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 16, 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, model selection, feature extraction, interpretability, and explainability, to ensure accurate, transparent, and actionable outcomes.

In Chapter 1 we introduce the burden of breast cancer, highlighting the need for personalized risk models. In it is also explained the importance of recurrence and survival inclusion, while also outlining this project goals. Next, in Chapter 2 we review the state of the art in breast cancer detection and prognosis from a machine learning perspective, paying a special attention to GAIL models.

We present the pipeline in Chapter 3, where we also detail the material we are using and the followed methodology. We evaluate diverse machine learning models, prioritizing predictive performance and fairness, considering protected attributes like age or race. Interpretability is a cornerstone of the pipeline, achieved through SHAP values and calibration plots, ensuring predictions are comprehensible to both clinicians and researchers. Moreover, survival analysis methods, specifically Cox proportional hazards modeling, are employed to predict recurrence and mortality.

Results are presented in Chapter 4, and validate the pipeline's effectiveness, with evaluation metrics demostrating its predictive power and reliability across diverse patient subgroups. The integration of recurrence and mortality predictions into the risk framework highlights its potential to inform both preventative and therapeutic decision-making. Finally, in Chapter 5 we conclude reflecting on the implications of our findings, acknowledging limitations, and suggestion a line for future research.

Additionally, in an attempt to make this thesis more reachable, we added a feature dictionary for both used datasets in Appendix A. On top of that, we also shared our project in the shape of a *GitHub* repository[1] (check Appendix B for a guide on it), so that people can take profit of this research if at all possible.

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.

[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 bring this project to the point you envisioned.

Next, to all people working for making *Universitat de Barcelona* a better place. It is partially thanks to you 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 to 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] and it represents about 12% of all new cancer cases and 25% of all cancers in women [2]. 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.

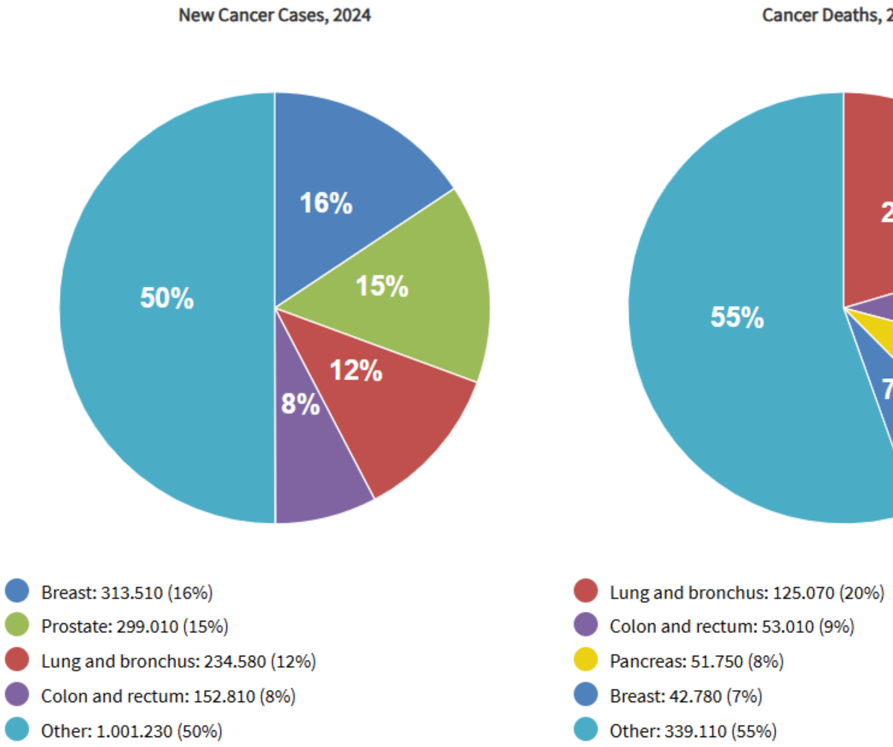


FIGURE 1.1: Statistics on new diagnosis and reported deceases for most common cancer types in 2024 for US citizens.

Risk prediction is accomplished by identifying characteristics that are associated with a high or low risk of developing a disease (traditional risk factors are listed in the following section), and then combining those characteristics in a statistical model to produce a probability estimate of developing the disease over a given 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 [3]. 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 [4].

There is increasing interest in using risk prediction models to 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 improving early detection of breast cancer [5].

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 [6] it was assessed each model's performance at the population level by comparing the number of women in their study who the model estimated [E] would develop breast cancer with the number of women who actually were 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 a matter of concern for researchers worldwide and new risk prediction *Machine Learning* (ML) algorithms 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.

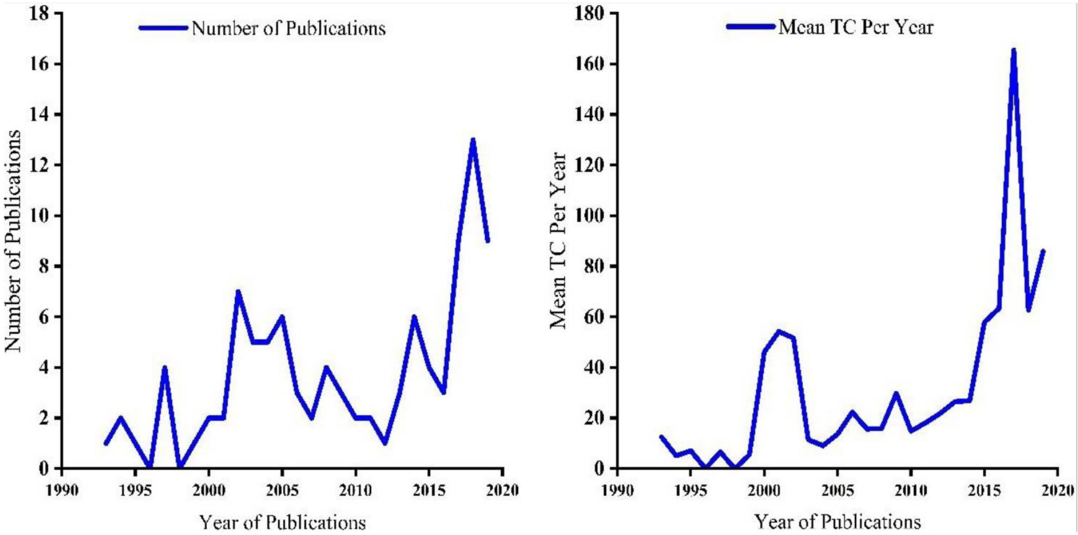


FIGURE 1.2: Annual growth of publications and mean of Total Citation Per Year (Mean TC Per Year) on AI and ML in cancer. As it was originally mentioned in [7], the results are for the top 100 cited articles indexed in Scopus database[2].

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

However, despite the success of the aforementioned methods, they still face significant limitations. For instance, incomplete risk profiling (excluding data, limiting their predicting power), inflexibility (static models, failing to adapt to a patient's changes over time) and population-specific biases (models calibrated for specific demographics, 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 [8].

Not only does human microbiome capture environmental and lifestyle influences, it also gives 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 predicting accuracy. For a more comprehensive and deeper explanation, one could find it in both [9] and [10].

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. The aim is for it to be as complete as possible, so we went all the way from data acquisition to model explainability, including both recurrence and mortality on it. Afterwards, the 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 to such datasets, as we mention later in Section 5.1.

1.2 Traditional risk factors

In order to get used to the features we used, in this section we list multiple breast cancer risk factors which have been identified over the years, as discussed in [11].

- **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 [12].
- **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 [13]. 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 [14]. The addition of breast density as a risk factor improves calibration and discrimination of various risk prediction models [15].
- **Radiation exposure:** Radiation exposure between the ages of 10 and 30 years (ie, in survivors of Hodgkin lymphoma) is a known risk factor [16].

- **Genetic factors:** Family history (in particular, an affected mother, sister, or male relative, early onset disease, and bilateral disease) is an established risk factor [11]. Inheritance of high-risk genetic mutations, such as BRCA1 and BRCA2, account for some but not all of this risk [17]. 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 [18].
- **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 [19]. 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) [20].
- **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 [11]. In postmenopausal obese women, the aromatase enzyme in adipose tissue converts androgens to estrogen, thus increasing breast cancer risk [21]. Premenopausal obese, however, have lower levels of serum estradiol [22]. Additionally, physical activity decreases breast cancer risk in a dose-dependent manner [23]. High levels of alcohol intake are associated with elevated breast cancer risk [11].

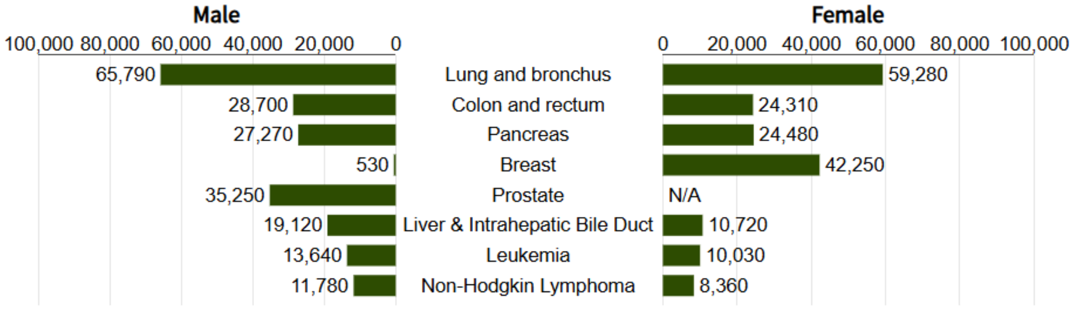
1.3 Integration of recurrence and mortality

In our context, incorporating recurrence and death 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 [24].

