has spread to the skin or nipple of the breast. This is an important clinical sign that indicates that the cancer has grown beyond the breast tissue itself and may be more advanced.
- **Staging (Metastasis):** Extent to which the cancer has spread, particularly to distant organs or tissues.
- **Staging (Nodes):** Extent to which cancer has spread to the lymph nodes, which are part of the body's immune system. When cancer cells break away from the primary tumor, they can travel through the lymphatic system and spread to nearby lymph nodes. This is a crucial factor in determining the stage of cancer.
- **Staging (Tumor Size):** Describes the size and extent of local invasion for the primary tumor. It provides important information about the stage of cancer, which helps guide treatment decisions and predict the patient's prognosis.
- **Surgery:** Whether or not a patient has undergone surgery.
- **Therapeutic or Prophylactic Oophorectomy as part of Endocrine Therapy:** Surgical removal of the ovaries, which is sometimes used in the treatment of hormone-sensitive conditions. This procedure is often considered as part of endocrine therapy to manage or reduce the risk of cancer recurrence, especially in hormone receptor-positive cancers.
- **Tumor Grade (M):** Degree of gravity for `Staging_(Metastasis)`.
- **Tumor Grade (N):** Degree of gravity for `Staging_(Nodes)`.
- **Tumor Grade (T):** Degree of gravity for `Staging_(Tumor_Size)`.
- **Tumor Location:** Side of cancer (left or right).

Appendix B

GitHub repository structure

Here we detail the structure of this project's *GitHub* repository. As one can observe in Figure B.1, it does not contain too many folders.

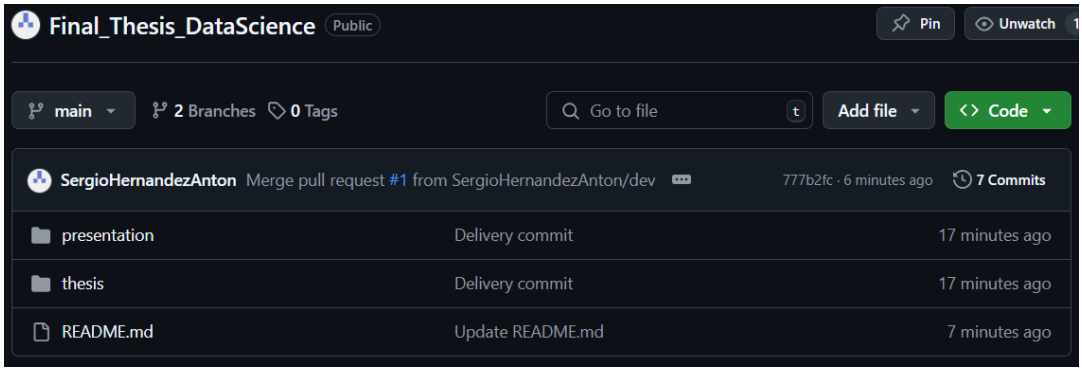

FIGURE B.1: Structure of the shared *GitHub* repository, where this document and experiments are stored.

First of all, in `README.md` you will find the abstract of the thesis and my sign attached. I also included a line where I leave constancy of the transfer of rights to *Universitat de Barcelona*.

Moving on to the folders, in `thesis` is stored all the necessary material to write this document, which is also there. Particularly, in the `code` folder appear the notebooks of my experiments (one for Lifestyle dataset and the other for Duke dataset), with the `utils.py` as mentioned in Chapter 3. On the other hand, in `data` are contained both datasets used (one can notice I retained the original name for the Lifestyle dataset).

Finally, in the `presentation` folder appear the slides I used for the thesis defense, while in the `material` folder you will find another a notebook file. It corresponds to the PLCO dataset mentioned in Chapter 5, which I used in the presentation to support my findings with a bigger dataset. It also marks the first step towards the proposed line of future work. As a remark, I cannot include the dataset due to access permission.

