POLITECNICO DI TORINO

Masters's Degree in ICT for Smart Societies

Masters's Degree Thesis

Machine learning analysis for bowel urgency in ulcerative colitis patients from IBD Tool web application

Supervisors

Prof. Carla Fabiana CHIASSERINI

Prof. Guido PAGANA

Candidate

Brayan MONTOYA RODRIGUEZ

July 2024

Summary

This thesis is based on the analysis of bowel urgency, which is a common and very problematic symptom associated with inflammatory bowel diseases, it is characterized by a sudden need to have a bowel movement, which manifests in different severities like a mild hurry and total incontinence. Bowel urgency represents one of the most undesirable symptoms of inflammatory bowel diseases and it is arguably the worst in terms of affecting the quality of life and the normal development of daily activities.

The IBD Tool is a telemedicine web application platform developed between LINKS and the Mauriziano Hospital in Turin, the objective is to aid in the monitoring of the patients diagnosed with any inflammatory bowel disease at a distance, the core functionality is to deliver a set of questionnaires which are designed by the international medical community in order to gather information about the impact and symptoms of the inflammatory bowel diseases.

The work of this thesis is focused on ulcerative patients, which is one of the most common types of inflammatory bowel diseases altogether with Crohn's disease. Ulcerative colitis is characterized by inflammation in only the colon and rectum, with inflammation limited to the innermost lining of the intestinal wall.

The objective of the thesis is the implementation of prediction models to estimate the degree of bowel urgency. The bowel urgency degree is obtained from the P-SCCAI questionnaire, which tracks the disease activity index; there is a set of questions that addresses this issue in a simple way for the patients, asking about how much of a hurry they feel whenever they need to go to the toilet. The IBD Tool project stores all the information in a centralized database using MongoDB, a document-oriented database management system, among the multiple collections of data stored, there are two particularly important for the work of this thesis, one storing all the information of the filled questionnaires, e.g. the scores, the answers to each question and the timestamp, and, on the other hand, the clinical data collection, which stores information associated with the pathology and medical record of the patients, such as disease extent, weight, height, pathology duration, therapies ongoing, particularly IBD Tool stores therapies of the pharmaceutical type.

Algorithms to be implemented are classifiers, particularly speaking, the chosen methods for the work are the logistic regression and the random forest. They are implemented following different strategies explained in the Results section to cope with issues such as class imbalance and similarities between classes. After all, the questionnaires are entirely filled out by patients, which gives room to improper responses that may dist from a medical point-of-view reality and with room to subjective thoughts.

During this thesis, strong relationships between questionnaires and medical therapies are found through the statistical methods employed. In particular, the implementation of the prediction models, with the objective of correctly classify the bowel urgency degree. Overall, random forest models demonstrated better performance in comparison with the logistic regression ones, reaching values of 93 % of global accuracy. The importance of this work lies on the fact that the prediction of bowel urgency may be an interesting and useful feature for future implementations in IBD Tool but also in any telemedicine solution that aims to deliver care to patients diagnosed with any inflammatory bowel disease, especially ulcerative colitis; which gives the possibility to help in the prevention of relapses of the patients and help medical staff to directionate treatment from different approaches.

The telemedicine is an ever-growing field and beneficial for all parties involved, it helps the medical infrastructure by avoiding unnecessary medical visits and also facilitates the communication between patients and physicians. It represents a bright future as a complement to traditional health services, however, it requires a strong encouragement to the patients to trust and use these tools.

Acknowledgements

I would like to express my gratitude to all the people who have supported me during my academic journey, especially my parents.

Many thanks to Prof. Carla Fabiana Chiasserini for allowing me to work on this thesis under her supervision. I am also deeply grateful to Prof. Guido Pagana for his assistance and support throughout this project.

I also want to extend my thanks to Dr. Marco Daperno, Prof. Rodolfo Rocca, and the staff of Fondazione IBD Onlus. Without their help, this work would not have been possible. Thanks for the collaboration.

Table of Contents

List of Tables

List of Figures

Acronyms

CD

Crohn's disease

EDA

Exploratory data analysis

GI

Gastrointestinal

IBD

Inflammatory bowel disease

IBD-DISK

Ulcerative colitis

MIAH-UC

Monitor IBD at home ulcerative colitis

P-SCCAI

Patient Simple Clinical Colitis Activity Index

RF

Random forest

ROC

Receiver operating characteristic

UC

Ulcerative colitis

Chapter 1

Introduction

1.1 Telemedicine

The term telemedicine refers to the science and practice of medicine over distance, through different communication channels; the operation of telemedicine then requires the help and assistance of a wide range of technologies, particularly, involving those who belong to the ICT field [1]. In the last decade, with the boost in the research and improvement of sectors such as video and image quality and other digital technologies, the growth of telemedicine is happening at a pace that was not expected [2]; overall, this field is developing as fast as the technologies whose support it; allow it to.

The primary objective of telemedicine is its integration into the dynamics of the conventional healthcare system, particularly providing support in tasks such as information transmission, disease diagnosis, patient monitoring, controls, and patient-physician communication. However, the concept of telemedicine is truly broad and ambiguous, considering it encompasses various technologies that may or may not have particularly different objectives; over time, concepts have emerged that target specific segments of telemedicine. One of these is known as telesupervision, where the physician can engage other doctors and specialists remotely in the diagnostic and clinical management process. On the other hand, telemanipulation refers to the remote use of medical devices in examinations or interventions, and finally, telemonitoring encompasses the monitoring and tracking of a patient's progress [3]. The field of telemedicine involves not only physicians and patients but also engineers, hardware manufacturing companies, researchers, and others. It is a highly multidisciplinary sector that involves the participation of different stakeholders.

An essential point to emphasize is that telemedicine per se does not aim to replace the traditional medical system, i.e., with the presence of healthcare personnel, but rather serves as an aid and support for both parties, the medical system, and patients. The primary characteristic of telemedicine is distance; from the patients' perspective, telemedicine can offer convenience and comfort, as it eliminates the need to travel to a medical center. This results in saved travel expenses and time, particularly addressing the problematic issue of patient travel, especially for those in rural and less urbanized areas where specialized medical centers may be less available. On the other hand, the medical system benefits from not being inundated with patients who do not require hospital beds, thus allowing prioritization of patients in critical conditions. Telemedicine also serves as an ideal tool for highly infectious diseases, where patient contact with third parties is not advisable. For instance, during the COVID-19 pandemic, remote medical consultations saw an exponential increase, highlighting the crucial role telemedicine played. Since then, the role of telemedicine has become more significant, given its substantial utility, allowing for further development and advancements in this field for the future [4]; lastly, another advantage to be noted about telemedicine is its suitability for the chronic disease management and supervision, which is one the case for this work, where the research is aimed to ulcerative colitis, a chronic inflammatory bowel disease.

1.2 Inflammatory bowel diseases

Inflammatory bowel disease (IBD) is a term used to describe a group of chronic inflammatory conditions that primarily affect the gastrointestinal (GI) tract. There are totally identified diseases related to IBD, Crohn's disease and ulcerative colitis. While the exact cause of IBD remains unclear, it is believed to involve a combination of genetic, environmental, and immune system factors. The precise origins of IBD are not fully understood, but several factors are thought to contribute to its development. Genetics play a significant role, as individuals with a family history of IBD are at a higher risk of developing the condition. Environmental factors such as diet, smoking, infections, and exposure to certain medications or pollutants may also trigger or exacerbate inflammation in susceptible individuals. Additionally, abnormalities in the immune system, particularly an overactive immune response targeting the GI tract, are believed to play a central role in the development of IBD [5].

Inflammatory bowel disease (IBD) presents with a variety of symptoms, including abdominal discomfort, chronic diarrhea often with blood or mucus, rectal bleeding, weight loss, and persistent fatigue. Additional symptoms may include fever, loss of appetite, nausea, vomiting, joint pain, skin problems, and eye inflammation. These symptoms can significantly impact quality of life and require comprehensive management by healthcare professionals. Diagnosis of IBD typically involves a combination of medical history, physical examination, laboratory tests, imaging studies (such as endoscopy and imaging scans), and sometimes, biopsy of the GI tract lining. Treatment for IBD aims to reduce inflammation, control symptoms, and improve quality of life. This may include medications such as anti-inflammatory drugs (e.g., corticosteroids, mesalamine), immunosuppressants, biologics (e.g., antibodies targeting specific molecules involved in inflammation), dietary changes, lifestyle modifications, and in some cases, surgery to remove diseased portions of the GI tract [5].

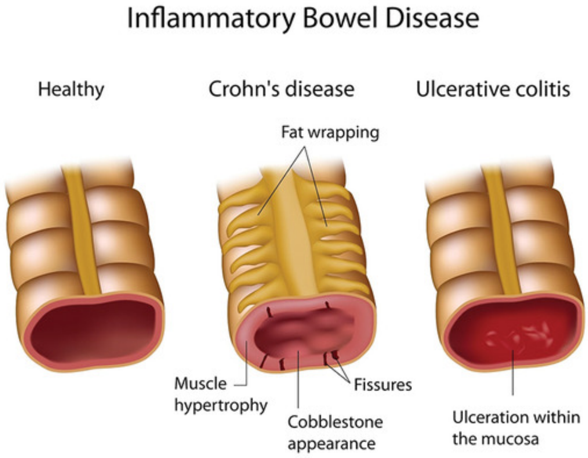


Figure 1.1: IBD: bowel overview and comparison among UC and CD.[6].

1.2.1 Crohn disease

Crohn's disease, a chronic inflammatory bowel disease (IBD), poses a significant challenge to patients and healthcare providers alike. This condition is characterized by inflammation that can affect any part of the gastrointestinal tract, from the mouth to the anus, leading to a diverse array of symptoms and complications. Manifestations of Crohn's disease vary widely among individuals, making diagnosis and management complex tasks. Central to the experience of Crohn's disease is the persistent abdominal discomfort that patients endure, often accompanied by cramping that can range from mild to debilitating. This abdominal pain, frequently localized in the lower right quadrant, becomes a constant reminder of the disease's presence and impact on daily life. Alongside this discomfort, chronic diarrhea plagues many individuals, disrupting normal bowel habits and contributing to feelings of urgency and distress.

The stool, in some cases, may exhibit blood or mucus, reflecting the underlying inflammation and damage to the intestinal lining. Furthermore, the insidious nature of Crohn's disease often leads to unintended weight loss despite adequate dietary intake, highlighting the metabolic toll of chronic inflammation and malabsorption. Beyond the physical symptoms, patients with Crohn's disease commonly experience persistent fatigue, which can be debilitating and challenging to manage. This fatigue, compounded by factors such as anemia and disrupted sleep patterns, further diminishes the quality of life and adds to the burden of living with a chronic illness. Overall, Crohn's disease presents a multifaceted clinical picture that demands a comprehensive approach to diagnosis, treatment, and ongoing management to address the diverse needs and challenges faced by patients [7].

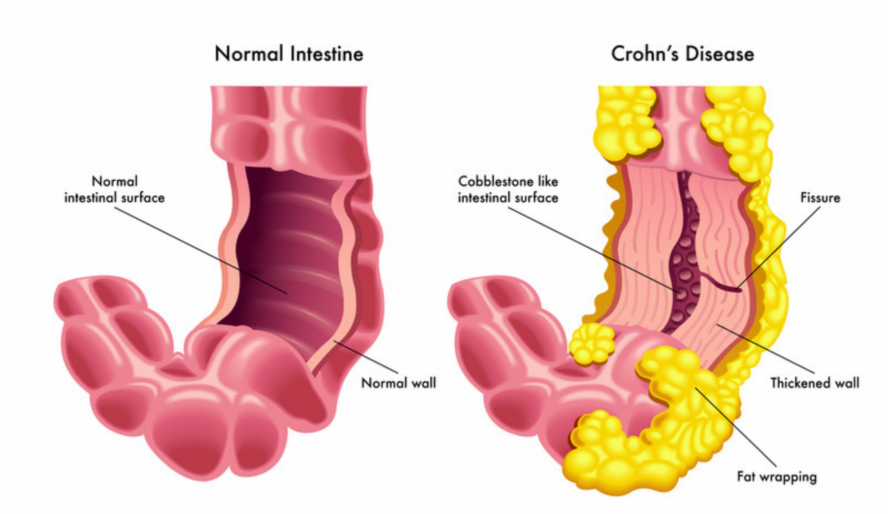


Figure 1.2: Crohn's disease intestine and healthy intestine comparison [8].

1.2.2 Ulcerative colitis

Ulcerative colitis is another common inflammatory bowel disease, it differs from Crohn's disease in the target of the gastrointestinal tract which ends up being affected; ulcerative colitis mainly targets the colon and the rectum, resulting in extensive inflammation and ulceration of the mentioned zones. The ulcerative colitis represents a critical challenge to both patients and physicians due to its chronic condition and degenerative nature [9].

The common symptoms of ulcerative colitis may encompass different affections. Bloody diarrhea, abdominal pain, urgency to defecate; from now on, we will to refer to the urgency to defecate as bowel urgency; an extremely important concept in this work. General fatigue and the feeling of no vitality are other usual clinical manifestations of ulcerative colitis. The conjunction of all these symptoms during the moments of high disease activity can profoundly impact the patients' quality of life; even from a psychological point of view [10].

Besides the commonly noted gastrointestinal symptoms, ulcerative colitis may present with other extra-colonic manifestations such as fever, weight loss, and arthralgia. All these symptoms contribute to the multi-faceted clinical presentation of the disease, highlighting the need for an extensive medical evaluation. Usually, the diagnosis of ulcerative colitis relies on a combination of different evaluations, laboratory investigations, endoscopic studies, and histological confirmation [11].