In one hand, recurrence refers 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 risk factors such as age, genetic predisposition or 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.



Source: Cancer Facts & Figures 2024, American Cancer Society (ACS), Atlanta, Georgia, 2024.

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 death 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 that can influence clinical decision-making, optimize treatment plans, and improve patient quality of life.

1.4 Aims of the research

As we have already mentioned in Section 1.1, our main goal is to define a pipeline for breast cancer using clinical data. Not only should it 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. In order to do this, we will require clinical data to get results good enough to support our findings. The main idea is to get predictions, extract the most important features for such predictions and see if the results are consistent by interpreting and explaining the aforementioned features. In the great scheme of things, the pipeline looks as Figure 1.4, though there are some details we omitted to enhance simplicity. Nonetheless, we are going to properly explain the whole procedure in Chapter 3.

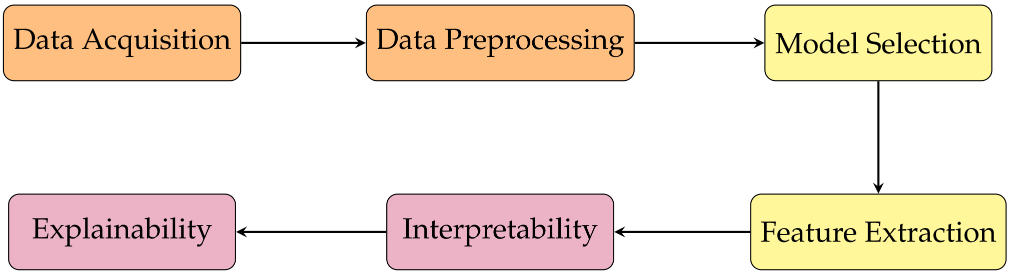


FIGURE 1.4: Pipeline for breast cancer risk prediction, which also can be used for recurrence and mortality. Blocks are partitioned by colors.

Additionally, we want the pipeline to be as robust as possible, so our intention is to test it with as many data and models as we can. For this reason, in our experiments we do not stick with a particular model, but rather select the best performing one for each case. Furthermore, several metrics have been defined in order to control the performance of the model from different perspectives; with these preventive measures we hope 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 a model which treats all protected classes equally[3]. 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.

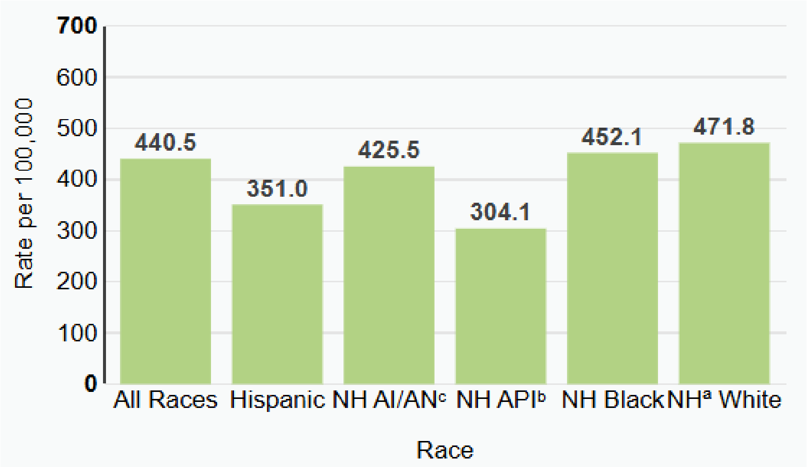


FIGURE 1.5: New cancer cases in US during 2017-2021 by race/ethnicity. [a]Non-Hispanic, [b]Asian/Pacific Islander, [c]American Indian/Alaska Native.

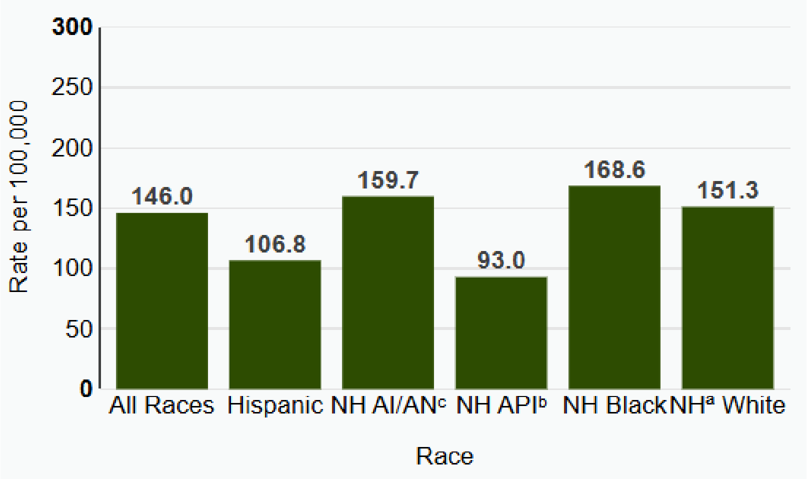


FIGURE 1.6: Reported deceases for cancer in US during 2018-2022 by race/ethnicity. [a]Non-Hispanic, [b]Asian/Pacific Islander, [c]American Indian/Alaska Native.

There are also groups which are discriminated due to the data available for us. 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 elegible as protected attributes.

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

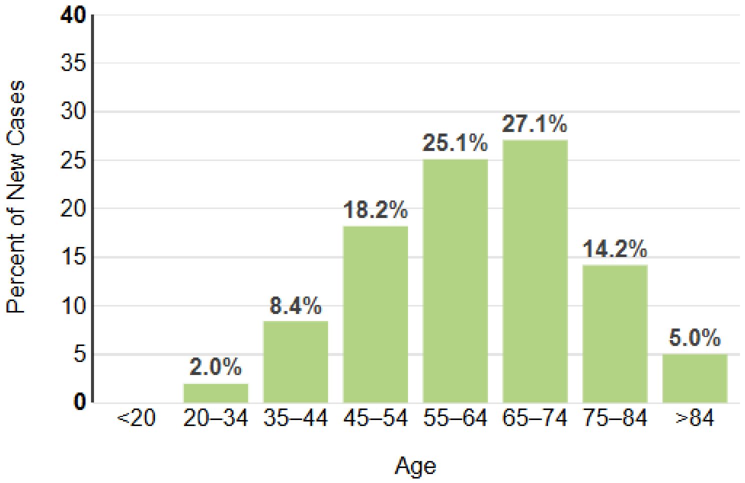


FIGURE 1.7: New confirmed breast cancer cases during 2017-2021.

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

2.1 Traditional approaches

The best-known model for predicting an individual woman's chance of being diagnosed with breast cancer is the Gail model (5 , 6) . This model includes the following risk factors: current age, race, age at menarche, age at first live birth, the number of firstdegree relatives with breast cancer, the number of previous breast biopsy examinations, and presence of atypical hyperplasia. The model predicts a woman's likelihood of having a breast cancer diagnosis within the next 5 years and within her lifetime (up to age 90 years). This and similar risk prediction models are readily available to clinicians and patients around the world through the Internet (7 , 8) . A version of the Gail model available on the National Cancer Institute's Web site (http://www.cancer.gov/bcrisktool/) is viewed 20,000 to 30,000 times each month (Rehmert JH: personal communication).

Regression Models Modified Gail Model/Breast Cancer Risk Assessment Tool The risk prediction model commonly known as the Gail model was developed in 1989 by Gail et al for women without prior breast cancer (42). The model has undergone several modifications over the years and is now available through the National Cancer Institute website as the BCRAT (43,44). The model is simple in design and can be used in various settings, including primary care. The first version, model 1, was composed of five questions focusing on reproductive and hormonal risk factors: age, age at menarche and at first live birth, number of previous breast biopsies, and number of first-degree female relatives with breast cancer (42). The risk factors and relative risks were based on case–control data from the Breast Cancer Detection Demonstration Project (BCDDP), a United States screening study conducted from 1973–1980, and the analysis used for the Gail model included data only from White women (42).

The second version, model 2, was developed in 1999 by the National Surgical Adjuvant Breast and Bowel Project investigators, in part, to determine eligibility criteria for the Breast Cancer Prevention Trial (43). This modified version uses incidence rates from the Surveillance, Epidemiology, and End Results (SEER) Program rather than data from the BCDDP and estimates only invasive cancer risk rather than invasive and in situ risk (43). Since then, the model, better known as the BCRAT, has undergone further modifications to include breast cancer incidence rates for African American and other non-White women and personal history of atypia as one of the risk factors (44). Other risk factors explored include breast density and weight (21).

The most recent version of the BCRAT provides estimates for five-year invasive cancer risk and lifetime invasive cancer risk for women who are at least 35 years of age

(44). In particular, the risk calculator is used to identify individuals with a five-year risk of at least 1.67

The Breast Cancer Surveillance Consortium (BCSC) model has a similar appearance to the BCRAT but also includes breast density. Earlier work on the Gail model using breast density as a continuous variable showed improved discrimination for invasive cancer risk in White women, and Tice et al demonstrated that the addition of Breast Imaging Reporting and Data System (BI-RADS)–based breast density improved calibration (21,45). The risk factors included in the BCSC model are similar to that of the Gail model/BCRAT and include age, race/ethnicity, family history of breast cancer in a first-degree female relative, and history of a breast biopsy with benign breast disease, in addition to BI-RADS breast density (46). The addition of other risk factors has been explored, such as polygenic risk scores and two sequential BI-RADS density measures (eg, breast density in 2007 and 2008) rather than one (47,78,79).