Bibliography

[1] U.S. Cancer Statistics Working Group. "US Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report". In: *Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute* (2012). URL: https://www.cdc.gov/uscs.

[2] Freedman AN et al. "Cancer risk prediction models: a workshop on development, evaluation, and application". In: *JNCI: Journal of the National Cancer Institute* (2005). URL: https://doi.org/10.1093/jnci/dji128.

[3] Joy JE et al. "Saving women's lives: strategies for improving breast cancer detection and diagnosis". In: *The National Academies Press* (2005). URL: https://doi.org/10.17226/11016.

[4] Decarli A et al. "Gail model for prediction of absolute risk of invasive breast cancer: independent evalucation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort". In: *JNCI: Journal of the National Cancer Institute* (2006). URL: https://doi.org/10.1093/jnci/djj463.

[5] Musa IH et al. "Artificial Intelligence and Machine Learning in Cancer Research: A Systematic and Thematic Analysis of the Top 100 Cited Articles Indexed in Scopus Database". In: *Cancer Control* (2022). URL: https://doi.org/10.1177/10732748221095946.

[6] Morgan XC et al. "Chapter 12: Human Microbiome Analysis". In: *PLOS Computational Biology* (2012). URL: https://doi.org/10.1371/journal.pcbi.1002808.

[7] Wu AH et al. "Gut microbiome associations with breast cancer risk factors and tumor characteristics: a pilot study". In: *Breast Cancer Research and Treatment* (2020). URL: https://doi.org/10.1007/s10549-020-05702-6.

[8] Huybrechts I et al. "The Human Microbiome in Relation to Cancer Risk: A Systematic Review of Epidemiologic Studies". In: *Cancer Epidemiology, Biomarkers Prevention* (2020). URL: https://doi.org/10.1158/1055-9965.epi-20-0288.

[9] Ban KA et al. "Epidemiology of breast cancer". In: *Surgical Oncology Clinics of North America* (2014). URL: https://www.surgonc.theclinics.com/article/S1055-3207(14)00028-3/abstract.

[10] Nelson HD et al. "Risk factors for breast cancer for women aged 40 to 49 years: a systematic review and meta-analysis". In: *Annals of Internal Medicine* (2012). URL: https://doi.org/10.7326/0003-4819-156-9-201205010-00006.

[11] Harvey JA et al. "Quantitative assessment of mammographic breast density: relationship with breast cancer risk". In: *Radiology* (2004). URL: https://doi.org/10.1148/radiol.2301020870.

[12] Engmann NJ et al. "Population-Attributable Risk Proportion of Clinical Risk Factors for Breast Cancer". In: *JAMA* (2017). URL: https://doi.org/10.1001/jamaoncol.2016.6326.

[13] Brentnall AR et al. "Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density". In: *JAMA* (2018). URL: https://doi.org/10.1001/jamaoncol.2018.0174.

[14] Travis LB et al. "Cumulative absolute breast cancer risk for young women treated for Hodgkin lymphoma". In: *JNCI: Journal of the National Cancer Institute* (2005). URL: https://doi.org/10.1093/jnci/dji290.

[15] Stratton MR et al. "The emerging landscape of breast cancer susceptibility". In: *Nature Genetics* (2008). URL: https://doi.org/10.1038/ng.2007.53.

[16] Gallagher S et al. "Association of a Polygenic Risk Score With Breast Cancer Among Women Carriers of High- and Moderate-Risk Breast Cancer Genes". In: *JAMA* (2020). URL: https://doi.org/10.1001/jamanetworkopen.2020.8501.

[17] Dupont WD et al. "Risk factors for breast cancer in women with proliferative breast disease". In: *The New England Journal of Medicine* (1985). URL: https://doi.org/10.1056/nejm198501173120303.

[18] Hartmann LC et al. "Benign breast disease and the risk of breast cancer". In: *The New England Journal of Medicine* (2005). URL: https://doi.org/10.1056/nejmoa044383.