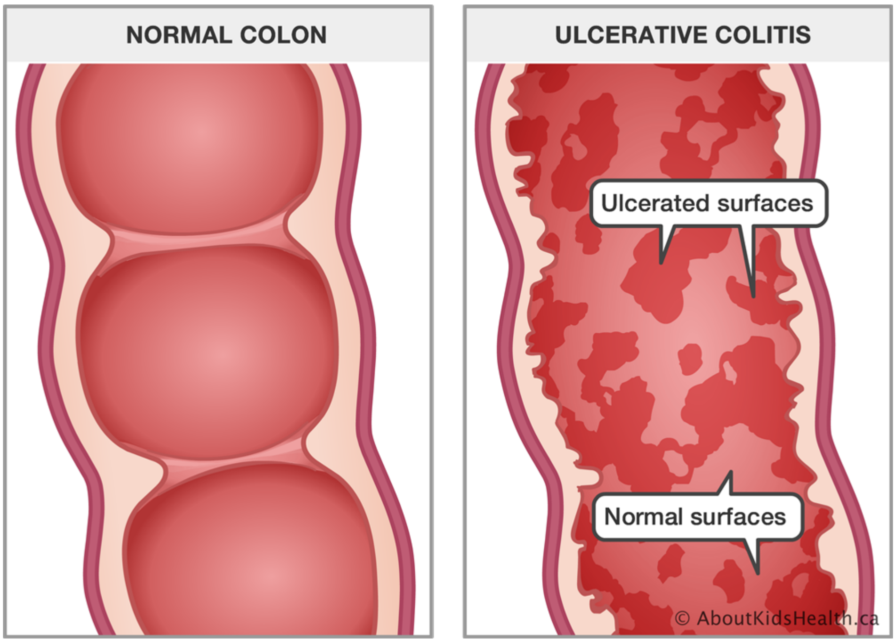


Figure 1.3: Ulcerative colitis comparison [12].

In summary, ulcerative colitis is a long-term chronic disease that causes inflammation in the colon and rectum. The symptoms can vary and may be present as colonic, such as diarrhea and bowel urgency, or extra-colonic, like fever and anemia. The treatment for ulcerative colitis involves different techniques, often given as a combination of medication, constant monitoring, and, in some cases, surgery. The causes of ulcerative colitis are still under study by the scientific and medical community, as well as research for a better understanding and improvement of the treatments.

1.3 Bowel urgency

Bowel urgency is a common symptom reported by patients diagnosed with ulcerative colitis, especially in periods of high disease activity. It can be defined as a sudden need to have a bowel movement; bowel urgency turns out to be a critical aspect related to the symptomatology given the fact that patients' quality of life and mental health may be compromised [13].

One of the common approaches to evaluate bowel urgency by the medical community is through the filling of questionnaires by the patients with a certain frequency; there are several types of questionnaires that try to assess different fields, such as the disease activity index, quality of life, quality of sleep, etc. This work has a big focus on one of them; the Simple Clinical Colitis Activity Index (SCCAI) was created in 1998 and one part of its scope addresses bowel urgency. Some patients report bowel urgency as the most unpleasant symptom and its absence as the top feature that improves the perceived quality of life [13], therefore a wide focus of the research of the pharmaceutical industry and medical field aims to cope with bowel urgency.

1.4 Ulcerative colitis: medical treatments

UC is one of the most common manifestations of IBD, it is a lifelong condition with degenerative characteristics. The medical treatments vary and are approached from a multidisciplinary point of view; the main objectives aim to reduce inflammation, manage colonic and extracolonic symptoms and maintain remission [14].

Different efforts are oriented toward the treatment and care of UC, such as surgery, and pharmaceutical medication; therapies are usually divided into induction and maintenance therapies. Also, the therapies are given based on the severity and the disease extent [15].

Regarding pharmaceutical therapy, there are two main types, biological and non-biological; the difference between these two concepts derives from the procedure in which they are created, biological drugs are derived from any biological source, meanwhile, in the creation of non-biological drugs, the source are mainly pharmaceutical chemical ingredients [16].

There are several pharmaceutical options from both types, from the non-biological side the most common are salicylates, corticosteroids, calcineurin inhibitors and immunomodulators [15]. On the other hand, regarding biological pharmaceutical therapy, the most common options are Anti TNF agents and Integrin antagonists [17].

There are several options available for therapy based on medication, the treatment supplied to the patients depends on their indicators and diagnosis. Nevertheless, drugs are not the only option for treating UC, there are more therapies involved in the UC treatment, for example, surgeries, exercise and diet, psychological follow-up, etc. Surgery, for instance, is usually the option in the most severe cases, some indications that it may be needed are toxic megacolon, perforation and severe colorectal bleeding, the common procedures in these situations are colectomy and ileostomy [18]. Some researchers affirm that a healthy and active lifestyle may contribute to the well-being, prevention, pathogenesis, and management of UC, as suggested in [19].

1.5 Telemedicine for IBD

Inflammatory bowel diseases are lifelong chronic diseases; they require constant monitoring to follow up on the patient's health status. Patients could register their health status on a particular system and if required, set an appointment with a physician. At the same time, the physicians may prescribe medication or other kinds of treatment and monitor the evolution of the patient. Overall, it helps as a tool of interaction, awareness and monitoring in the patient-physician interaction, which does not replace in-person medical treatment, if ever necessary.

1.5.1 IBD Tool

IBD Tool is a telemedicine application intended for IBD patients; it establishes a direct channel of communication between patients and physicians. The core of IBD Tool is the monitoring of patient health status by compiling a set of questionnaires related to well-being, the evolution of user responses is the key for physicians to evaluate the evolution of the illness, but also, it also becomes an accessible communication channel between patients and physicians.

1.5.2 MyIBDCoach

[H] MyIBDCoach is a digital health developed in the Netherlands, this platform supports patients diagnosed with any IBD. This tool aims to improve disease management, enhance communication between patients and healthcare providers, and ultimately improve patient outcomes.

Figure 1.4: MyIBDcoach project logo [20].

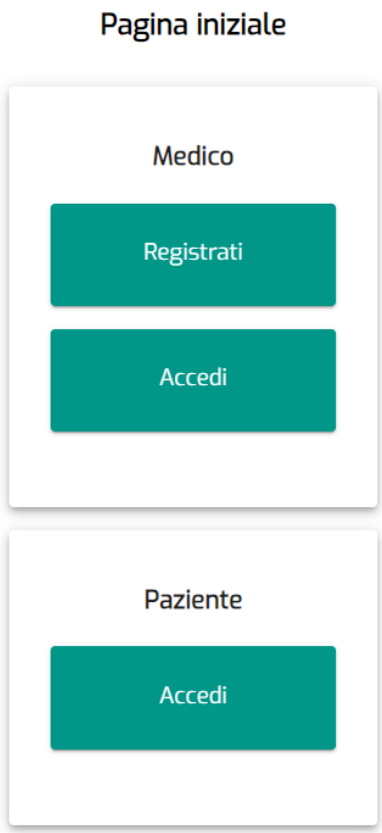


Figure 1.5: IBD Tool web application log-in interface

1.6 Objectives

[H] The purpose of this work is to analyze and understand the behavior of the bowel urgency degree in patients diagnosed with UC or an IBDU who are active users of IBD Tool web application.

The IBD Tool platform avails a set of variables associated to the diagnosis and more medical features from the patients; and, also, contains the information about the questionnaires that are sent to patients to collect information about their well-being and symptoms directly and indirectly associated with the diagnosis of the IBD.

Bowel urgency is particularly problematic since it affects the lifestyle and general perception of the quality of life of patients severely; the idea is to create different prediction models that can classify the bowel urgency registered by a patient based on the responses given in their questionnaires, medical data about their diagnosis and data about their undergoing pharmaceutical treatment. This will give insights into how all the available features are going to affect the degree of bowel urgency of a patient.

The importance of understanding bowel urgency points towards the well-being and general satisfaction of the patients; and in future works; a closer collaboration between classification and prediction stages of bowel urgency together with different strategies of therapy by the physicians may improve the welfare of the patients.

Chapter 2

Architecture of the system

2.1 Databases

A database is an organized collection of structured information or data, typically stored electronically in a computer system. It is designed to efficiently manage, retrieve, and update data according to various requirements and constraints.

2.1.1 Relational and non-relational databases

Relational databases are structured systems that organize data into tables with rows and columns, and they rely on SQL for data management and querying. These databases excel in scenarios where data relationships are well-defined and require strong consistency and integrity, such as in financial transactions or inventory management systems. The ACID properties ensure that transactions are reliable and maintain data integrity, which is crucial for many enterprise applications.

On the other hand, non-relational databases, often referred to as NoSQL databases, offer more flexibility in handling diverse data types and structures. They are designed to manage unstructured or semi-structured data more efficiently than relational databases. NoSQL databases come in various types, including document-oriented, key-value, columnar, and graph databases. Each type is optimized for specific use cases, such as storing and retrieving JSON documents, handling high-velocity data streams, or modeling complex relationships between entities.

Non-relational databases are well-suited for applications requiring horizontal scalability and distributed architectures, as they can easily scale out across multiple servers or nodes. They sacrifice some of the strict consistency guarantees of relational databases in favor of increased scalability and performance. NoSQL databases are commonly used in web applications, big data analytics, real-time data processing, and other scenarios where traditional relational databases may struggle to meet the performance demands or handle the diverse nature of the data.

In summary, relational databases offer strong consistency and data integrity, making them suitable for structured data and transactional systems. Meanwhile, non-relational databases provide flexibility, scalability, and performance advantages for handling diverse, unstructured, or semi-structured data in distributed and high-traffic environments. The choice between relational and non-relational databases depends on the specific requirements of the application, including the nature of the data, scalability needs, and performance considerations.

2.1.2 Example: Mongo-DB

All the information regarding the IBD Tool is stored in a MongoDB database. MongoDB is an open-source, non-relational database management software. Traditionally, databases used to follow a table/schema structure completely rigid, where each field (column) of each record (row) can be visualized clearly. Relational databases are also known as SQL databases; SQL stands for Standard Query Language, which is a domain-specific language and serves as the interface for communication and interaction with relational databases, this permits the user to modify, read, and do more complex queries with the data.

On the other hand, Non-relational databases, also called No-SQL (Not-only Structured Query Language), are not structured in a traditional row and column fashion but in other different data models which depend on the nature of the data, some forms of structure are documents, key-values, graphs, etc. This plays a big role in cases where the information is large-scale and unstructured. For MongoDB, the data model is document-oriented. This data model allows for have flexible schema that can be modified as the application evolves and the data recorded even if it needs to change; as of this flexibility, non-relational databases are widely used in a variety of industries and use cases.

In document-oriented databases, a document is a record that stores any information about an object; documents are stored in a field-value fashion, and they can support different data types such as strings, numbers, dates, arrays, tuples, etc. In MongoDB, documents are stored in a JSON-like format called BSON (Binary JSON). JSON stands for JavaScript Object Notation and it is a markup language for object interchange. In MongoDB, a group of documents makes up a collection, as MongoDB is a non-relational database the documents stored in the very same collection may not follow the same fields or are structured [21].

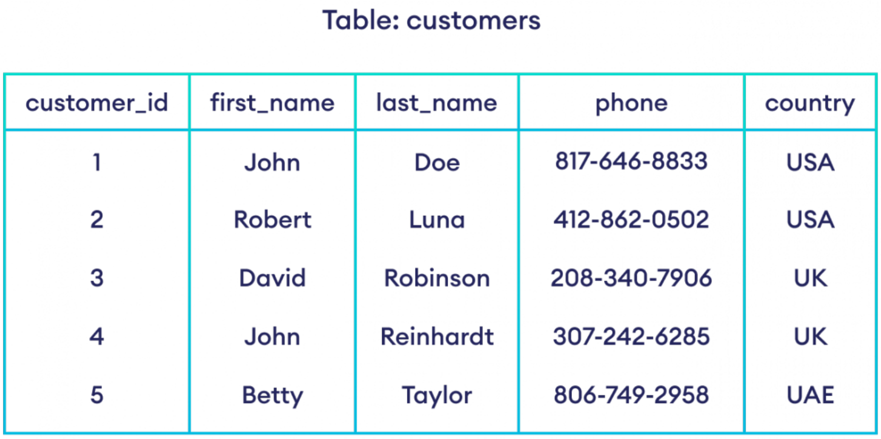


Table: customers

customer_id	first_name	last_name	phone	country
1	John	Doe	817-646-8833	USA
2	Robert	Luna	412-862-0502	USA
3	David	Robinson	208-340-7906	UK
4	John	Reinhardt	307-242-6285	UK
5	Betty	Taylor	806-749-2958	UAE

Figure 2.1: A typical relational database is made up of rows and columns [22].

Figure 2.1 shows a classical relational database, they are filled like a matrix, with rows and columns; one of the conflicting issues one may think about this architecture is the horizontal scalability for large volume data, and also, key for the context of this work, unstructured data handling, which is not handled particularly well. On the other hand, Figure 2.2, displays the architecture of a document-oriented database, like MongoDB, perfectly suited for applications where the data may not be strictly and consistently structured.

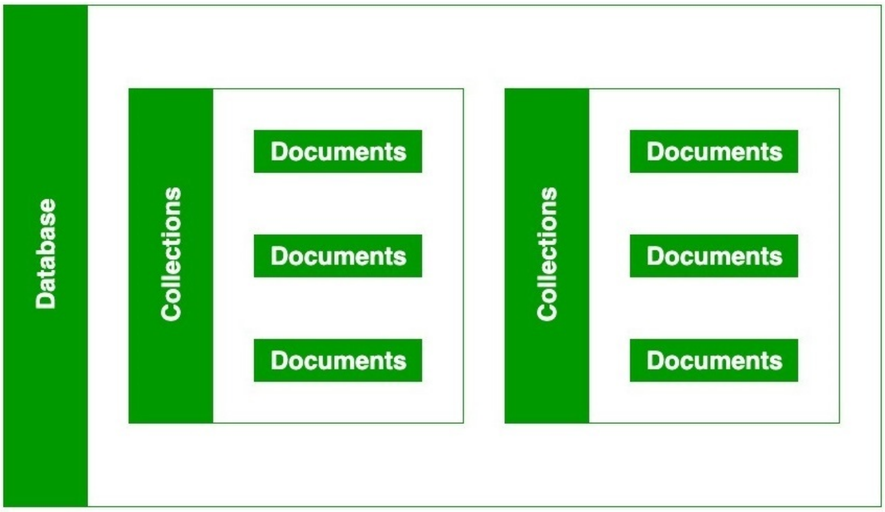


Figure 2.2: Structure of a MongoDB database [21].