The BCSC model provides estimates for five- and 10-year invasive cancer risks for women who are at least 35 years of age without prior breast cancer, mastectomy, or breast augmentation (46). In the BCSC study population, the model slightly underestimated risk in younger women aged 40 to 44, Asian women, and Hispanic women (45). Subsequently, it was validated in cohorts from the Mayo Mammography Health Study and Metro Chicago Breast Cancer Registry (47,48). In the cohort of women in Chicago, 26

Rosner–Colditz Model The Rosner–Colditz model was based on the Pike model of breast tissue age, which was described in 1983 (51,80). The Pike model proposes that breast tissue age largely depends on estrogen and progesterone levels. Specifically, first full-term pregnancy at an early age is associated with reduced breast cancer risk, due to terminal differentiation of the mammary gland (which makes it less susceptible to carcinogens), while subsequent pregnancies are associated with transient increases in risk, due to the growth-enhancing effects of estrogens on premalignant cells (81). Following menopause, hormone levels depend on the peripheral conversion of androgens into estrogen by fat metabolism.

Using Nurses' Health Study data, Rosner and Colditz extended the Pike model by incorporating the following features into their risk prediction model: age at menarche, age at first birth and at each subsequent birth, and age at menopause (51). In 2000, the following risk factors were included: first-degree family history of breast cancer, benign breast disease, type of menopause, postmenopausal hormone use, body mass index (BMI), height, and alcohol consumption (82). The addition of those risk factors was shown to improve the model's AUC from 0.57 to 0.63 (52). Other risk factors identified in the Nurses' Health Study, such as breastfeeding, vegetable intake, physical activity, and breast density, were subsequently incorporated, which improved discriminatory statistics (83).

The Rosner–Colditz model predicts invasive cancer risk in women up to 70 years of age without prior breast cancer (67). It was validated based on California Teachers Study data, at which time it was revamped with newer data from the Nurses' Health Study, and performed similarly when applied to the California Teachers Study data (AUC of 0.59) and to the newer Nurses' Health Study data (AUC of 0.60) (53). It performed best in women aged 47 to 69 when estimating five-year risk. The Rosner–Colditz model highlighted the effects of modifiable lifestyle risk factors on breast cancer incidence and placed less importance on chronologic age; however, its clinical use is limited due to its modest discriminatory statistics and lack of availability through a website platform.

Genetic Risk Models Tyrer–Cuzick (IBIS) Model Developed in 2004, the Tyrer–Cuzick model, or IBIS model, is among the most well-known and widely used tools (56). It is based on data from the IBIS conducted in the UK and combines a genetic segregation model for familial risk and a regression model for other risk factors (40). The genetic segregation model assumes a two-locus genetic model, with one locus for BRCA1 or BRCA2 and the other locus for an unknown, low penetrance gene (67). Risk factors considered in the model include: age at menarche, age at first live birth, age at menopause, parity, height, BMI, atypical hyperplasia/lobular carcinoma in situ, hormone replacement therapy, benign breast disease, family history of breast and ovarian cancer in first- and second-degree relatives, and age at diagnoses (40). The latest additions are breast density and polygenic risk scores (22–25,40). A UK study demonstrated that polygenic risk scores based on a large number of single-nucleotide variants lead to improved risk stratification when combined with Tyrer–Cuzick risk and breast density (84).

The computer program for the Tyrer–Cuzick model displays a chart that shows a woman's breast cancer risk until 85 years of age, in addition to 10-year and lifetime risks (37). It also calculates the likelihood of having a BRCA mutation or a hypothetical autosomal dominant gene mutation assumed to have a low penetrance but a high frequency in the population (37,57). The risk estimates can be used to identify women who would benefit from chemoprevention and/or MRI screening (2,8). The Tyrer–Cuzick 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 (57). It demonstrates good calibration and discrimination when used in high-risk populations, and recent evidence suggests that models with multigenerational family history, such as Tyrer–Cuzick, better estimate risk even for women with below-average or average breast cancer risk (55,67).

Claus Model Developed in 1991, the Claus model is based on data from the Cancer and Steroid Hormone Study, which was conducted by the Centers for Disease Control (85,86). The study population was composed of 4730 White women aged 20 to 54 years with breast cancer and 4688 matched controls, registered between 1980 and 1982 at eight SEER centers (85,86). The original purpose of the model was to calculate familial breast cancer risk in women with a known family history of the disease. The model thus focused on family history of breast cancer (including age at diagnoses and including paternal history) and subsequently also incorporated family history of ovarian cancer (87). It does not include nongenetic risk factors.

The Claus model predicts risk of invasive cancer and ductal carcinoma in situ in women without a genetic mutation and can be used to identify women who would qualify for supplemental screening with MRI (2,37). Based on the assumption that breast cancer is transmitted as an autosomal dominant trait, the Claus model was the first model based on familial cancer history with a single dominant hereditary genetic mutation as a cause (85–87). The results of the authors' studies laid a foundation for the existence of a heritable, germline mutation as a cause of breast cancer in women with family history, as their studies predated the identification of the BRCA1 and BRCA2 mutations. Following the discovery of the BRCA genes and their link to ovarian cancer, the model was revised and has now been integrated into a pedigree drawing software (Cyrillic) (88). It calculates the likelihood of carrying a genetic mutation and the cumulative risk of developing breast cancer. However, discrepancies exist between published tables and the extended Claus model available through the software package, possibly because the tables make no adjustments for unaffected relatives (10). Overall, the Claus model does not perform as well as other genetic risk prediction models (58,59).

BRCAPRO Model Developed in 1997, the BRCAPRO model is a genetic risk prediction model that uses Bayes' theorem to estimate the likelihood of carrying a BRCA mutation in patients with a family history of breast and/or ovarian cancer (60,61). Whereas the Claus model predated the discovery of the BRCA mutations and assumed a genetic cause for familial breast cancer, the BRCAPRO model estimates the likelihood of carrying a BRCA mutation based on a Mendelian inheritance pattern and relies on published data about gene penetrance and prevalence (60). Initially, the original model assumed two-allele and autosomal dominant inheritance of BRCA1 genes, with the expectation that the BRCA2 gene would be incorporated later with further evidence, and regarded other genes as sporadic. As such, consideration for BRCA2 was later added and is included in the current version (89,90). Since the first version, other factors such as race/ethnicity and tumor markers (eg, estrogen receptor and progesterone receptor status) have been added, which improved its performance, and the current model consistently shows good discrimination between carriers and noncarriers (55,58,62,63,90).

The BRCAPRO model can be used to determine whether a patient would benefit from genetic testing (38). In addition, lifetime risk above 20

Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm The BOADICEA is a genetic risk model that is based on breast cancer genetic susceptibility being attributed to the effects of the BRCA1 and BRCA2 mutations and the assumption that residual clustering within families is due to the multiplicative effects of many genes (polygenic component) (64,65). The specific pathogenic mutations it now incorporates are BRCA1, BRCA2, PALB2, CHEK2, and ATM (94). The BOADICEA was developed in 2002 with data from the Anglian Breast Cancer Study (which was later renamed SEARCH and included women with breast cancer diagnosed before the age of 55 who were registered in the East Anglian Cancer Registry) and from multiple case families in the UK (which included families with two or more breast cancer cases, one of which was diagnosed before the age of 50) (64,65). It was then updated using data from the UK National Case Control Study, the Manchester Study, and pooled pedigree data from 22 studies (65,95–97). Similar to the Tyrer–Cuzick model, the BOADICEA incorporates nongenetic risk factors; more recently, tumor pathology and breast density were added (66,98).

The BOADICEA predicts the probability of carrying a BRCA1 or BRCA2 mutation (or the proposed polygenic component) and also predicts breast cancer and ovarian cancer risk (65). The model is available online, through which family history can be entered beyond second-degree relatives (66). Unlike other genetic risk models, family history is not limited to particular relatives or degrees (67). Compared to other models, the BOADICEA performs well with good calibration and discrimination (58,67).

Myriad Model The Myriad model is based on gene sequencing analyses performed by Myriad Genetic Laboratories (Salt Lake City, UT). Created in 1997, Myriad I (or the Shattuck-Eidens model) is an empirical model that was developed with 798 unrelated individuals from the United States and Europe thought to be at high risk of a BRCA1 mutation (68). Myriad I estimates the risk of harboring a BRCA1 mutation based on the following risk factors: personal history of unilateral/bilateral breast cancer or ovarian cancer, patient age at first diagnosis of cancer, Ashkenazi Jewish descent, and number of relatives with breast or ovarian cancer (68).