[19] Toniolo PG et al. "A prospective study of endogenous estrogens and breast cancer in postmenopausal women". In: *JNCI: Journal of the National Cancer Institute* (1995). URL: https://doi.org/10.1093/jnci/87.3.190.

[20] Potischman N et al. "Reversal of relation between body mass and endogenous estrogen concentrations with menopausal status". In: *JNCI: Journal of the National Cancer Institute* (1996). URL: https://doi.org/10.1093/jnci/88.11.756.

[21] McTiernan A et al. "Physical Activity in Cancer Prevention and Survival: A Systematic Review". In: *Medicine & Science in Sports & Exercise* (2019). URL: https://doi.org/10.1249/mss.0000000000001937.

[22] Akcay M et al. "Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy". In: *Advances in Radiation Oncology* (2020). URL: https://doi.org/10.1016/j.adro.2020.07.007.

[23] Rockhill B et al. "Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention". In: *JNCI: Journal of the National Cancer Institute* (2001). URL: https://doi.org/10.1093/jnci/93.5.358.

[24] Gail MH et al. "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually". In: *JNCI: Journal of the National Cancer Institute* (1989). URL: https://doi.org/10.1093/jnci/81.24.1879.

[25] Costantino JP et al. "Validation studies for models projecting the risk of invasive and total breast cancer incidence". In: *JNCI: Journal of the National Cancer Institute* (1999). URL: https://doi.org/10.1093/jnci/91.18.1541.

[26] Gail MH et al. "Projecting individualized absolute invasive breast cancer risk in African American women". In: *JNCI: Journal of the National Cancer Institute* (2007). URL: https://doi.org/10.1093/jnci/djm223.

[27] Tyrer J et al. "A breast cancer prediction model incorporating familial and personal risk factors". In: *Statistics in Medicine* (2004). URL: https://doi.org/10.1002/sim.1668.

[28] Mavaddat N et al. "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes". In: *The American Journal of Human Genetics* (2019). URL: https://doi.org/10.1016/j.ajhg.2018.11.002.

[29] Vachon CM et al. "The contributions of breast density and common genetic variation to breast cancer risk". In: *JNCI: Journal of the National Cancer Institute* (2015). URL: https://doi.org/10.1093/jnci/dju397.

[30] Cuzick J et al. "Selective oestrogen receptor modulators in prevention of breast cancer: an updated meta-analysis of individual participant data". In: *Lancet* (2013). URL: https://doi.org/10.1016/s0140-6736(13)60140-3.

[31] Tice JA et al. "Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model". In: *Annals of Internal Medicine* (2008). URL: https://doi.org/10.7326/0003-4819-148-5-200803040-00004.

[32] Ho J et al. "Generative adversarial imitation learning". In: *Advances in neural information processing systems* (2016). URL: https://doi.org/10.48550/arXiv.1606.03476.

[33] Wu X et al. "Personalized Prognostic Prediction Models for Breast Cancer Recurrence and Survival Incorporating Multidimensional Data". In: *JNCI: Journal of the National Cancer Institute* (2017). URL: https://doi.org/10.1093/jnci/djw314.

[34] Hussain S et al. "Breast cancer risk prediction using machine learning: a systematic review". In: *Frontiers in Oncology* (2024). URL: http://dx.doi.org/10.3389/fonc.2024.1343627.

[35] Cooke GE et al. "Moderate Physical Activity Mediates the Association between White Matter Lesion Volume and Memory Recall in Breast Cancer Survivors". In: *PLoS One* (2016). URL: https://doi.org/10.1371/journal.pone.0149552.

[36] Chawla NV et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* (2002). URL: https://doi.org/10.1613/jair.953.

[37] Scott SB et al. "Memory lapses in daily life among breast cancer survivors and women without cancer history". In: *Psychooncology* (2020). URL: https://doi.org/10.1002/pon.5357.

[38] Somani A et al. "Interpretability in Deep Learning". In: *Springer Cham* (2023). URL: http://dx.doi.org/10.1007/978-3-031-20639-9.