2.2 Questionnaires for ulcerative colitis

The core functionality of IBD Tool is that the platform revolves around administering questionnaires. When a user-doctor registers a new patient-user and specifies their condition, an automated process initiates to send out questionnaires. These questionnaires are structured into monthly, quarterly, and half-yearly intervals. They serve various purposes and functionalities, and the selection of questionnaires isn't uniform across all users or consistent in terms of administration frequency [23]. There are several types of questionnaires oriented to patients diagnosed with any IBD. Initially; the allocation of the questionnaires depends on the disease diagnosed.

The IBD Tool web application is used by the physicians of the gastroenterology area from Mauriziano hospital in Turin, Italy; another section regarding the questionnaires allocation depended on the classification of the patients; originally a patient was asked to be classified voluntarily as a "TELEMEDICINE" patient; which would mean to receive the set of assigned questionnaires monthly and a special telemonitoring assessment from part of the physicians; otherwise, the classification of a patient would be "STANDARD", which receive the questionnaires three-monthly. However; for the following research of this work this classification was suggested by the head physician of the hospital to be ignored since the compliance and objectives associated with the TELEMEDICINE class were not met.

2.2.1 SCCAI

The Simple Clinical Colitis Activity Index is a questionnaire used as a clinical tool to evaluate the severity and track the symptoms associated with ulcerative colitis in adults; it is composed of a set of questions that address different fields that can be affected due to the activity of UC. There are two types of SCCAI questionnaires, one is denominated as Clinician-based Simple Clinical Colitis Activity Index, and the other, Patient-based Simple Clinical Colitis Activity Index. They both ask the patients about the symptoms of the previous week, however, the patient-based type is intended to be written more simply, avoiding medical terminology that may not be familiar to the patient [24].

Questions
1. On average per day (24 hours), how many times did you use the toilet for defecation during the previous week? Blood and slime discharge is also considered as defecation. ☐ 0 to 3 times (score 0) ☐ 4 to 6 times (score 1) ☐ 7 to 9 times (score 2)

☐ More than 9 times (score 3)

2. On average per night, how many times did you get out of bed to use the toilet for defecation during the previous week?
☐ None (score 0)
☐ 1 to 3 times (score 1)
☐ More than 3 times (score 2)

3. During the previous week, were you able to hold up your stool for 15 minutes or longer, when you felt the urge to use the toilet?
☐ Yes (score 0)
☐ No (score 1)
☐ I do not know* (score 0)

4. During the previous week, did you have to make adjustments to your activities, to ensure that there was a toilet nearby?
☐ Yes (score 0)
☐ No (score 1)
☐ I do not know* (score 0)

5. During the previous week, have you found stools in your underwear?
☐ Yes (score 1)
☐ No (score 0)
☐ I do not know* (score 0)

6. During the previous week, how many times did you see blood in your stool?
☐ Never (score 0)
☐ Less than half of the times (score 1)
☐ At least half of the times (score 2)
☐ I do not know* (score 0)

7. How would you rate your general well-being during the previous week by giving it a number, what number would you choose? (1 = very bad, 10 = perfect)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

8. During the previous week, did you have joint pain which was worse at rest than after activity?
☐ Yes (score 1)
☐ No (score 0)
☐ I do not know* (score 0)

9. During the previous week, were your joints red or swollen?
☐ Yes (score 1)
☐ No (score 0)
☐ I do not know* (score 0)

10. During the previous week, have you ever woken up from joint pain?
☐ Yes (score 1)
☐ No (score 0)

□ I do not know* (score 0)
11. During the previous week, have you had a skin disorder that has been diagnosed as erythema nodosum by your treating specialist? □ Yes (score 1) □ No (score 0) □ I have a skin disorder but have not seen my specialist for it or do not know what the disorder is called.* (score 0)
12. Do you momentarily have an eye infection, that you have seen an eye-specialist for and for which your treating specialist diagnosed as uveitis? □ Yes (score 1) □ No (score 0) □ I have an eye infection but have not seen an eye specialist for it or do not know what the infection is called.* (score 0)

Table 2.1: Patient-based SCCAI questionnaire [24].

Tables 2.1 and 2.2 display the questions that compose both questionnaires and their scoring system, the final calculation of the score is an arithmetic sum. The relative health status of the patient can be derived from the result, a score lower than 5 means remission, and a score equal to or higher than 5 means relapse.

2.2.2 MIAH-UC

MIAH-UC stands for Monitor IBD at Home - ulcerative colitis; it is a questionnaire intended to monitor the colonic activity in adult patients diagnosed with ulcerative colitis.

Table 2.3 shows the questions associated to the MIAH-UC questionnaire, the calculation of the final score follows the equation:

$$X = \frac{10 \cdot (e - 0.611 - 0.160 \cdot \text{answer domain } 1 + 2.103 \cdot \text{answer domain } 2 + 0.167 \cdot \text{answer domain } 3) - 0.184 \cdot \text{answer domain } 4}{1 + e - 0.611 - 0.160 \cdot \text{answer domain } 1 + 2.103 \cdot \text{answer domain } 2 + 0.167 \cdot \text{answer domain } 3 - 0.184 \cdot \text{answer domain } 4}$$

The state of the patient is obtained by comparing the result of the equation portrayed above with the threshold of 3.6 points; if the score registered by the user is equal to 3.6 or higher it means relapse, otherwise, it means remission. Domain 4 refers to questions 4 and 5 of table 2.3 weighted by 0.5 and summed.

2.2.3 IBD-DISK

The Inflammation-Bowel-Disease Disability Index (IBD-DISK) is oriented for patients diagnosed with any IBD, and it groups several colonic and non-extracolonic features derived from the symptoms of any IBD.

Variable	Description	Scoring
1	Bowel frequency (day)	n (1 per occurrence) 0 – 3 (score 0) 4 – 6 (score 1) 7 – 9 (score 2) > 9 (score 3)
2	Bowel frequency (night)	0 (score 0) 1 – 3 (score 1) 4 – 6 (score 2)
3	Urgency of defecation	None (score 0) Hurry (score 1) Immediately (toilet nearby) (score 2) Incontinence (score 3)
4	Blood in stool	None (score 0) Trace (score 1) Occasionally frank (<50% of defecation) (score 2) Usually frank (>50% of defecation) (score 3)
5	General well-being (0 – 10)	≥ 7 – very well (score 0) 6 – slightly below par (score 1) 5 – poor (score 2) 4 – very poor (score 3) < 4 – terrible (score 4)
6	Extracolonic features	1 per manifestation: Arthritis Yes = 1 No = 0 Uveitis Yes = 1 No = 0 Erythema nodosum Yes = 1 No = 0 Pyoderma gangrenosum Yes = 1 No = 0

Table 2.2: Clinical-based SCCAI questionnaire [24].

2.2.4 Other questionnaires

The questionnaires SCCAI, MIAH-UC and IBD-DISK are the most important for the analysis of bowel urgency, mainly due to their aimings; since there are objective questions related to bowel urgency and as well as colonic features closely related to it and also, these questionnaires are filled monthly, so their availability is high.

Questions	**Answers**	**Scores**
If you had to rate your general health status by assigning a number, what number would you choose?	Scroll bar (0=very bad, 10=perfect)	VAS scale 0 to 10
Rectal bleeding	• Yes • No	10 0
Number of bowel movements per day	Empty field, possibility to indicate an integer number	$< 4 = 0$, $>= 3 = 10$
Defecation urgency	• No • Yes, urgent	0 10
Abdominal pain	Scroll bar (0=very bad, 10=perfect)	0 to 10
Final score		Solve the algorithm

Table 2.3: MIAH-UC questionnaire [25].

Questions	Answers	Scores
1. Abdominal pain: Have you had stomach or abdominal pain?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
2. Bowel control: Have you had difficulty in controlling and coordinating your bowel movements, including choosing a suitable place to defecate and cleaning up afterwards?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
3. Interpersonal interactions: Have you had difficulty with personal relationships or participating in the community?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
4. Education and work: Have you had difficulties with schoolwork or study, or with work or chores at home?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
5. Sleep: Have you had difficulty sleeping, for example, with falling asleep, waking up frequently at night, or waking up too early?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
6. Energy: Have you not felt rested during the day, feeling tired and lacking energy?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
7. Emotions: Have you felt sad, down, or depressed, or worried or anxious?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
8. Body image: Have you not been happy with how your body or parts of your body look?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
9. Sexual function: Have you had difficulties with sexual thoughts or physical aspects?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
10. Joint pain: Have you had pain in your joints?	Scale (0: Totally disagree, 5: Neither agree nor disagree, 10: Totally agree)	0 to 10
Final score	Sum of the ten answers scores	0 to 100

Table 2.4: IBD-DISK questionnaire [25]

However, IBD Tool web application also delivers other questionnaires to UC patients, they address other fields and are filled out less frequently than the previous three mentioned:

- PHQ9 (Patient Health Questionnaire-9): questionnaire is filled out every 3 months and evaluates the symptoms of depression in IBD patients.
- IPAQ-SF (International Physical Activity Questionnaire Short Form): filled out every 3 months and evaluates the amount and type of physical activity performed in the last seven days.
- WPAI (Work Productivity and Activity Impairment Questionnaire): filled out every 3 months addressing the impact of the disease in the work and daily activities.
- MMAS8 (8-item Morisky Medication Adherence Scale): filled out every 3 months addressing the compliance with the medical treatment given to the patient.
- TSQM (Treatment Satisfaction Questionnaire for Medication): filled out every 3 and evaluates the satisfaction of the patient with the supplied medication.
- PSQI (Pittsburgh Sleep Quality Index): filled out every 6 months and evaluates the sleep quality of the patient diagnosed with any IBD.
- EQ5D5L (5-level EQ-5D (European Quality - version 5D - 5 Levels): filled out every 6 months and monitors the measurement of the quality of life, work, personal care, physical and psychological health, usual activities, pain and discomfort.

2.3 IBD Tool Database

IBD Tool uses MongoDB as database management service. This database is called *ibdtool* and has 9 collections:

- The collection *announcements* contains all important and novel announcements published by medical doctors [26].

```
_id: " "
doctorName: ""
doctorSurname: " "
date:
text: " "
```

```
_class: "com.backend.web.ibdtool.entity.Announcement"
```

- The collection *assignedPatients*, in which each document represents a medical doctor and an array that contains the patients assigned to them.

  ```
  _id: " "
  patients: Array
  _class: "com.backend.web.ibdtool.entity.AssignedPatients"
  ```

- *chatMessages* collection stores all messages between patients and medical doctors [26].

  ```
  _id: " "
  sender: " "
  recipient: " "
  patientEmail: " "
  patientName: " "
  patientSurname: " "
  doctorName: " "
  doctorSurname: " "
  date:
  text: " "
  read: bool
  _class: "com.backend.web.ibdtool.entity.ChatMessage"
  ```

- Then there is the collection *clinicalDataPatients*. This is a crucial collection since it contains all the clinical data retrieved from the patients during visits to their hospital, such as examinations, ongoing treatments, history of evolution and more, there are multiple arrays inside the object *bodystat* where all those registers are stored.

  ```
  _id: " "
  bodyStat: Object
  lastLogin: 2022-06-22T20:55:43.042+00:00
  category: "TELEMEDICINA"
  registrationTime: 1596823125
  _class: "com.backend.web.ibdtool.entity.ClinicalDataPatients"
  ```

- The collection *pending* stores documents about questionnaires still to be read by medical doctors or fulfilled by patients [26].

  ```
  _id:  " "
  uuid: " "
  type: "PATIENT-SCCAI"
  doctorID: " "
  patientSSN: " "
  date:
  _class: "com.backend.web.ibdtool.entity.Pending"
  ```

- The collection*questionnaires* contains the information about a certain questionnaire fulfilled by a patient, like the answer to each question, the type of questionnaire, also if a warning is given to the patient.

  ```
  _id: " "
  type:"PATIENT-SCCAI"
  doctorID: " "
  patientSSN: " "
  compiled: true
  date: 2020-08-14T16:56:42.057+00:00
  results: Array
  finalScore: 5
  read: true
  evaluation: false
  warning:false
  _class: "com.backend.web.ibdtool.entity.QuestionnairePatientSCCAI"
  ```

- Collection *questionnairesToNotify* stores all the results of questionnaires that must be sent to the doctor [23].

  ```
  _id: " "
  type: "PSQI"
  score: 0
  doctorID: " "
  patientSSN: " "
  patientName: " "
  patientSurname: " "
  ripresaAttivita: false
  _class: "com.backend.web.ibdtool.entity.QuestionnaireToNotify"
  ```

- The collection *userNotifications* stores all the relevant messages that IBD Tool users both patients and medical doctors need to be aware of.

```
_id: " "
notifications: Array
_class: "com.backend.web.ibdtool.entity.UserNotifications"
```

- Finally, the last collection is *users*, which stores all information about users registered in IBD Tool, whether they are patients or medical doctors.

```
_id: " "
name: " "
surname: " "
SSN: " "
phoneNumber: " "
password: " "
role: Object
enabled: true
emailValid: true
registrationTime:
doctorID: " "
chatNotificationEnabled: false
birthDate:
birthPlace: " "
deadlineQuestionnaireModified: false
timestampLastMonthlyQuestionnaire: 0
timestampLastThreeMonthlyQuestionnaire: 0
timestampLastBiannualQuestionnaire: 0
doctorRole: "JUNIOR"
tokens: Array
lastLogin: 2023-06-08T18:57:17.605+00:00
usageTime: 920918
_class: "com.backend.web.ibdtool.entity.UserEntity"
```

However, for this study, only two collections will be used. The rest of them contain no relevant information for exploring the bowel urgency degrees in selected patients with UC or IBDU.