The current empirical model, Myriad II by Frank et al, was developed in 1998 based on women with breast cancer diagnosed before the age of 50 or ovarian cancer diagnosed at any age and at least one first- or second-degree relative with either breast or ovarian cancer (69). The model was then refined and tested in 2002 with the results of 10 000 gene sequence analyses, which were performed to identify deleterious BRCA1 or BRCA2 mutations and three specific Ashkenazi Jewish founder mutations (69,70). The current model is based on the following risk factors: personal history of breast cancer, Ashkenazi Jewish descent, and family history of a first- or second-degree relative with breast cancer diagnosed before the age of 50 or ovarian cancer diagnosed at any age (71). The mutation prevalence tables are separated according to Ashkenazi Jewish ancestry, as this population has higher rates of BRCA1 and BRCA2 mutations attributed to three founder mutations (70). Myriad II is used in high-risk women based on family history but is only able to include risk from up to two relatives and attributes the same level of risk to all breast cancers diagnosed before age 50 (eg, breast cancer diagnosed in the 20s is treated similarly to cancer diagnosed in the 40s) (67).

2.2 Including microbiome data

Chapter 3

Materials and methods

The most difficult part of Data Science is arguably obtaining qualitative and quantitative data. This is also the case for the Health domain, where you have to take further considerations in order to access data. We will discuss in Chapter 5 the particular inconveniences we had with it, while in this one we are going to show which data were used for our experiments and the methodology we followed.

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 best our interests.

Unfortunately, the public datasets we found did not have control subjects (that is, patients which did not contract breast cancer). Thus, we decided to use the one in [25], 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 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 death risk prediction) it was easier, we just reviewed all the datasets we searched and selected the one with more subjects and qualitative variables. This dataset is included in the *Cancer Imaging Archive* collection[1]. Because we focused on prediction from clinical data (no use of medical imaging), we only used the `.xlsx` file with title *Clinical and Other Features*.

3.2 Data summary

We will refer to the datasets used in [25] and from *Cancer Imaging Archive* as Lifestyle and Duke datasets, respectively. As a reference when looking at our experiments, we included a feature dictionary for both datasets in Appendix A.

3.2.1 Lifestyle dataset

As explained in the referenced article, 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 we do not need to preprocess 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 ended 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 MRI and US tests.

Since we had a total of 98 features, we removed all data from imaging tests alongside unnecessary variables 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 started our experiments using the 69 variables that were 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 we discussed after obtaining the data was which models we should use. Even though neural networks would most likely obtain the best results, we decided to discard 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 took some of the most popular machine learning classifiers and see which of them worked best in our scenarios.

These models were not selected arbitrary, we particularly wanted to include *Logistic Regression* and *K-Nearest Neighbors* ($K = 5$ by default), but also probabilistic and tree models were used. After discussing it again and again, we ended up using the aforementioned ones, in addition to *Decision Tree*, *Random Forest*, *XGBoost* and *SVM*. With the exception of *K-Nearest Neighbors*, all these models accept a `random_state` parameter (initialized at 42), which will allow us to reproduce the results we obtain.

3.3.2 Data preprocessing

We already mentioned in the previous section Lifestyle dataset is already clean, so we only have to focus on the Duke dataset. Some of the models we use (*Logistic Regression*, 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.

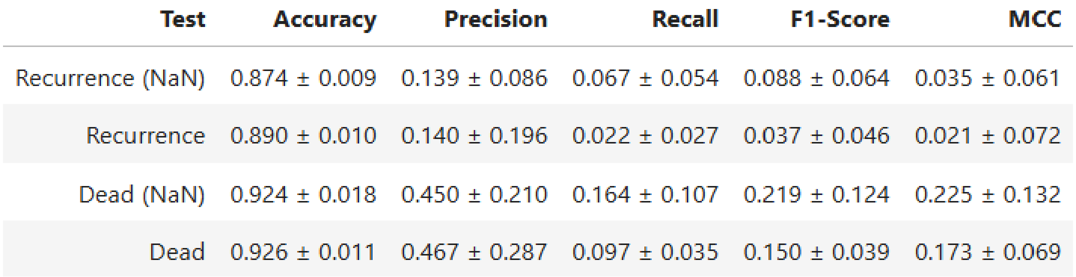
After changing the feature `Age` from days to years, we use the mean value of each feature to fill missing data. Although this is a typical approach, we will see it indeed improves the performance of models with the capacity of dealing with `NaN`, such as *XGBoost*. However, we only filled the columns where it made 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 used a label encoder to transform categorical string data into numeric types.

Due to the nature of our data, we only had columns specifying the days until our target features. Thus, we had to manually add columns which will serve us as labels for our classifications. Moreover, we combined both `Days_to_local_recurrence` and `Days_to_distant_recurrence` into a single column `Days_to_recurrence` because there was not enough data to treat each variable separately. Finally, we filled 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 prefered 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 selected 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 a 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 diseases like breast cancer.
- **F1-Score:** We define it 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 also is 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 *XGBoost*.

Since we both wanted to check these metrics and whether our preprocessing was 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 *XGBoost*. 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 were successful, we decided it would be best to first compute metrics for a simple model with data before preprocessing. In the case of Duke dataset specifically, this means we retain features with more than 600 `NaN`.

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

3.3.5 Experiments

As we have already mentioned, our goal is to assess both risk prediction for breast cancer and, additionally, to its recurrence and mortality. For the latter ones, we also tried to predict the time before the events occur, though it obviously is a much more difficult problem and for its performance to be acceptable it requires a quality and quantity of data we did 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. All hand-crafted methods were 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 wanted to perform a 80-20 split, we took 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, since we are showing the mean and standard deviation from all 5 computations, which we thought is a natural way to express it.

Additionally, we scaled 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 [26]. Thus, we tested this technique to hopefully 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 was to split data in privileged and unprivileged classes. In the case of `Mol_Subtype` and `Race_and_Ethnicity`, we counted each of the values and noticed there was a predominant value, which we considered 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 were 0 for the former (associated to `luminal-like`) and 1 for the latter (label for `white`).

We wanted to follow a similar approach for `Age` even though it is not a categorical feature. Hence, by defining a threshold we created a new feature (denoted as `Age_Group`) splitting data into two classes. Based on the literature, we decided to 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 tradeoff 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 SHAP (SHapley Additive exPlanations) 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 tried to predict the time before these events occurred using both the target (`Recurrence`, `Dead`) and duration (`Days_to_recurrence`, `Days_to_death`) features using a Cox's proportional hazard model[3]. Although the results were not promising (see Chapter 4), we considered it an essential part of the pipeline and decided to include it regardless.

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

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

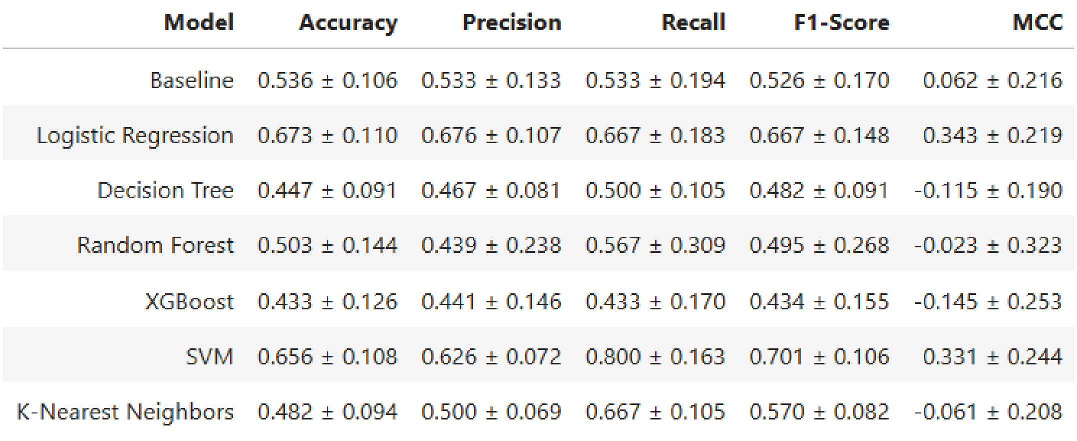
Chapter 4

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 switched their places instead. We encourage you to check out the *GitHub* repository if you face any trouble.

4.1 Breast cancer risk prediction

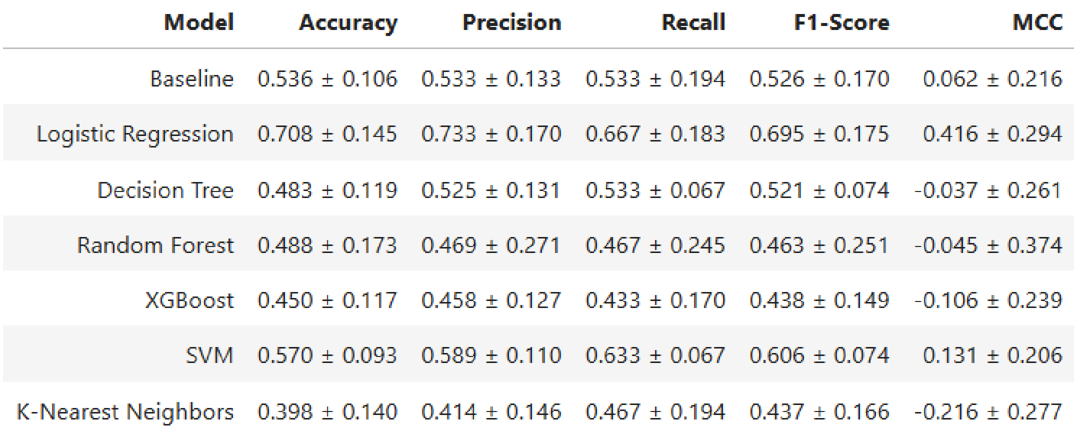
Following the process detailed in Chapter 3, we first compare all models using the metrics we defined before. In Figure 4.1 you can see the performance for the models without balancing the dataset through *SMOTE*, while in Figure 4.2 we show the same visualization using this strategy. One can notice 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 *Logistic Regression*. As we can see in both figures, an increase in precision (which means higher rate while predicting positives) led to a better overall performance. Thus, 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).