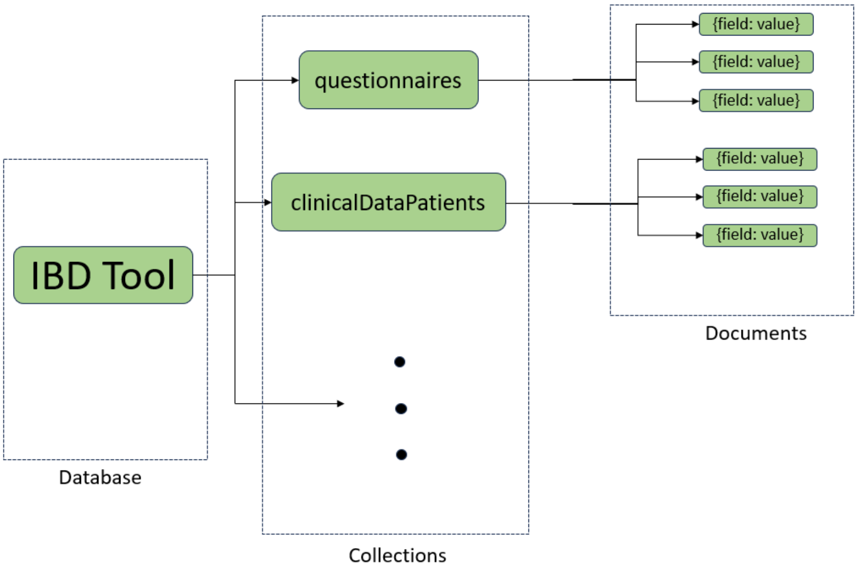


Figure 2.3: Structure of IBD Tool database with collections of interest

2.3.1 Questionnaires collection

The questionnaire collection is a key part of this work since it stores a big part of the information that is used in the analysis, and on the other hand, the output variable, the bowel urgency degree is encapsulated inside them. Nevertheless, as explained before, the focus is on three questionnaires, MIAH-UC, P-SCCAI and IBD-DISK.

```
_id: " "
type:" "
doctorID: " "
patientSSN: " "
compiled: true
date: 2020-08-14T16:56:42.057+00:00
results: Array
finalScore: 5
read: true
evaluation: false
warning:false
_class: "com.backend.web.ibdtool.entity.Questionnaire****"
```

The detailed fields of the documents inside the questionnaire collections store information about the patient identifier, the assigned doctor, the filing date, the

Type	Count
PSQI	871
EQ5D5L	890
SIBDQ	1
MIAH-CD	4149
IBDQ	1890
CLINICAL-HBI	749
WPAI	1809
CLINICAL-SCCAI	649
CLINICAL-PRISM	1399
IPAQ-SF	1897
TSQM	1705
Gradimento	473
HBI	4308
MIAH-UC	3941
LARS	1018
PATIENT-SCCAI	4116
IBD-DISK	8117
PRISM	2902
MMAS8	1804
PHQ9	1945

Table 2.5: Count of questionnaires per type inside the collection.

final score, and the evaluation. The field results contain an array associated with the response given by the patient to each of the answers to the corresponding questionnaire. Table 2.5 reports the count of questionnaires inside the IBD Tool project database.

2.3.2 Clinical data collection

The clinical data collection contains the medical information associated with the patients, it registers it as a dictionary in the field bodyStat.

```
_id: " "
bodyStat: Object
lastLogin: 2022-06-22T20:55:43.042+00:00
category: "TELEMEDICINA"
registrationTime: 1596823125
_class: "com.backend.web.ibdtool.entity.ClinicalDataPatients"
```

Field	**Value**
`_id`	`Tax code`
`bodyStat.weight`	Float
`bodyStat.height`	Integer
`bodyStat.pathology`	String
`bodyStat.age`	Integer
`bodyStat.sex`	Boolean
`bodyStat.dateOfDiagnosis`	Data object
`bodyStat.ageOfDiagnosis`	Integer
`bodyStat.familiarity`	Integer
`bodyStat.pathologyDuration`	Integer
`infiammatorioCD`	Array
`stenosanteCD`	Array
`penetranteCD`	Array
`malattiaPerianaleCD`	Array
`locatColonCD`	Array
`locatIleumCD`	Array
`locatUpperGI`	Array
`localizzazioneUC`	Array
`storicheEIMS`	Array
`attiveEIMS`	Array
`interventi`	Integer
`numResezioni`	Integer
`stomia`	Array
`therapy5ASAOS`	Array
`therapyTOPICA`	Array
`therapyGCSOS`	Array
`therapyIMM`	Array
`therapyBIO`	Array
`lastLogin`	Date object
`category`	String
`registrationTime`	Timestamp

Table 2.6: Information stored within clinical data collection.

Table 2.6 shows all the information associated to the clinical information of a patient stored within the clinical data collection of the IBD Tool database. There are different and various fields like weight, age, and sex, but also, fields that specify about medical parameters of the diagnosed disease, and about the therapies underdoing.

Chapter 3

Methods and materials

This chapter informs about a detailed level of the methodology employed to undergo through the thesis problems.

3.1 Data extraction

A critical step of the work carried out addresses data extraction. All the information regarding the IBD Tool web application project is stored within a centralized Mongo DB database. The data extraction process requires particular attention since it is essentially the input for further stages; in this scenario, a statistical representation of the data and different machine learning models will use the data extracted.

The work is focused on two collections from the IBD Tool database; questionnaire collection, which contains all the data related to the questionnaires filled by the users such as answers to each question, submission date, user identifier, and type of questionnaire; the other important collection is clinical data from the patients, this last encapsulates clinical information such as weight, height, the IBD diagnosis, sex, pathology duration, disease extent, age of diagnosis, and, extraintestinal manifestations of IBD, stomia, surgeries, and, in some particular cases, the information about the medical treatment undergoing.

3.1.1 Inclusion criteria

The inclusion criteria for a patient are:

- The patient has a confirmed UC/IBDU diagnosis
- The patient has not been operated on
- The patient has at least one p-SCCAI and one MIAH-UC pair of questionnaires completed within one week.

The first two criteria can be verified for each patient by observing its associated document in the clinical data collection, where all the health-related information is stored, on the other hand, confirming the existence of a pair of questionnaires needs to be done by looking at the collection of the questionnaires.

To downgrade the computational burden and exploit numerous Python libraries and resources, the relevant collections for the thesis are locally stored and represented as data frames using the Pandas library.

The pipeline for the data extraction begins by filtering the patients who meet the first two criteria in the clinical data collection, a confirmed diagnosis of UC or IBDU and no interventions undergone.

The following step is to exploit the collection of the questionnaires and extract all the questionnaires associated with users whose diagnosis is either UC or IBDU and who have not undergone any surgery.

3.1.2 Bowel Urgency: output variable

The output variable is the bowel urgency, and the value is extracted from the P-SCCAI questionnaire. The P-SCCAI questionnaire is a modified version of the original C-SCCAI questionnaire where the questions are written in a simplified manner, avoiding the user of complicated medical terminology that may not be comprehensible for a portion of the patients.

In the clinical SCCAI, the question that addresses bowel urgency is:

Question #3: Urgency of defecation Scoring:

- None (score 0)
- Hurry (score 1)
- Immediately (score 2)
- Incontinence (score 3)

On the other hand, the P-SCCAI, which is the questionnaire that the patients of IBD Tool fill up, the bowel urgency is composed of the following questions:

Question #3: During the previous week, were you able to hold up your stool for 15 minutes or longer, when you felt the urge to use the toilet?

Scoring:

- Yes (score 0)
- No (score 1)
- I don't know (score 0)

Question #4: During the previous week, did you have to make adjustments to your activities, to ensure that there was a toilet nearby?
Scoring:

- No (score 0)
- Yes (score 1)
- I don't know (score 0)

Question #5: During the previous week, have you found stools in your underwear?
Scoring:

- No (score 0)
- Yes (score 1)
- I don't know (score 0)

Overall, the bowel urgency is then addressed by the patients answering questions 3, 4 and 5 of the P-SCCAI questionnaire. The value which corresponds then to the bowel urgency at that timestamp when the patient filled and sent the questionnaire is then:

$$\text{Bowel Urgency} = \text{Question 3} + \text{Question 4} + \text{Question 5} \tag{3.1}$$

Bowel urgency equals the arithmetic sum of questions 3, 4 and 5. The main reason behind the split of the original bowel urgency question into 3 different ones is the fact that the new 3 questions address the problem not in an explicit fashion as the C-SCCAI question does but with slightly more implicit questions which, particularly asking about situations derived from the bowel urgency such as the affection to daily activities, making it easier for the patient to evaluate his/her bowel urgency degree with more attention.

3.2 Statistical review of the dataset

The section 3.2.1 reports some insights into the extracted data of all records that meet the data extraction criteria explained at 3.1.1.

Nevertheless, before that, some critical details about the initial availability of the data need to be explained. Even though the clinical data collection, which is one of the essential collections as the MongoDB database of the IBD Tool project contains fields related to the medical treatment of the patients, they are essentially empty, the reason behind this is the management of medical information of the patients that could not be digitalized into the database at the time they were supplied with any medical treatment, therefore, at the first stage, only few medical data was available in the database, mostly related to the basal characteristics of the patients, in detail, they are:

Variable	**Description**
Sex	Binary, 1 = female, 2 = male
Age	Integer
Age of diagnosis	Integer
Pathology duration	Integer
Strocihe EIMS	Integer
Stomia	Integer
Disease extent	Integer [1, 2, 3] depending on the location of the UC

Table 3.1: Table of initially available variables and their type

On top of the variables present in the table above, some others are not present in the clinical data collection but still are available in the questionnaires collection, they are the result of the P-SCCAI and MIAH-UC, which were mandatory criteria to created the records, eventually, bowel urgency is also present, since it is composed based on some answers from P-SCCAI questionnaire; and finally, the IBD-DISK questionnaire is one of the variables in consideration.

However, the analysis and potential machine learning models require more extensive data to function effectively. Therefore, an enrichment step was undertaken, involving the addition of several medical features primarily related to medication. This task was carried out by the staff at Mauriziano Hospital. Due to the sensitivity of medical records, the inclusion of this information into the dataset had to be handled by the hospital staff.

3.2.1 Characteristics of the extracted data

The extracted data counts 2768 rows, where all the criteria are met. Figure 3.1 displays the proportion of the sex based on the level of bowel urgency registered in each record. Bowel urgency varies from 0 to 3, each level corresponding to some degree of severity being 0 no urgency at all, 1 referring to hurry, 2 referring to the sudden need to go to the toilet and the most severe level, 3 referring to incontinence. The plot shows firstly that there are more records associated to women since the bars represent the contribution from records associated either to males or females to each degree of bowel urgency. Particularly, there are more records of females reporting at least any degree of urgency, i.e. from level 1 to 3, meanwhile, for cases of no urgency at all, there are more records that belong to male patients.

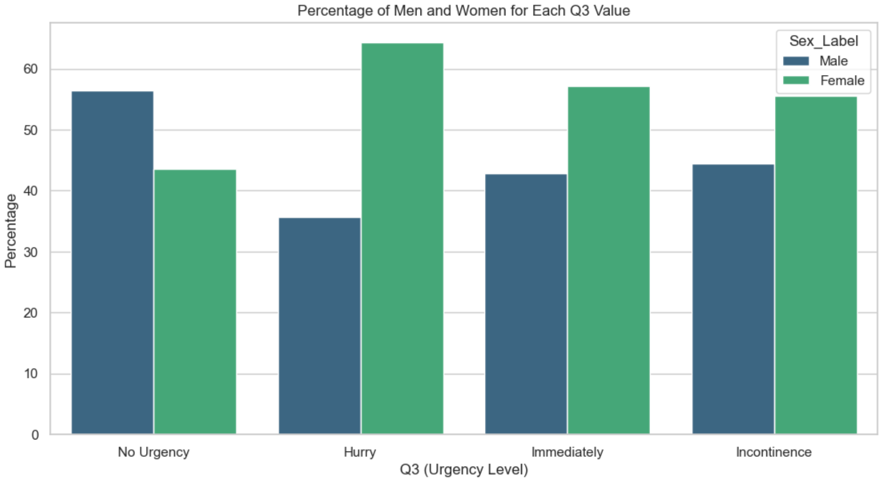


Figure 3.1: Distribution bowel urgency based on sex of patients from extracted dataset

Similarly, Figure 3.2 displays the information about the MIAH-UC score above or under 3.6. The reason behind choosing this particular value for the plot is rather simple, the MIAH-UC algorithm has a threshold to indicate the remission or relapse of a patient, in other words, the MIAH-UC questionnaire, which tracks the colon activity, marks a score higher than 3.6 as relapse (the patient having more negative effects), meanwhile, 3.6 is cataloged as remission (the patient feels better). Overall, there are more records where the score is lower than 3.6, however, the amount of relapses is quite high too.

Figure 3.3 displays the distribution of the ages associated with the records obtained from the data extraction process, particularly, the highest category is the patients between 20-40 years old, almost topping 50% of all the records. Since the IBD Tool is a project oriented toward the monitoring of the evolution of patients

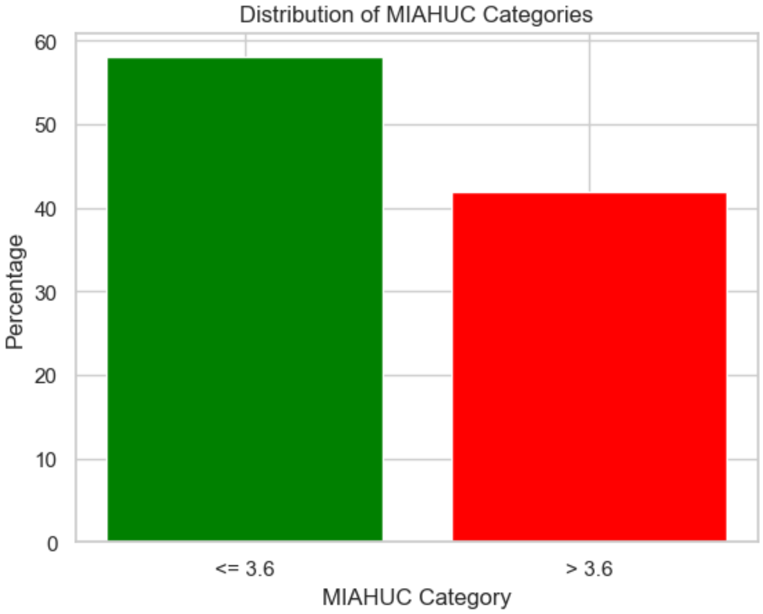


Figure 3.2: Distribution of scores from MIAH-UC questionnaires based on remission/relapse.

over 18 years old, the group under 20 years old is quite small but still relevant. The groups that represent ages over 40 are highly relevant, for example, records associated with patients between 40 and 50 years old roughly make the 25% of the population, the age group between 50 and 60 goes near to the 22%, and finally, patients over 60 years old still contribute with a population of records which makes close to 13%.

A view of the distribution of the different degrees of bowel urgency in the different age groups is visible in Figure 3.4. Starting with the youngest group, the patients under 20, particularly it has the highest rate of records associated with incontinence, which is the most severe degree of bowel urgency, however, just as explained before with Figure 3.3, since IBD Tool project's scope is adult patients, the representation of patients between 18 and 20 years old is not large, still, the high rate of incontinence is quite concerning, especially considering that the rate of immediately hurry, which is the second most severe degree is also considerably high. However, all groups have one characteristic in common, no urgency level being the most predominant degree of bowel urgency, especially in the age group of patients over 60 years old. Another interesting insight is the low rate of incontinence amongst records associated to patients between 50 to 60 years old, it is the lowest among all age groups, also, in contribution to that, the very same age group has the largest rate of records associated with a bowel urgency equal to 1.

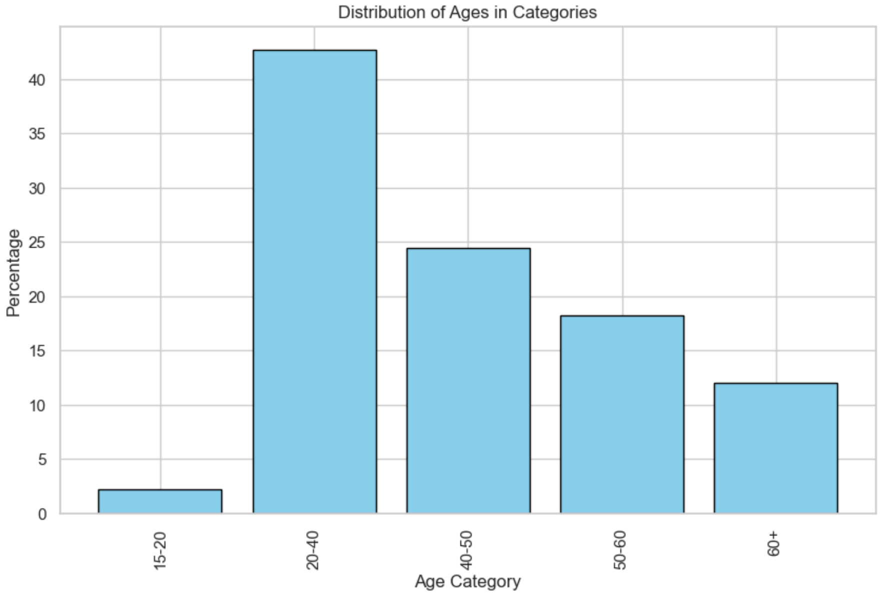


Figure 3.3: Distribution of age among records of the extracted data

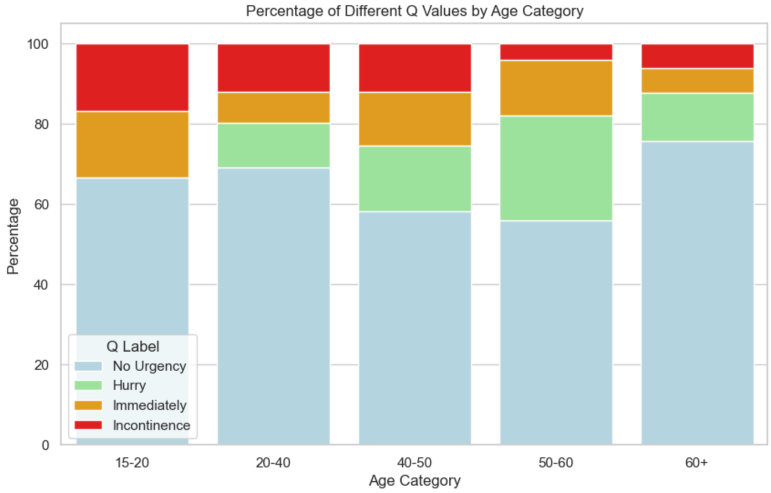


Figure 3.4: Distribution of bowel urgency degrees per age group

Figure 3.5 refers to the statistical distribution of the pathology duration of the patients who contribute to records in the extracted data, in other words, they have at least one pair of questionnaires MIAH-UC and P-SCCAI filled within one week, are diagnosed either with UC or IBDU and have been under any surgeries. Most of these patients have been diagnosed with ulcerative colitis during 10 to 20 years, this group makes up around 27% of the patients, the next two most prominent groups are patients whose pathology duration is between 0 and 5 years, and between 5 and 10 years, both of them contribute to a percentage close to 22% and 23%. On the other hand, the patients whose pathology duration is between 20 and 30 years is close to 17%. Moreover, there are two other groups, with pathology duration between 30 and 40 years, and the last one with more than 40 years, they sum up 8% and 3%, respectively.

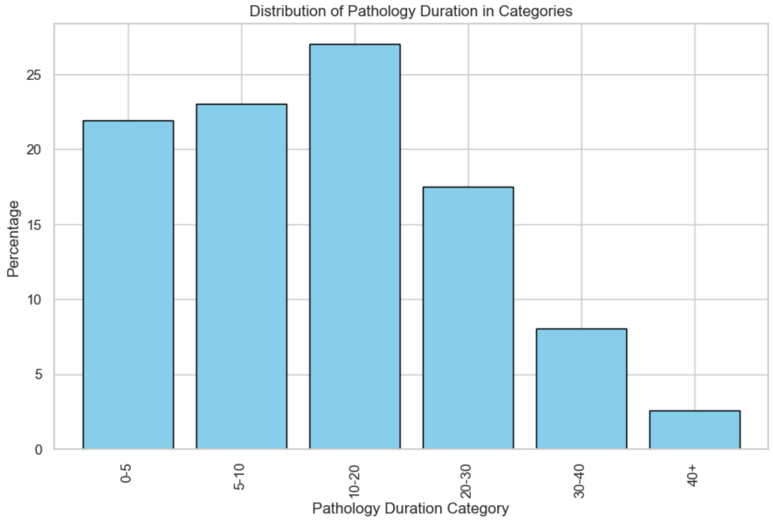


Figure 3.5: Distribution of the pathology duration variable.

Figure 3.6 complements Figure 3.5, it displays the composition of bowel urgency reported by each group based on the pathology duration. All groups follow a similar distribution of records associated with no urgency at all, this is, bowel urgency is non-existent, although the biggest contrast is found in the group between 5 to 10 years old pathology duration, where the numbers of records associated with incontinence are considerably high, close to an 18%. Another noticeable issue is the absence of records of incontinence of patients with pathology duration over 40 years, however, the main cause behind this is their low compliance and participation in the project. It is important to remember that the distribution of the figures is just an aisled case, a very particular one which regards participants of patients, the one which is under study and can not be used to conclude the general behavior of all patients who may be diagnosed of a UC.

The scores associated with the IBD-DISK questionnaires are shown in Figure 3.7, the results are quite sparse, there may be several reasons behind this, firstly, this questionnaire leans toward a subjective evaluation of each patient, where different fields like feelings, sleep quality, daily activities, self-perception are evaluated, therefore, patients may perceive their disease in totally different ways, however, it is still valuable information for bowel urgency analysis; as patients report that bowel

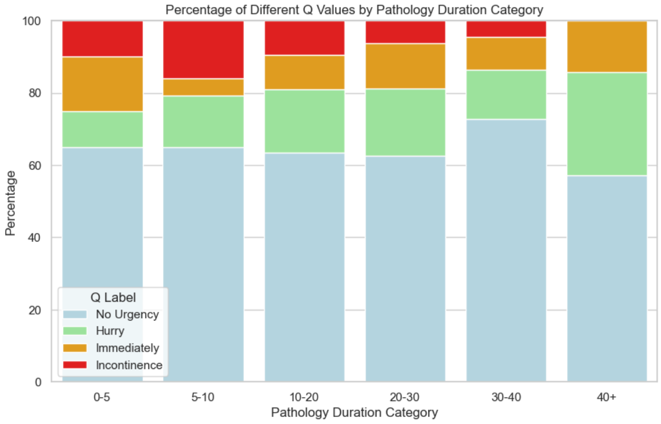


Figure 3.6: Percentage of different bowel urgency values by pathology duration category

urgency and incontinence are the most undesired and unsatisfactory symptoms raised from ulcerative colitis.

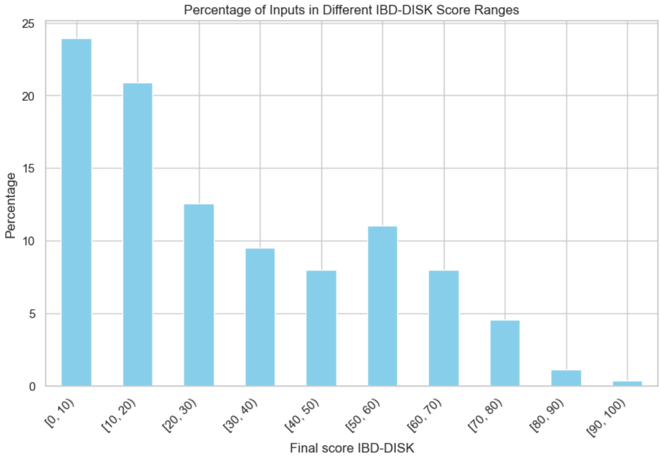


Figure 3.7: Distribution of scores associated to IBD-DISK questionnaire.

3.2.2 Correlation among variables: initial data

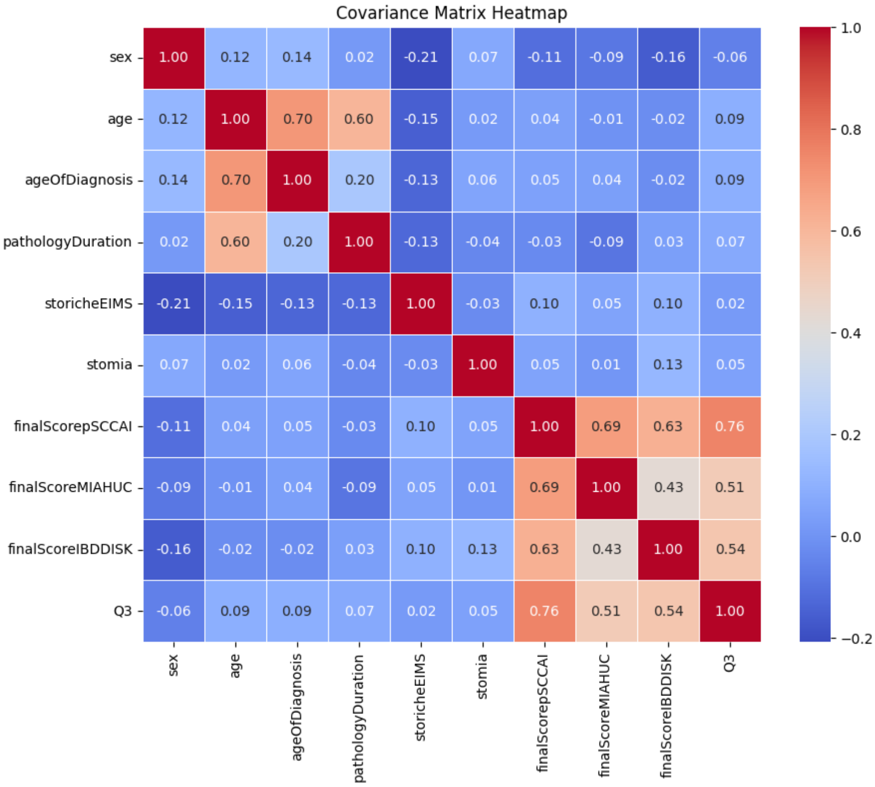
One of the most common approaches to understanding the nature and correlation of the variables within a system is to verify their correlation. Essentially, the covariance can be easily understood by thinking of how much two variables change together, e.g. it is how much is the variation of one feature concerning the other. In machine learning algorithms, which are the ones this work will be based on, the covariance among variables ends up being a topic of interest; the reason behind this is that by checking the covariance one can make an idea of the importance of the features with respect to the output variable, in simpler words, it is possible to check features whose variation follows the same pattern as the output variable, for instance, the output variable increases or decreases when a given variable does it so, the process of analyzing the features and selecting some of them and form a subset is known as feature selection, it is a good practice to understand the nature of the variables and, in some cases to form a subset of variables to work, those who are more informative; this may contribute in the performance of the model.

The collinearity is another important concept that arises from the analysis of the covariance of the variables, particularly speaking, is not a good idea to have multiple features that are linear combinations of others, this contributes to unstable models and can lead to undesired conditions such as overfitting.

The covariance matrix is built like this:

$$\Sigma = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & \operatorname{cov}(X_n, X_n) \end{pmatrix}$$

In the context of this thesis, the covariance matrix with the initial available features is shown in Figure 3.8.



	sex	age	ageOfDiagnosis	pathologyDuration	storicheEIMS	stomia	finalScorepSCCAI	finalScoreMIAHUC	finalScoreIBDDISK	Q3
sex	1.00	0.12	0.14	0.02	-0.21	0.07	-0.11	-0.09	-0.16	-0.06
age	0.12	1.00	0.70	0.60	-0.15	0.02	0.04	-0.01	-0.02	0.09
ageOfDiagnosis	0.14	0.70	1.00	0.20	-0.13	0.06	0.05	0.04	-0.02	0.09
pathologyDuration	0.02	0.60	0.20	1.00	-0.13	-0.04	-0.03	-0.09	0.03	0.07
storicheEIMS	-0.21	-0.15	-0.13	-0.13	1.00	-0.03	0.10	0.05	0.10	0.02
stomia	0.07	0.02	0.06	-0.04	-0.03	1.00	0.05	0.01	0.13	0.05
finalScorepSCCAI	-0.11	0.04	0.05	-0.03	0.10	0.05	1.00	0.69	0.63	0.76
finalScoreMIAHUC	-0.09	-0.01	0.04	-0.09	0.05	0.01	0.69	1.00	0.43	0.51
finalScoreIBDDISK	-0.16	-0.02	-0.02	0.03	0.10	0.13	0.63	0.43	1.00	0.54
Q3	-0.06	0.09	0.09	0.07	0.02	0.05	0.76	0.51	0.54	1.00

Figure 3.8: Covariance matrix represented as a heat map with the first initial features.