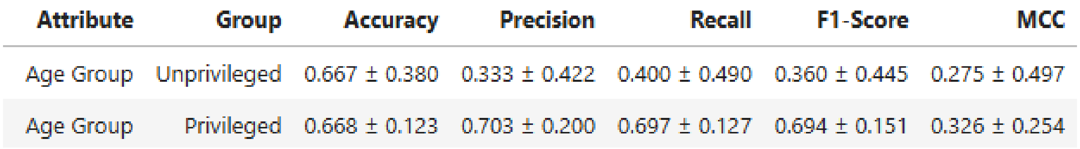
Model	Accuracy	Precision	Recall	F1-Score	MCC
Baseline	0.536 ± 0.106	0.533 ± 0.133	0.533 ± 0.194	0.526 ± 0.170	0.062 ± 0.216
Logistic Regression	0.708 ± 0.145	0.733 ± 0.170	0.667 ± 0.183	0.695 ± 0.175	0.416 ± 0.294
Decision Tree	0.483 ± 0.119	0.525 ± 0.131	0.533 ± 0.067	0.521 ± 0.074	-0.037 ± 0.261
Random 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 *Random Forest* and *SVM* some metrics improve while other diminish, but any of these changes is of big magnitude. Nonetheless, *K-Nearest Neighbors* experiences a decreasing in all metrics, reflecting the fact that *SMOTE* does indeed not work for all models in this particular scenario.

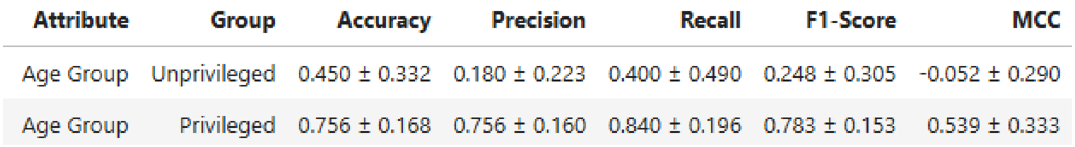
Taking a deeper look to Figure 4.1, the best performing models in this first phase are without a doubt *Logistic Regression* and *SVM*, which move on to the fairness test. Remember that for the Lifestyle dataset we only considered `Age_Group` as a protected attribute, so the comparison will be much easier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

For *Logistic Regression*, 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 *Logistic Regression* 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 *Logistic Regression*. 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 *Logistic Regression* 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 *Logistic Regression*, 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 [27] 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.

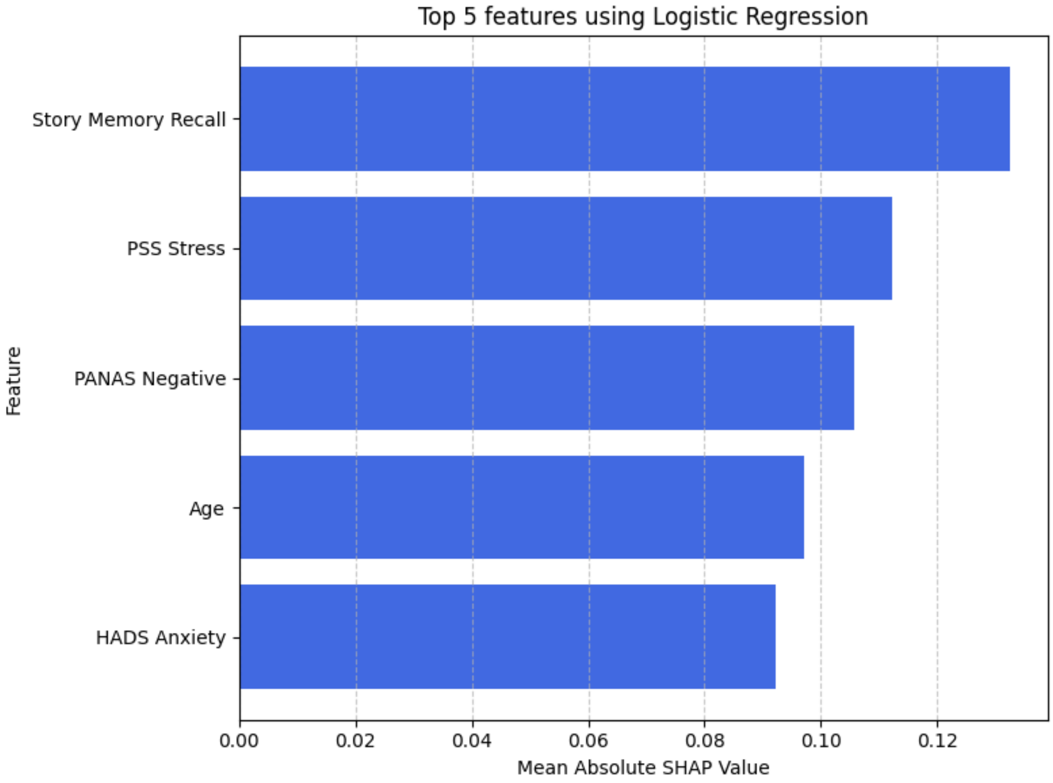


FIGURE 4.5: Plot of the 5 most relevant features for breast cancer risk prediction when using *Logistic Regression* (using SHAP values).

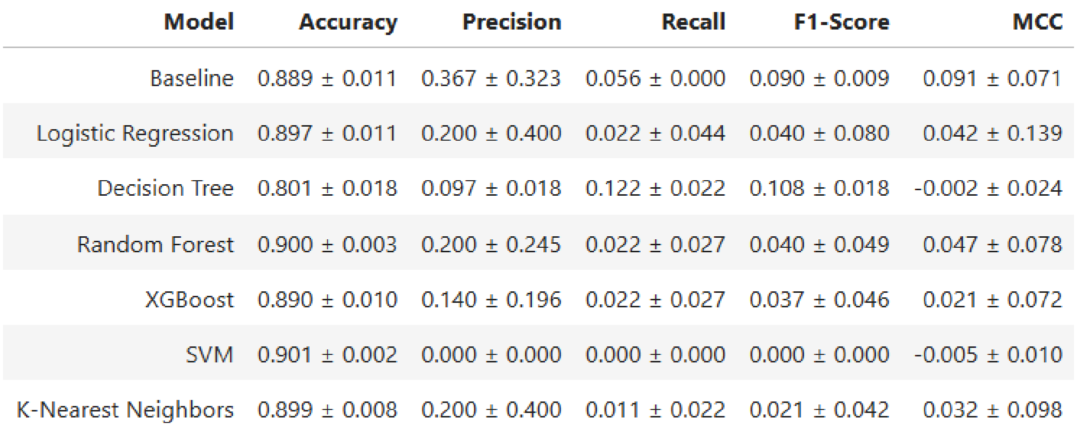
On the other hand, the rest of the features (expect `Age`) reflect psychosocial stressors, which are intricately linked to physical health. In particular, chronic stress and negative affect are thought to exacerbate cancer risk through immune suppression, chronic inflammation and unhealthy lifestyle 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.

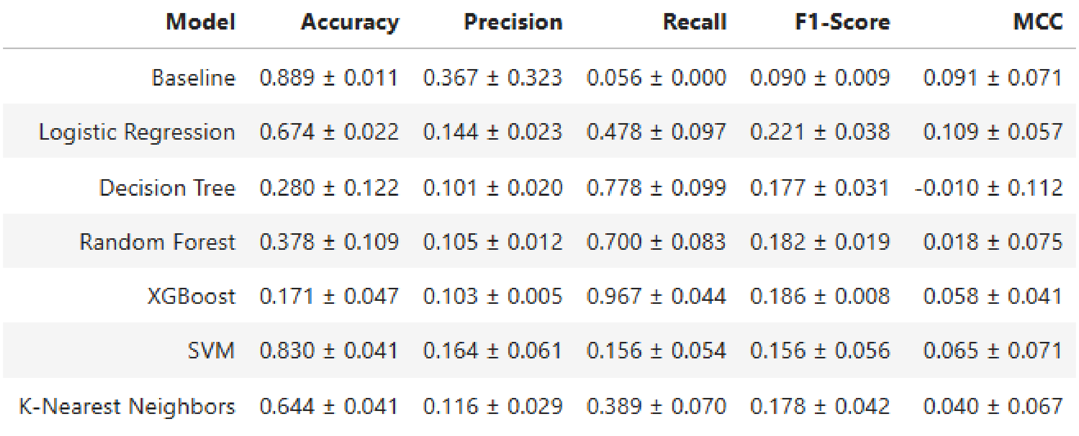
4.2.1 Target 1. Recurrence

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

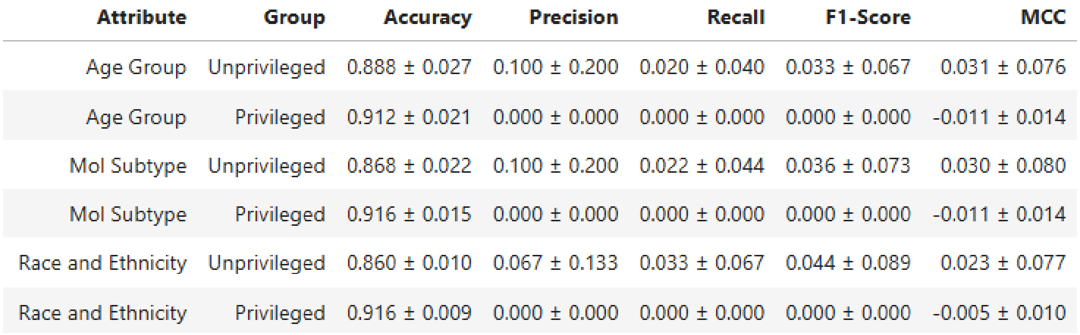
FIGURE 4.7: Comparison between models while predicting the risk of recurrence (with *SMOTE*).