As seen in Figure 3.8, Q3 represents the value of bowel urgency, up to this moment, the highest correlated features are the questionnaires, P-SCCAI, MIAH-UC and IBD-DISK.

3.3 Kaplan-Meier curves

The Kaplan-Meier curves are a common statistical resource widely used in the medical field to portray the rate of survival from event-driven data. However, the Kaplan-Meier curves are not only used in survival analysis strictly, the key requirement is that the data involves tracking the time until an event of interest occurs, the event can be anything, in the scenario of this work, a nice idea is to think about the remission or relapse of a patient, based on the changes registered in the scores of their questionnaires. In fact, this kind of analysis is commonly found in literature related to data analysis of UC and related IBD, for example, in [27], Kaplan-Meier curves are used to predict the relapse of UC patients after using a particular medication.

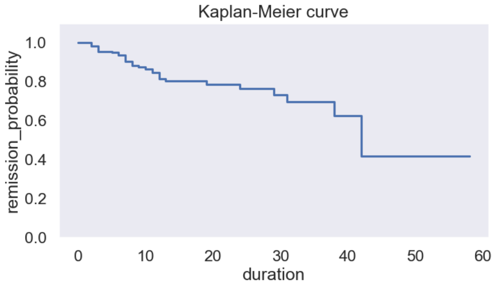


Figure 3.9: Kaplan-Meier curve: remission

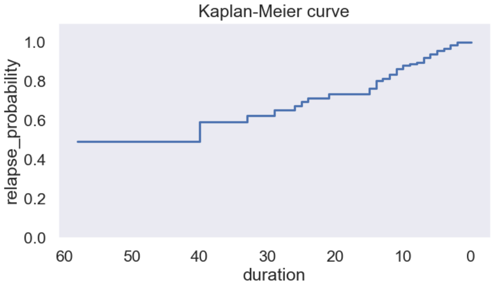


Figure 3.10: Kaplan-Meier curve: relapse

Figures 3.9 and 3.10 show the Kaplan-Meier curves for the context of this thesis. One aspect to remark on, is that the availability of the data is not synchronous, this means, a user may send a questionnaire today and the next one in 5 months, so in some cases is particularly problematic.

The left plot shows the remission probability, starting at 1.0 (indicating all subjects are initially in remission) and decreasing stepwise over time. This decline reflects the occurrence of events where subjects lose remission. By the end of the observation period, the remission probability is around 0.4, indicating that approximately 40

The right plot depicts the relapse probability, starting low and increasing over time, indicating that the probability of relapse grows as time progresses. The stepwise increases mark when relapse events occur, approaching a probability of 1.0 by the study's end, suggesting nearly all subjects have relapsed.

These curves complement each other: as remission probability decreases, relapse probability increases. They highlight significant time points where many events happen and provide a detailed view of the dynamics between remission and relapse over time within the studied population.

3.4 Machine learning algorithms

3.4.1 Logistic regression

Logistic regression is a statistical method widely used in classification tasks; particularly in the prediction of binary outputs. It models the relationship between the predictor variables (independent variables) and the log odds of the dependent variable.

Based on [28], the logistic regression model for predicting $P(Y = 1 \mid X)$ is given by:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

where:

- $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients,
- $X_1, X_2, \ldots, X_p$ are the predictor variables,
- $P(Y = 1 \mid X)$ is the probability of Y being 1 given the predictors X.

In the scenario of this thesis, even though the output variable is not strictly binary since bowel urgency may have different degrees between 0 and 3, it can certainly be binarized according to some criteria. For example, one may think about creating binary classifications dividing the previous labels into presence or absence of urgency, e.g. 0 as 0 and 1, 2, 3 as 1. Another approach is to create the new two labels based on the severity perceived by the user, being 0 the antiques 0 and 1, and 1 the previous 2 and 3. Logistic regression is a common method used in the analysis of patients with IBD, as seen is [29].

One of the many issues that come arise from classification problems is class imbalance, where records from one particular class outnumber the other(s) class, this condition tends to create biases when training a model which derives in a poor generalization performance; this thesis copes with that problem, in the scenario of IBD Tool, the vast majority of records are associated with 0 degree of urgency, therefore records with any other degree are outnumbered; the logistic regression implementation helps by decreasing the difference within the imbalance; and, according to the medical experts consulted during the carried out work, has more medical sense.

3.4.2 Random forest

Random forest is a popular learning method, heavily characterized by its robustness in classification tasks. It is a regression technique based on trees and randomization of predictors to achieve good accuracy values [30].

The random forest technique is widely used in similar problems to study different consequences of IBD. The robustness offered by the random forest makes it a preferred choice. For example, in [31] RF is used to predict responses to a biological drug in Crohn's disease based on clinical parameters. Similarly, in [32], RF is used to predict the result of endoscopies of patients in remission of ulcerative colitis with vedolizumab, a common biologic monoclonal antibody used as a medication for UC.

3.5 Data enrichment

One of the key parts for further analysis is the data enrichment, particularly, to the previous medical features already available in the clinical data of the patients, new features were added to the extracted data by the side of the Mauriziano hospital staff; these new features particularly address medical treatment.

Therapy	**Description**
CRP	Marker for inflammation, elevated during active disease.
Calprotectin	Protein indicating gut inflammation, measured in stool.
No treat	No treatment.
Oral 5ASA	Reduces colon inflammation, used for mild to moderate cases.
Topical treatment	Suppositories/enemas targeting rectum/lower colon inflammation.
Low GCS oral	Short-term anti-inflammatory for acute flare-ups.
Sys GCS oral	Systemic anti-inflammatory for moderate to severe cases.
IMM	Immunosuppressants to reduce colon inflammation and prevent flare-ups.
Anti TNF	Biologic blocking TNF, used for moderate to severe cases.
VDZ	Vedolizumab, blocks gut-specific integrin to reduce inflammation.
UST	Ustekinumab, targets IL-12 and IL-23 to reduce inflammation.
Anti-IL 23	Biologics targeting IL-23 for inflammation control.
JAKi	Janus kinase inhibitors reducing inflammation by blocking signaling pathways.

Table 3.2: Therapies for ulcerative colitis corresponding to the enriched data

Chapter 4

Results

4.1 Event-based approach

In the event-based approach, the database view's present form consists of each patient and his/her urgency counts at each time urgency is measured. In simpler words, in the event-based approach, the whole dataset available is analyzed, not associating the bowel urgency measured with the corresponding patient.

The main objective is to understand the interaction of all the variables available, at this point, the dataset includes new features associated with the medical treatment; analyzing the correlation of the novel information concerning the output feature, bowel urgency degree, and also, the creation of a machine learning model which can predict and classify well the degree of bowel urgency of a patient based, so that an idea of how the medical features can help in the prediction and even, in the prescription of medical treatments.

The Table 4.1 displays information about the variable's name, non-null count, and the data type of every feature that composes the dataset, which is stored as a comma-separated file. The total rows of available data are equal to 2768; as the Bowel Urgency feature tells us. As with any real-world problem, and particularly one that depends on the compliance of information submitted by users, there are some issues to solve with missing data. By looking at the amount of non-null data entries, some features are essentially useless since they are barely present, for example, Durat UST, Durat JAKi, Durat Anti iL 23, etc. Overall, some criteria need to be implemented to cope with this missing information.

The procedure to clean the dataset begins by filtering the rows e.g. features that contain at least 1000 non-null entries; by doing that, features with very low contribution of entries are cleaned up, the resulting dataset is described in Table 4.2.

Column	**Non-Null Count**	**Dtype**
Bowel Urgency	2768	int64
SCCAI	2768	int64
MIAH-UC	2768	float64
IBD-DISK	2665	float64
IBDQ	476	float64
PHQ9	437	float64
MMAS8	428	float64
LARS	216	float64
sex	2768	int64
age	2765	float64
height	2670	float64
weight	2734	float64
ageOfDiagnosis	2730	float64
pathologyDuration	2768	int64
storicheEIMS	2768	int64
attiveEIMS	2768	int64
stomia	2768	int64
CRP	1752	object
calprotectin	652	float64
No_treat	2736	float64
oral_5ASA	2736	float64
topical_treatm	2736	float64
low_GCS_oral	2736	float64
Durat_L_GCS	109	float64
Sys_GCS_oral	2736	float64
Durat_S_GCS	172	float64
IMM	2736	float64
AntiTNF	2736	float64
AntiTNF_opt	506	float64
Durat_antiTNF	506	float64
VDZ	2733	float64
VDZ_Opt	407	float64
Durat_VDZ	406	float64
UST	2736	float64
Durat_UST	27	float64
AntiIL23	2736	float64
Durat_AntiIL23	3	float64
JAKi	2736	float64
Durat_JAKi	122	float64
diseaseExtent	2723	float64

Table 4.1: Data Summary

Column	**Non-Null Count**	**Dtype**
Bowel Urgency	2768	int64
SCCAI	2768	int64
MIAH-UC	2768	float64
IBD-DISK	2665	float64
sex	2768	int64
age	2765	float64
height	2670	float64
weight	2734	float64
ageOfDiagnosis	2730	float64
pathologyDuration	2768	int64
storicheEIMS	2768	int64
attiveEIMS	2768	int64
stomia	2768	int64
No_treat	2736	float64
oral_5ASA	2736	float64
topical_treatm	2736	float64
low_GCS_oral	2736	float64
Sys_GCS_oral	2736	float64
IMM	2736	float64
AntiTNF	2736	float64
VDZ	2733	float64
UST	2736	float64
AntiIL23	2736	float64
JAKi	2736	float64
diseaseExtent	2723	float64

Table 4.2: Data summary after first filter

By looking at Table 4.2, the number of entries among all variables is relatively consistent, none of them goes below 2700, being the feature with fewer entries is the disease extent, with 2763 out of 2768 possible entries. Some features are not considered for prediction models, they are: height, weight, and age of diagnosis, after some discussion with physicians the conclusion is that there is no evidence about the correlation of them concerning bowel urgency, therefore they should not be considered in further steps. Table 4.3 displays the final set of features used for the prediction models, including the output variable, Bowel Urgency.

The querying explained above comprehends the procedure regarding the net number of entries per feature, however, there may be rows with several features missing, since this may be problematic for feeding the prediction models, the approach to handle them is by removing rows with 7 or more missing features. After applying this criterion, the total number of rows to work within the implementation of the prediction models is 2584.

Column	Non-Null Count	Dtype
Bowel Urgency	2768	int64
SCCAI	2768	int64
MIAH-UC	2768	float64
IBD-DISK	2665	float64
sex	2768	int64
age	2765	float64
pathologyDuration	2768	int64
storicheEIMS	2768	int64
attiveEIMS	2768	int64
stomia	2768	int64
No_treat	2736	float64
oral_5ASA	2736	float64
topical_treatm	2736	float64
low_GCS_oral	2736	float64
Sys_GCS_oral	2736	float64
IMM	2736	float64
AntiTNF	2736	float64
VDZ	2733	float64
UST	2736	float64
AntiIL23	2736	float64
JAKi	2736	float64
diseaseExtent	2723	float64

Table 4.3: Data Summary after second filter

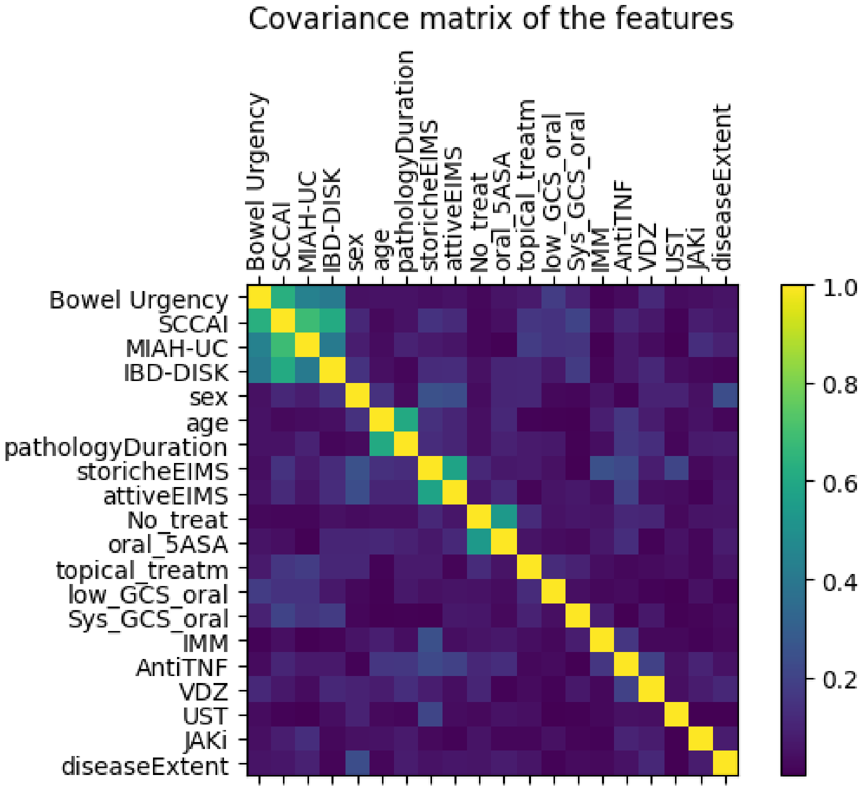


Figure 4.1: Covariance matrix of the filtered features represented as a heatmap

Figure 4.1 displays the covariance matrix as a heat map among the present variables, meanwhile Figure 4.2 represents the correlation coefficient as a line plot. Particularly speaking, bowel urgency is highly associated with the scores of the questionnaires (P-SCCAI, IBD-DISK AND MIAH-UC) and also with some other features regarding the medical therapies, especially the GCS treatment, the topical treatment and VDZ.

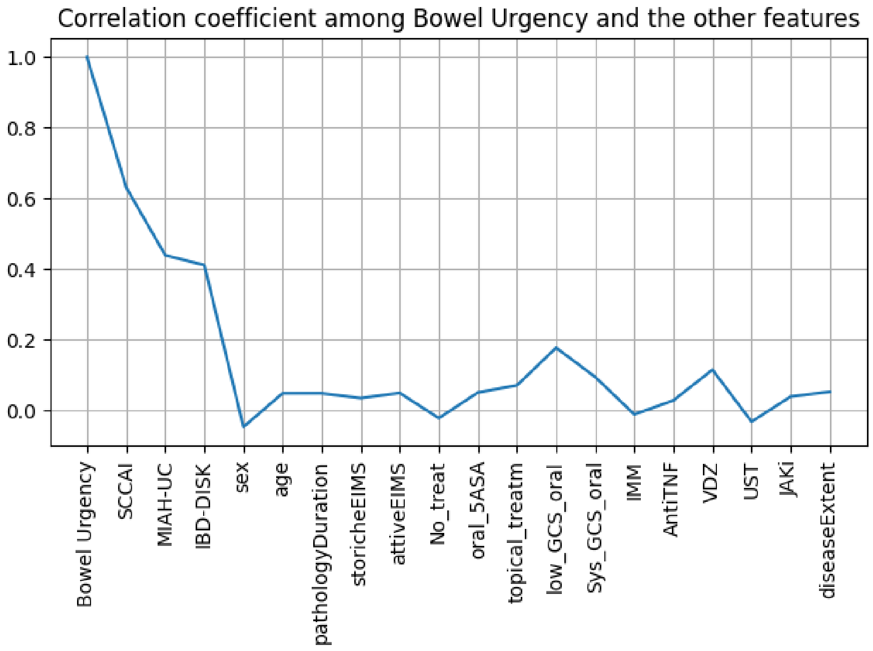


Figure 4.2: Correlation coefficient among variables concerning bowel urgency

4.1.1 Logistic regression models

The logistic regression algorithm is oriented into binary outcomes, however, bowel urgency degrees go from 0 to 3. Therefore, to implement a logistic regression model; some modifications need to be performed in the way the output variable is constructed.

- **First logistic regression model:**

$$Y = \begin{cases} 0 & \text{if } X \in \{0,1\} \\ 1 & \text{if } X \in \{2,3\} \end{cases}$$

In this scenario, the previous bowel urgency categories are mapped in a binary manner, the new level 0 corresponds to previous levels 0 and 1, and the new level 1 corresponds to previous levels 2 and 3. The data of 2584 rows is split into a training and test dataset in an 80 % and 20 % manner. The solver used is Limited-memory Broyden–Fletcher–Goldfarb–Shanno. Also using penalty L1 to see features that contribute some and adding some class weights.

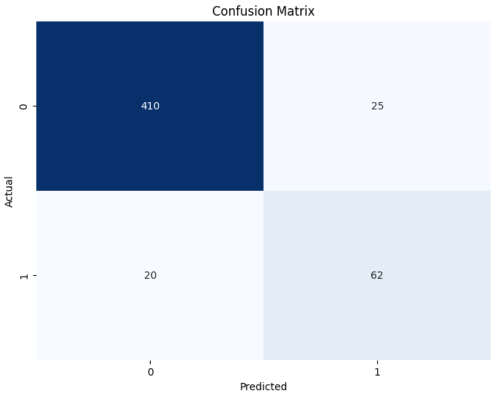


Figure 4.3: Confusion matrix associated with the first logistic regression model

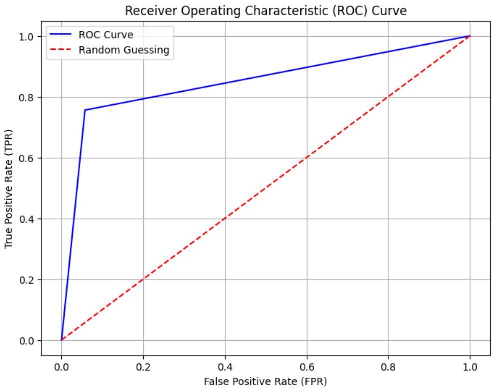


Figure 4.4: ROC curve associated with he first logistic regression model

Figure 4.3 displays the confusion matrix associated with the first logistic regression model implemented, particularly, the model classifies samples belonging to class 0 pretty well, with a precision close to 0.9425, however, on the other hand, samples belonging to class 1 are not predicted with the same performance, the precision in this scenario is around 0.756. The overall correctness is 0.913, and the precision of the model, which tells the prevalence of true positives, e.g. class 1 classified as class 1 is 0.713; regarding the trade-off between precision and recall the F1 score is 0.734, the performance of this model is quite acceptable but could be better, particularly in the precision and recall; by the precision side this is important to minimize false positives, which can lead to unnecessary stress and procedures, and the recall because is crucial in medical diagnostics to ensure critical conditions are not missed. Figure 4.4 shows the ROC curve associated to the model, which tells the ability of the model to differentiate between both classes. The summary of the metrics is shown at Table 4.4.

Metric	Value
Accuracy	0.913
Precision	0.713
Recall	0.756
F1-score	0.734
ROC-AUC score	0.849
Specificity	0.920

Table 4.4: First logistic regression model performance metrics, considering class 1

- **Second logistic regression model:**

$$Y = \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{if } X \in \{1, 2, 3\} \end{cases}$$

In this context, the previous bowel urgency categories have been redefined into a binary classification scheme: the new category 0 encompasses the original categories 0 and 1, while the new category 1 includes the original categories 2 and 3. The dataset, consisting of 2584 rows, is partitioned into training and test sets using an 80% and 20% split, respectively. The logistic regression model is configured with the liblinear solver, chosen for its suitability with L1 regularization and the ability to handle sparse matrices efficiently. Additionally, class weights are assigned to address potential class imbalance, with class 0 weighted as 1 and class 1 weighted as 1.5.

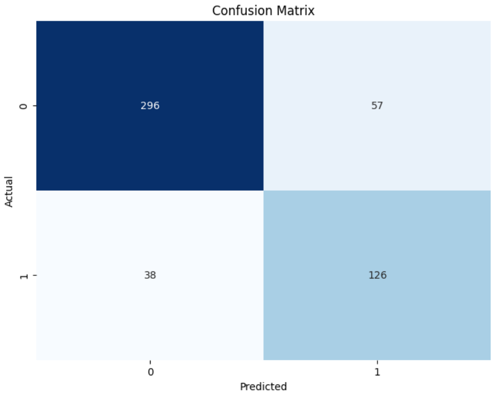


Figure 4.5: Confusion matrix associated with the second logistic regression model

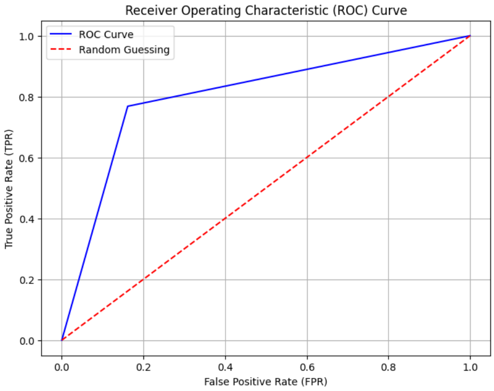


Figure 4.6: ROC curve associated with the second logistic regression model

Figure 4.5 displays the classification work performed by the second model using the logistic regression method. This model differs from the first one in constructing the new binary labels. One of the biggest concerns about the class imbalance is somewhat solved since now the ratio between the two classes is smaller; nevertheless, a weight was assigned to class 1 to make the statistical model consider and mind the associated characteristics that may represent samples belonging to class 1; the weight of 1.5 was decided after a grid search algorithm in which different parameters were iterating until finding the suitable ones. In this scenario, the model's overall accuracy is around 0.88 and the trade-off with the precision and recall, reflected in the F1 score is 0.755. This model still has room for improvement; therefore. Table 4.5 summarizes the performance metrics from the model.

Metric	**Value**
Accuracy	0.879
Precision	0.761
Recall	0.744
F1-score	0.755
ROC-AUC score	0.807
Specificity	0.831

Table 4.5: Second logistic regression model performance metrics, considering class 1

- **Third logistic regression model:**

$$Y = \begin{cases} 0 & \text{if } X \in \{0,1\} \\ 1 & \text{if } X \in \{2,3\} \end{cases}$$

The logistic regression model was trained using the liblinear solver, chosen for its efficiency in handling sparse data and suitable for binary classification tasks. The model was further enhanced with L2 regularization (penalty='l2') to penalize large coefficients and prevent overfitting. To address potential class imbalance in the dataset, class weights were incorporated during training. Specifically, class 0 was weighted as 1 meanwhile class 1 was weighted as 2.5.

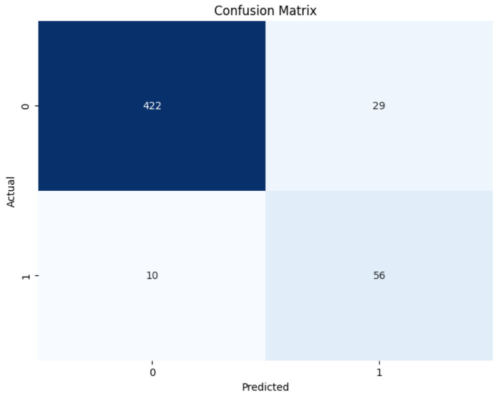


Figure 4.7: Confusion matrix associated with the third logistic regression model

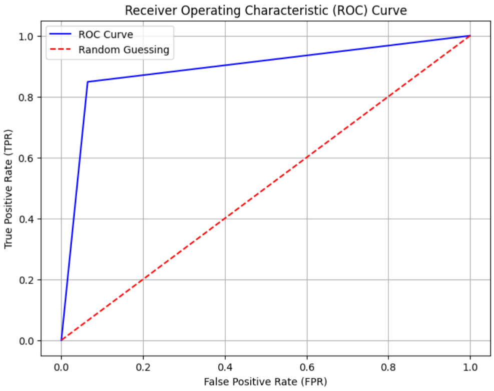


Figure 4.8: ROC curve associated with the third logistic regression model

This model represents the best option using the logistic regression algorithm. Figure 4.7 displays the confusion matrix obtained after executing the model. The overall accuracy of the model is 0.923, which is a value typically accepted as a good performance indicator. The precision, on the other hand, is in the order of 0.659, however, the recall is 0.848, particularly the biggest interest is towards the recall since it is very important to classify correctly samples that actually belong to class 1, even though have high precision is important as well, the context of this project, being a telemonitoring tool gives some flexibility on this metric. This model differentiated between classes the better, as seen with the ROC score of 0.892.

Metric	**Value**
Accuracy	0.923
Precision	0.659
Recall	0.849
F1-score	0.742
ROC-AUC score	0.892
Specificity	0.936

Table 4.6: Third logistic regression model performance metrics, considering class 1

Without any doubt, the feature with the biggest information contribution to the classification models are the three questionnaires, P-SCCAI, MIAH-UC and IBD-DISK, the covariance matrix shown in Figure 4.1 confirms the statement;

therefore, an interesting insight is a statistical distribution and distinction about some features and their classes, to see how discriminatory they can be for the classification.

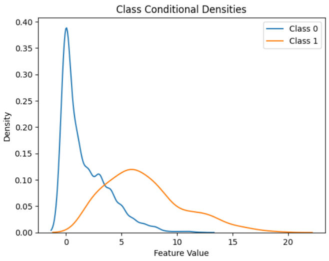


Figure 4.9: Class-conditional density model 1: P-SCCAI

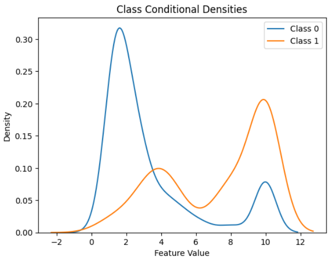


Figure 4.10: Class-conditional density model 1: MIAH-UC

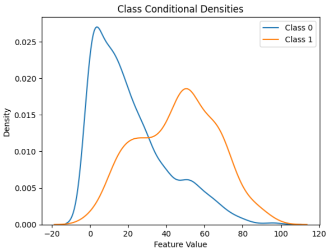


Figure 4.11: Class-conditional density model 1: IBD-DISK

Figures 4.9, 4.10 and 4.11 display the class-conditional densities comparing the statistical distribution of classes 0 and 1 from the first logistic regression for the three questionnaires. P-SCCAI shows moderate separation between the classes with significant overlap. MIAH-UC demonstrates better separation with less overlap, indicating it might be more effective in distinguishing between the two classes based on the feature. IBD-DISK shows considerable overlap with complex distributions, suggesting it might have difficulty distinguishing between the classes based on the feature.

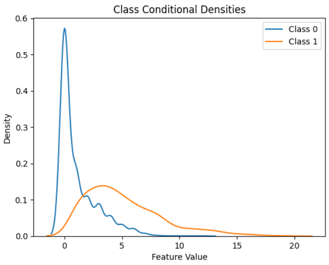


Figure 4.12: Class-conditional density model 2: P-SCCAI

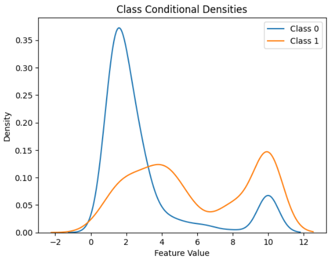


Figure 4.13: Class-conditional density model 2: MIAH-UC

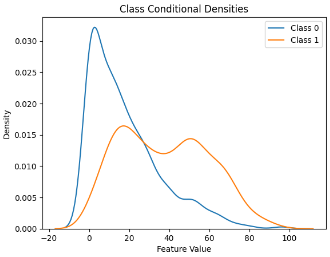


Figure 4.14: Class-conditional density model 2: IBD-DISK

Figures 4.12, 4.13 and 4.14 on the other hand, show the class-conditional densities of these questionnaires in the second model. The behavior for the P-SCCAI and MIAH-UC questionnaires is quite similar to the approach developed in the first model, however, the IBD-DISK questionnaire is a little bit diverse, indicating that it may not be as informative as in the previous model to discriminate between classes; in other words, it has a bigger overlap.