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 *Decision Tree* 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 *Decision Tree* is superior, so we also discard it.

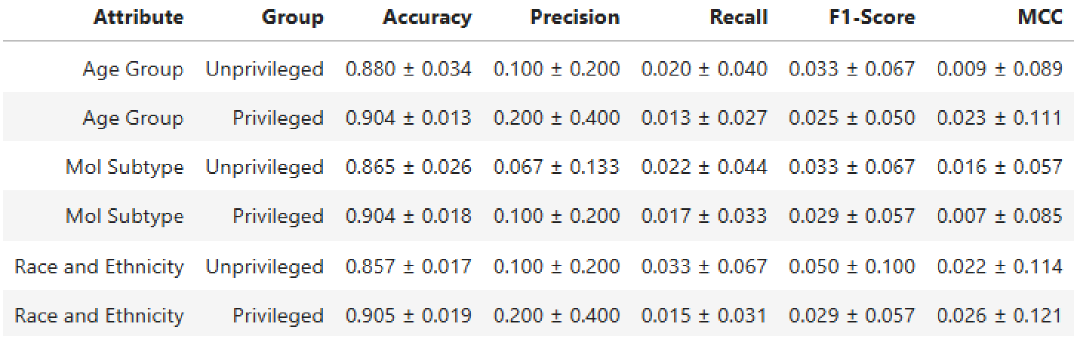
Due to the fact that for Duke dataset we consider three protected attributes, we should discard at least one more model. Since the rest of them are better than the others in some metric, we ruled *Logistic Regression* out. The reason being it gets almost the same values as *Random Forest*, but in all 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 *Random Forest*, *XGBoost* and *K-Nearest Neighbors*, respectively.



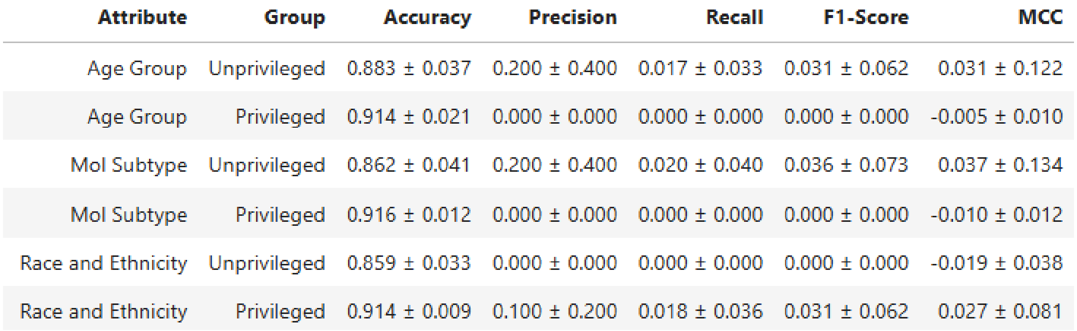
Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.888 ± 0.027	0.100 ± 0.200	0.020 ± 0.040	0.033 ± 0.067	0.031 ± 0.076
Age Group	Privileged	0.912 ± 0.021	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	-0.011 ± 0.014
Mol Subtype	Unprivileged	0.868 ± 0.022	0.100 ± 0.200	0.022 ± 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 *Random Forest* 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 *XGBoost* 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 *K-Nearest Neighbors* 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 *Random Forest* and *K-Nearest Neighbors*, one of the classes gets horrible metrics. One could argue 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 *XGBoost*, so we discard both *Random Forest* and *K-Nearest Neighbors*.

We now proceed to the interpretability step, where we show the most important features for our predictions. In Figure 4.11 it is represented the top 5 variables while using *XGBoost*. 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.

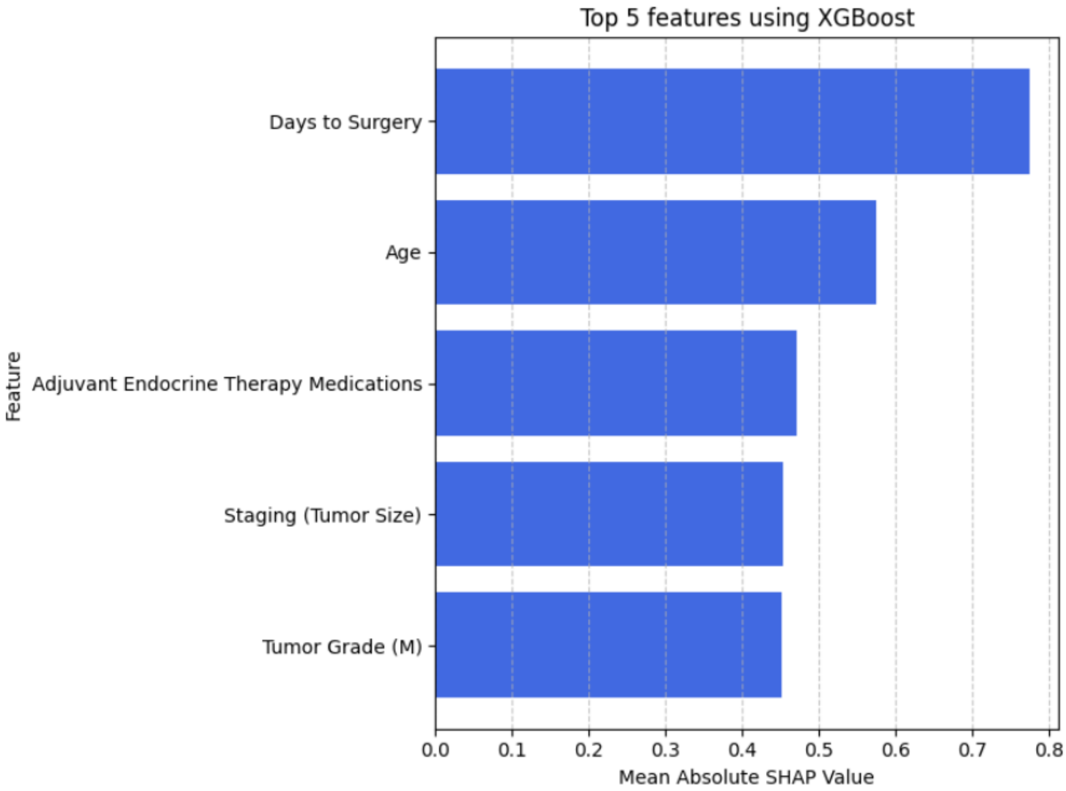


FIGURE 4.11: Plot of the 5 most relevant features for recurrence prediction when using *XGBoost* (using SHAP values).

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 to, 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 we 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.

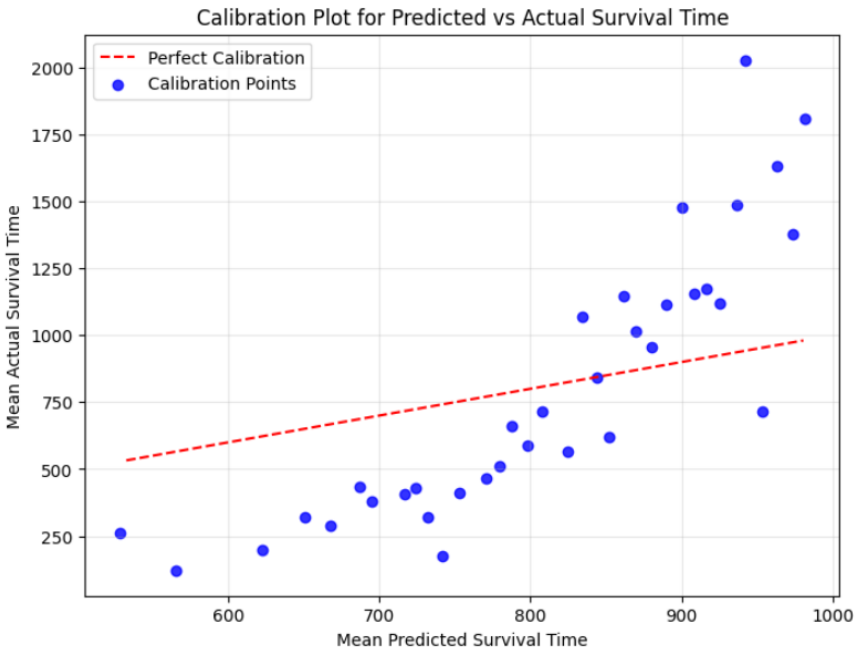
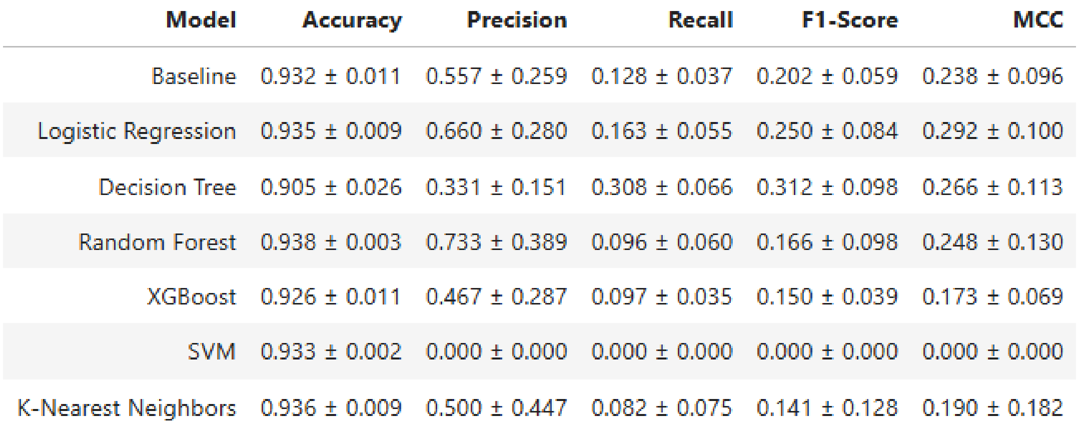


FIGURE 4.12: Plot of time estimation before `Recurrence` occurs.

We already foreshadowed in Chapter 3 the performance will not be great, the reason behind it being these types of models require more quantity of data that 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.

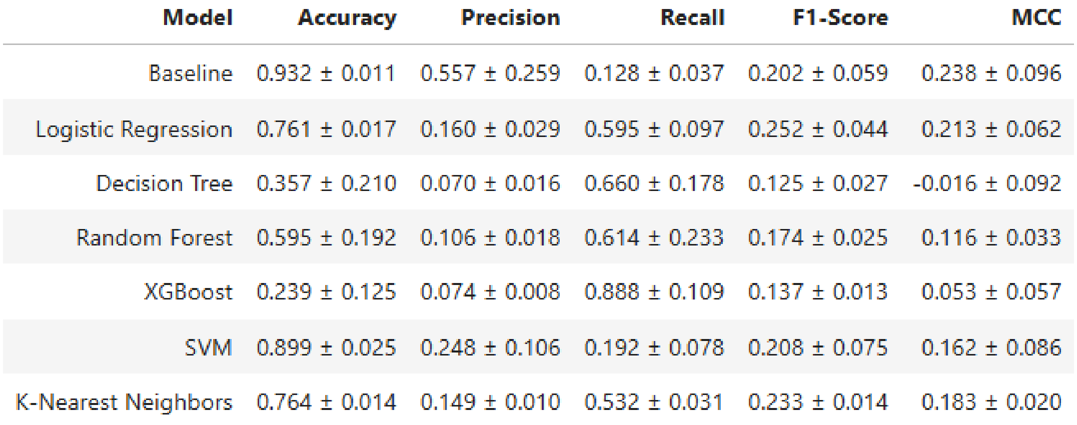
4.2.2 Target 2. Mortality

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

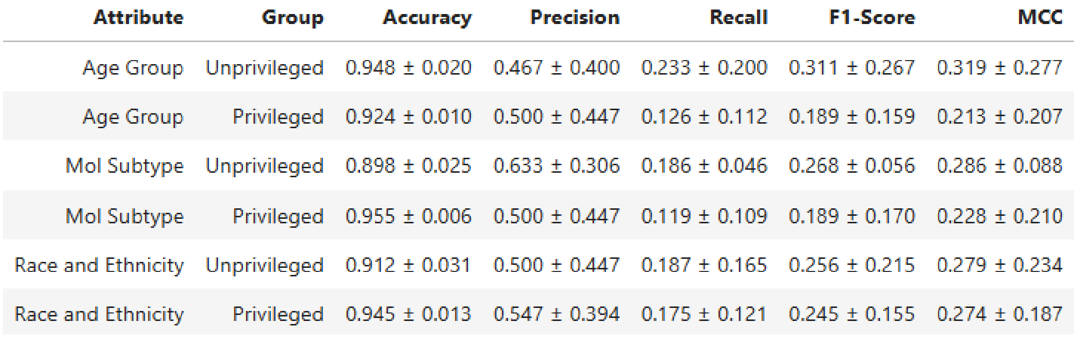
FIGURE 4.14: Comparison between models while predicting the risk of death (with *SMOTE*).

Here we observe a similar behaviour as 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 *K-Nearest Neighbors* almost ties in accuracy with *Logistic Regression*, it is inferior to the latter for all other metrics, so we also rule it out.

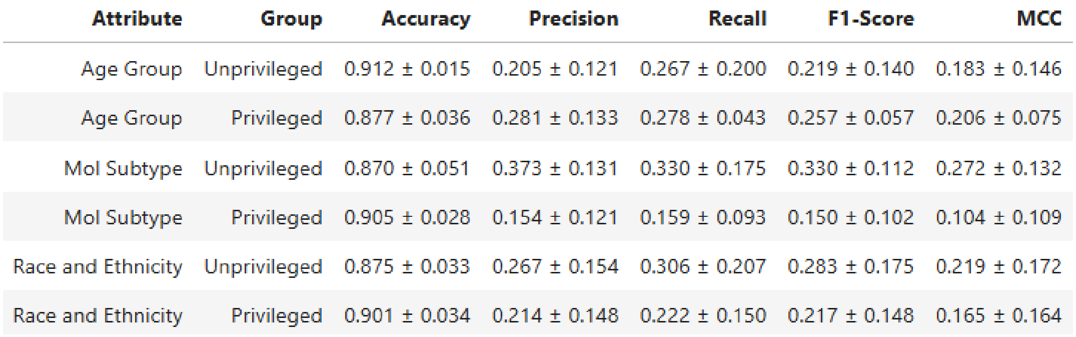
Following the same logic as for recurrence, we should discard at least another model. Notice that *Logistic Regression* and *Random Forest* lead both accuracy and precision, while *Decision Tree* is on top for recall and f1-score. Thus, we remove *XGBoost* from our selection.

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 *Logistic Regression*, *Decision Tree* and *Random Forest*, 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.



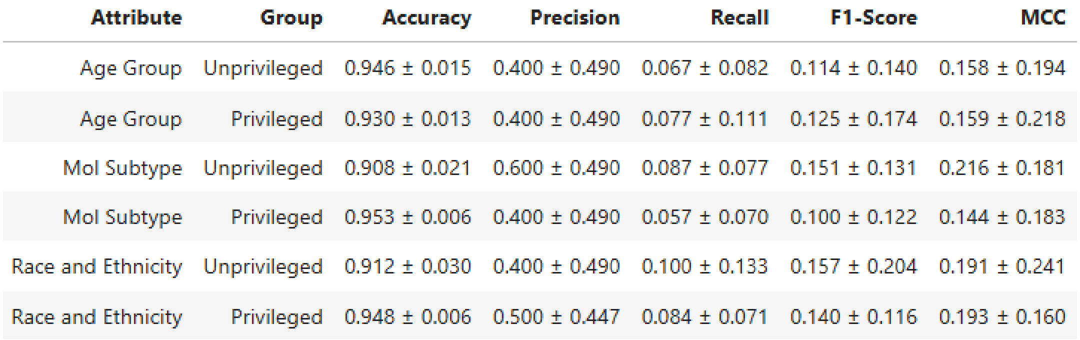
Attribute	Group	Accuracy	Precision	Recall	F1-Score	MCC
Age Group	Unprivileged	0.948 ± 0.020	0.467 ± 0.400	0.233 ± 0.200	0.311 ± 0.267	0.319 ± 0.277
Age Group	Privileged	0.924 ± 0.010	0.500 ± 0.447	0.126 ± 0.112	0.189 ± 0.159	0.213 ± 0.207
Mol Subtype	Unprivileged	0.898 ± 0.025	0.633 ± 0.306	0.186 ± 0.046	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 *Logistic Regression* 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 *Decision Tree* 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 *Random Forest* 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 will rely on particular observations to select the best. For example, *Decision Tree* 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, *Logistic Regression* slightly outperforms *Random Forest*, while the latter beats the former for Age_Group. Finally, in Mol_Subtype we can see bigger difference in values, with *Logistic Regression* getting better metrics.

In spite of the fact that *Random Forest* 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. 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 *Random Forest* as our model, we now proceed to interpretate its predictions. Following the same procedure as in both previous experiments, we show in Figure 4.18 the top 5 relevant features while predicting mortality and, then, try to justify their appearance. Notice that, as we were expecting, Age happens to also be an important fatality factor.

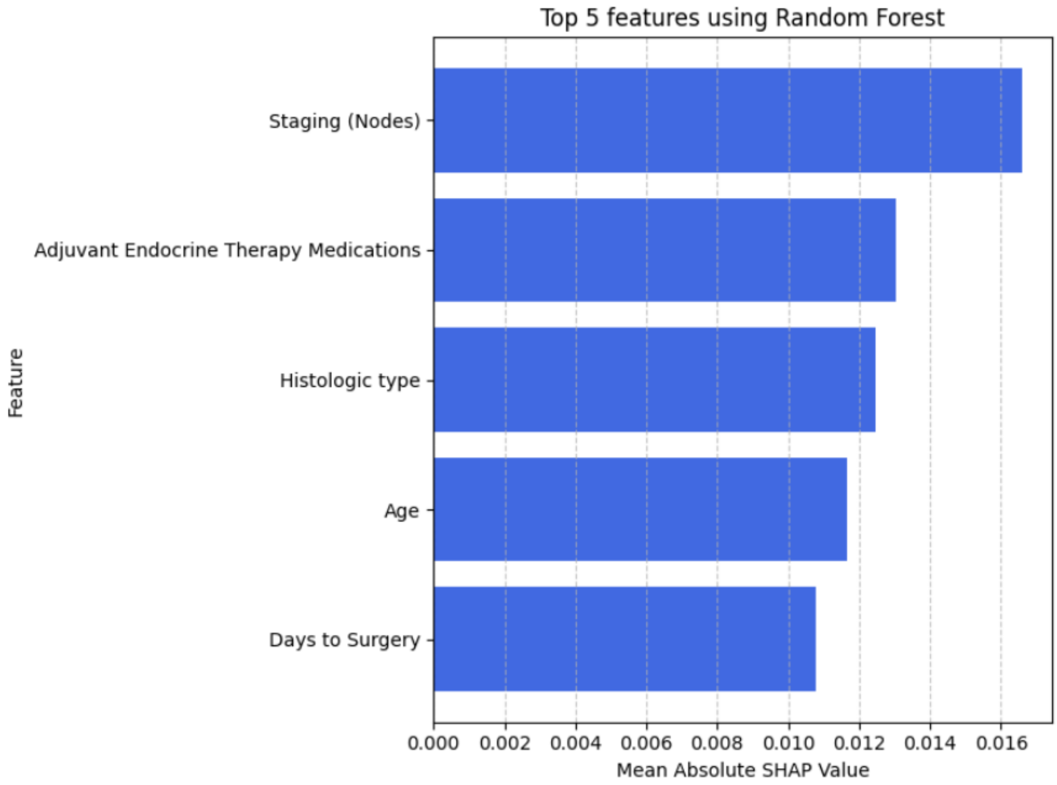


FIGURE 4.18: Plot of the 5 most relevant features for mortality prediction when using *Random Forest* (using SHAP values).

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.

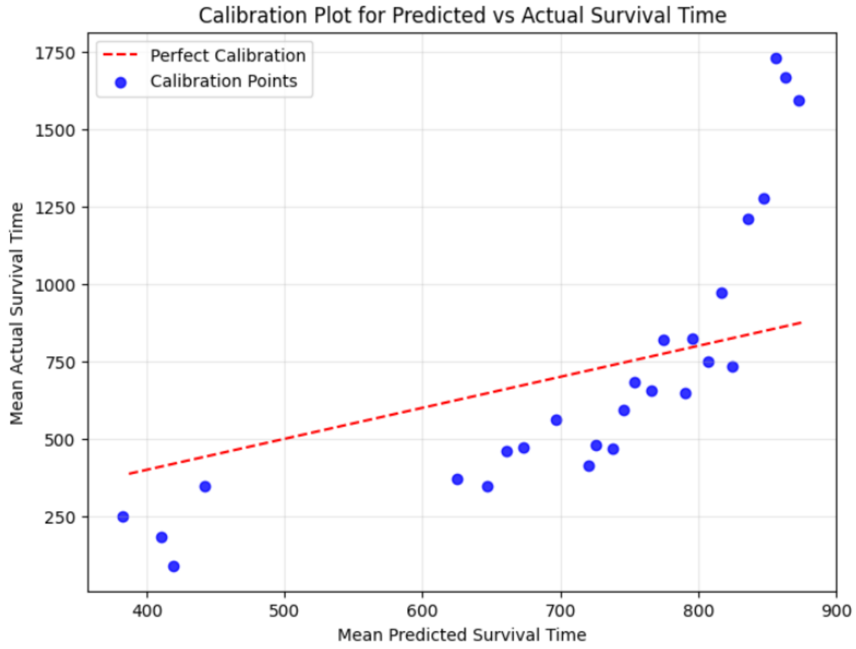


FIGURE 4.19: Plot of time estimation before the event `Dead` occurs.

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 wanted to test the robustness of our pipeline for these specific cases. This is also one of the reasons why we selected *Random Forest* for mortality prediction, since we wanted to show the performance of the whole process with as many models as possible.

Summing up our findings, *Logistic Regression* was the top performer for breast cancer risk prediction, *XGBoost* for recurrence prediction and *Random Forest* 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 aspects which could be improved, which we will include in 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 often delayed for one to two weeks. 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. However, by this time we were almost at the start of Christmas holidays, and the few people responding us also 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 issues, 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 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 seventy thousand subjects and more than a hundred of features. Hence, we did not have enough time to understand and process 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 [28]), 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 before in this thesis, for us it is crucial readers comprehend what we are referring to while observing 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

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] Unknown. “Globocan 2012 - Home”. In: *International Agency for Research on Cancer (IARC)* (2012). URL: http://globocan.iarc.fr/Default.aspx.

[3] Isaacs C. Peshkin B. “Risk assessment in women with an inherited predisposition to breast cancer”. In: *UpToDate* (2023). URL: http://uptodateonline.com/utd/content/topic.dotopicKey=breastcn/7619.

[4] Gail MH Hartge P Colditz GA Ballard-Barbash R et al Freedman AN Seminara D. “Cancer risk prediction models: a workshop on development, evaluation, and application”. In: *J Natl Cancer Inst* (2005).

[5] Penhoet EE Petitti DB Joy JE. “Saving women’s lives: strategies for improving breast cancer detection and diagnosis”. In: *The National Academies Press* (2004).

[6] Masala G Specchia C Palli D Gail MH Decarli A Calza S. “Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort”. In: *J Natl Cancer Inst* (2006).

[7] Zamit I Musa TH Musa HH Tassang A Akintunde TY Li W Musa IH Afolabi LO. “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).

[8] Xochitl C. Morgan and Curtis Huttenhower. “Chapter 12: Human Microbiome Analysis”. In: *PLOS Computational Biology* (2012). URL: https://doi.org/10.1371/journal.pcbi.1002808.

[9] Anna H Wu et al. “Gut microbiome associations with breast cancer risk factors and tumor characteristics: a pilot study”. In: *Breast cancer research and treatment* (2020).

[10] Loobuyck A Vandenbulcke Z Vogtmann E Pisanu S Iguacel I Scalbert A Indave I Smelov V Gunter MJ Michels N Huybrechts I Zouiouich S. “The Human Microbiome in Relation to Cancer Risk: A Systematic Review of Epidemiologic Studies”. In: *Cancer Epidemiol Biomarkers Prev* (2020).

[11] Godellas CV Ban KA. “Epidemiology of breast cancer”. In: *Surg Oncol Clin N Am* (2014).

[12] Cantor A et al Nelson HD Zakher B. “Risk factors for breast cancer for women aged 40 to 49 years: a systematic review and meta-analysis”. In: *Ann Intern Med* (2012).

[13] Bovbjerg VE Harvey JA. “Quantitative assessment of mammographic breast density: relationship with breast cancer risk”. In: *Radiology* (2004).

[14] Miglioretti DL Sprague BL Kerlikowske K; Breast Cancer Surveillance Consortium Engmann NJ Golmakani MK. “Population-attributable risk proportion of clinical risk factors for breast cancer”. In: *JAMA Oncol* (2017).

[15] Buist DSM Bowles EJA Brentnall AR Cuzick J. “Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density”. In: *JAMA Oncol* (2018).

[16] Dores GM et al Travis LB Hill D. “Cumulative absolute breast cancer risk for young women treated for Hodgkin lymphoma”. In: *J Natl Cancer Inst* (2005).

[17] Rahman N Stratton MR. “The emerging landscape of breast cancer susceptibility”. In: *Nat Genet* (2008).

[18] Wagner S et al Gallagher S Hughes E. “Association of a polygenic risk score with breast cancer among women carriers of high- and moderate-risk breast cancer genes”. In: *JAMA Netw Open* (2020).

[19] Page DL Dupont WD. “Risk factors for breast cancer in women with proliferative breast disease”. In: *The New England Journal of Medicine* (1985).

[20] Frost MH et al Hartmann LC Sellers TA. “Benign breast disease and the risk of breast cancer”. In: *The New England Journal of Medicine* (2005).

[21] Zeleniuch-Jacquotte A et al Toniolo PG Levitz M. “A prospective study of endogenous estrogens and breast cancer in postmenopausal women”. In: *J Natl Cancer Inst* (1995).

[22] Siiteri P Hoover RN Potischman N Swanson CA. “Reversal of relation between body mass and endogenous estrogen concentrations with menopausal status”. In: *J Natl Cancer Inst* (1996).

[23] Katzmarzyk PT et al.; 2018 Physical Activity Guidelines Advisory Committee McTiernan A Friedenreich CM. “Physical activity in cancer prevention and survival: a systematic review”. In: *Med Sci Sports Exerc* (2019).

[24] Celik O Akcay M Etiz D. “Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy”. In: *Adv Radiat Oncol* (2020).

[25] Nathan C. Wetter Gillian E. Cooke 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://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149552`.

[26] L. O.Hall W. P. Kegelmeyer N. V. Chawla K. W. Bowyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* (2002).

[27] Sliwinski MJ Jim HSL Small BJ Scott SB Mogle JA. "Memory lapses in daily life among breast cancer survivors and women without cancer history". In: *Psychooncology* (2020).

[28] Dilip K. Prasad Ayush Somani Alexander Horsch. "Interpretability in Deep Learning". In: *Springer Cham* (2023).