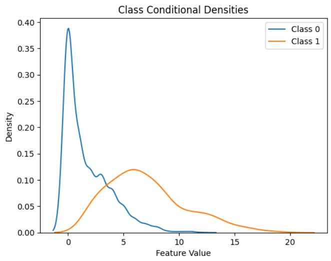


Figure 4.15: Class-conditional density model 3: P-SCCAI

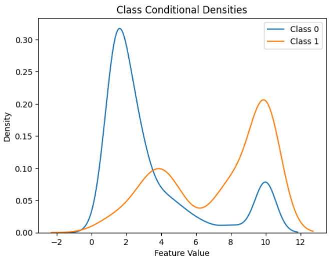


Figure 4.16: Class-conditional density model 3: MIAH-UC

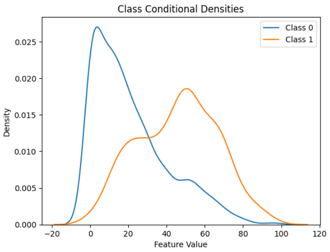


Figure 4.17: Class-conditional density model 3: IBD-DISK

Regarding the third logistic regression model, Figures 4.15, 4.16 and 4.17 display the conditional densities for the third logistic regression model. The behavior is quite similar to the second model except for the fact that IBD-DISK questionnaire is quite more informative and therefore, helpful for the classifier to discriminate properly between classes. As a matter of fact, the third model possesses the best performance among the three.

4.1.2 Random forest models

The random forest model does not restrict the binarization of the classes just as the logistic regression does, therefore, it is possible to implement a model with the objective to predict all four degrees of bowel urgency degree.

- **First random forest model**

$$Y = \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{if } X = 1 \\ 2 & \text{if } X = 2 \\ 3 & \text{if } X = 3 \end{cases}$$

The first model created using random forest did not map the original bowel urgency degrees. However, since the problem of the class imbalance is known (where the samples associated with level 0 outnumber the rest), in addition to a particularity noted during this work in the classification of samples belonging to class 1 and 2, some weights were designated to each class in order to fit the model and obtain an acceptable performance. Classes 1 and 2 are weighted as 2, meanwhile, classes 0 and 3, the ones the classifier performs the best are kept as 1.

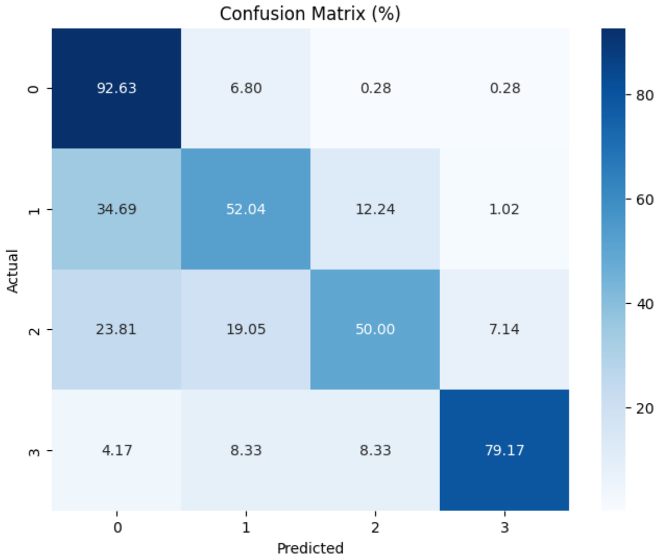


Figure 4.18: Confusion matrix displaying classification and misclassification rate of first RF model.

Figure 4.18 displays the confusion matrix of the first classifier expressed as percentages. Even though the implementation of some strategies to sort out the issue with the classification of classes 1 and 2, the performance is not quite well; unlike classes 0 and 3, essentially, the extreme classes, where the classifier performs a better job. Nevertheless, according to the evaluation of physicians in the head of IBD Tool, predicting the four levels of urgency may not be an appropriate procedure; a further model addresses this issue. The overall accuracy of this model is 0.809. The performance metrics are summarized in Table 4.7.

	Precision	Recall	F1-score	Support
0	0.880	0.930	0.900	353
1	0.600	0.520	0.560	98
2	0.580	0.500	0.540	42
3	0.790	0.790	0.790	24
Accuracy		0.814		
Macro avg	0.710	0.680	0.700	517
Weighted avg	0.800	0.810	0.800	517

Table 4.7: Performance metrics of the first random forest model

- **Second random forest model**

$$Y = \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{if } X = 1 \\ 2 & \text{if } X = 2 \\ 3 & \text{if } X = 3 \end{cases}$$

The second model in which the random forest model is implemented is done to check whether there is any difference in the quality and performance of the generalization when utilizing a different set of variables, unlike all of them as in the first RF model.

Figure 4.19, displays the correlation coefficient between the output variable and the rest. Based on this, selecting the set of features with the highest coefficients may affect somehow the performance of the model, therefore a new selection of variables is carried out. In this model, the sub-set of variables is created by grouping: P-SCCAI, MIAH-UC, IBD-DISK, pathology duration, age, disease extent, VDZ and topical treatment.

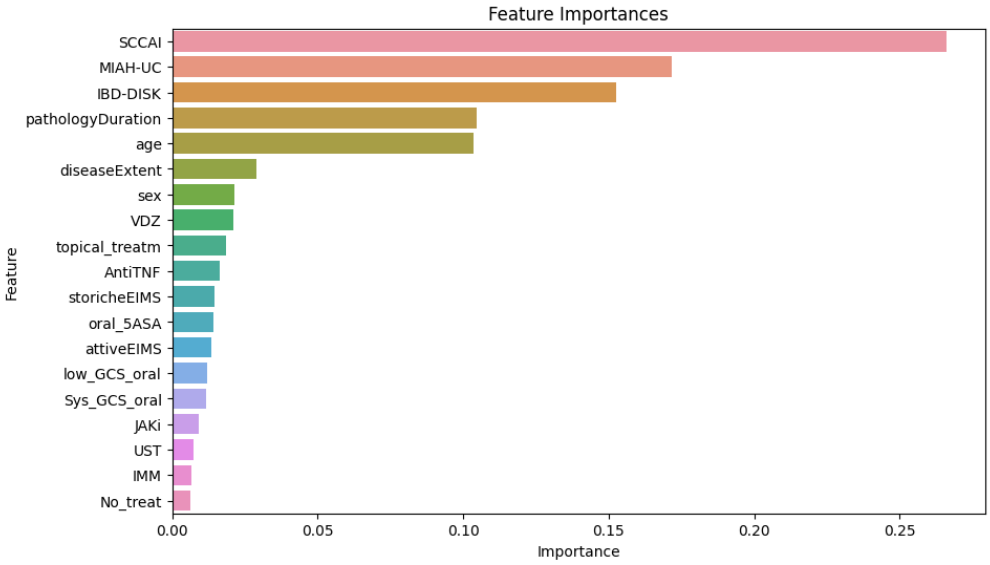


Figure 4.19: Correlation coefficient among features with respect to bowel urgency.

The result of this new configuration is summarized in the confusion matrix visible in Figure 4.20, the performance metrics are summarized in Table 4.8. The overall performance of the classifier is better, but the changes are rather small, the success in the prediction of classes 1 and 2 indeed improved by a small percentage; however, the overall accuracy of the model is of 0.83, therefore it could be improved. The way to go is to generalize and map classes, just as suggested by physicians.

	Precision	Recall	F1-score	Support
0	0.910	0.920	0.910	353
1	0.630	0.640	0.640	98
2	0.620	0.550	0.580	42
3	0.860	0.750	0.800	24
Accuracy		0.839		
Macro avg	0.750	0.720	0.730	517
Weighted avg	0.830	0.830	0.830	517

Table 4.8: Performance metrics of the second random forest model

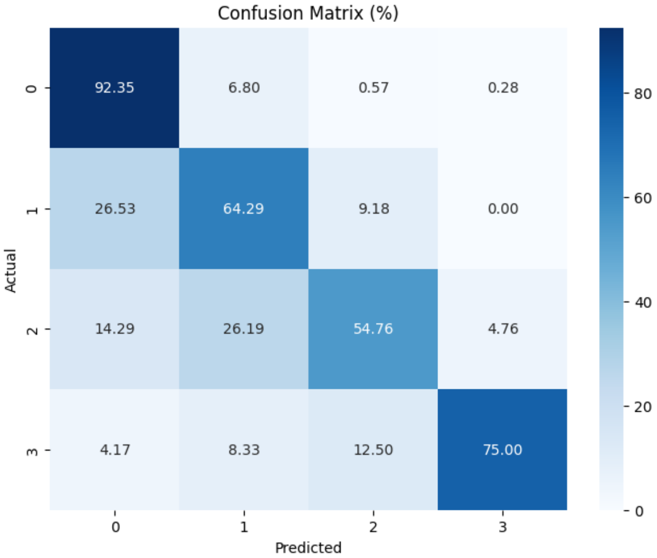


Figure 4.20: Confusion matrix displaying classification and misclassification rate of second RF model with a subset of features.

Figure 4.21 displays the ROC curve from the different classes of the second model developed with random forest. Once again, the best predictions are obtained in the generalization of classes 0 and 3, meanwhile, 1 and 2 are particularly problematic. One of the main reasons behind this is the similarity between the symptoms developed by these two classes; it is important to mention that the filling of the questionnaires is a procedure done by the patients and therefore some components of subjective thinking may come across.

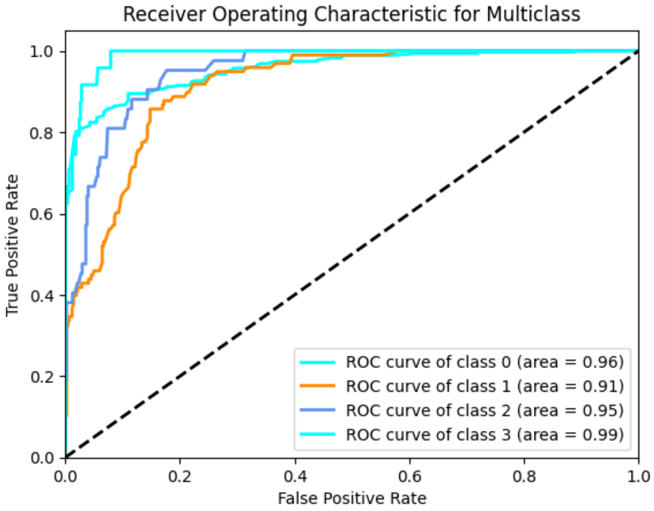


Figure 4.21: ROC curve of the second model of Random Forest.

- **Third random forest model**

$$Y = \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{if } X \in \{1, 2, 3\} \end{cases}$$

This model is essentially constructed in a detection fashion, which aims to predict whether there will be at least some degree of bowel urgency or not. The first step in its construction is the mapping of the classes.

The core of this classifier is the implementation of a parameter grid for its tuning, some key aspects to cover are the complexity of the regularization and the max depth of the decision trees. These parameters collectively influence model performance by adjusting the trade-off between bias and variance, aiming to optimize predictive accuracy for the given dataset and task.

	Precision	Recall	F1-score	Support
0	0.930	0.910	0.920	353
1	0.880	0.860	0.870	164
Accuracy		0.918		
Macro avg	0.900	0.890	0.900	517
Weighted avg	0.920	0.920	0.920	517

Table 4.9: Performance metrics of the third random forest model

Overall, this model is the best one, it has a global accuracy of 0.92. Figure

4.22 displays the confusion matrix of the classifier expressed as percentages. Table 4.9 summarizes the performance metrics obtained by the model and Figure 4.23 portrays the associated ROC curve.

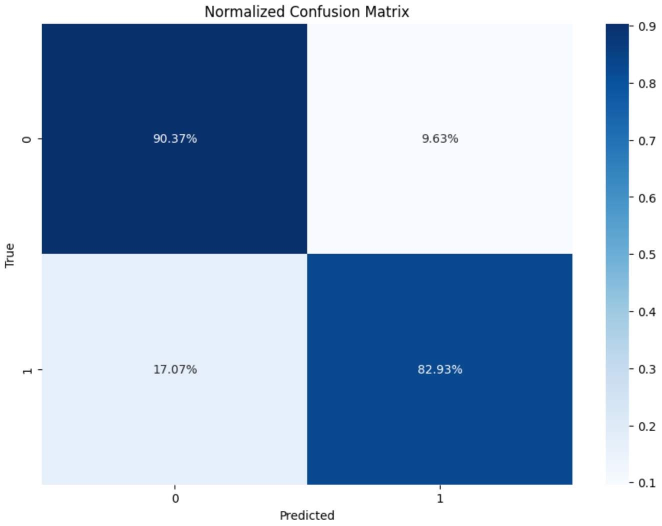


Figure 4.22: Confusion matrix displaying classification and misclassification rate of third RF model with a subset of features.

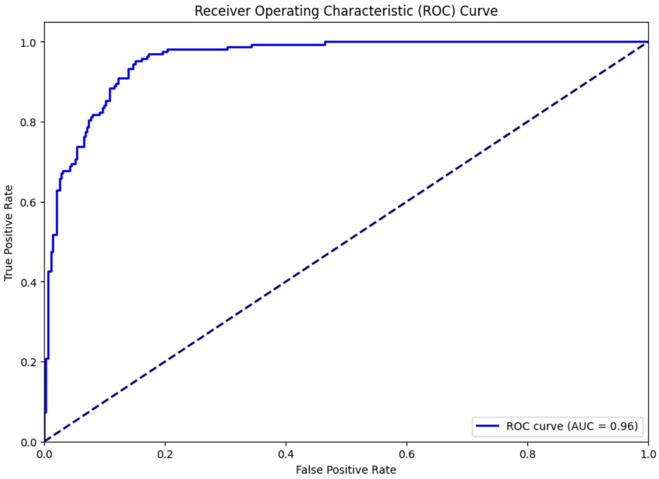


Figure 4.23: ROC curve of the third model of Random Forest.

4.2 Patient-based approach

The patient-based approach is certainly an approach where the main focus is on the samples associated with each patient. The main objective is to understand and analyze which features contribute the most to the remission or relapse of the patients; even though it is already known that the most informative features can be drawn from the covariance matrix heat map in Figure 4.1 and from the bar plot in Figure 4.19, another thing of interest is to recognize which kind of variables associated to the pharmaceutical therapies contributed the most in the changes of the bowel urgency degree for each patient.

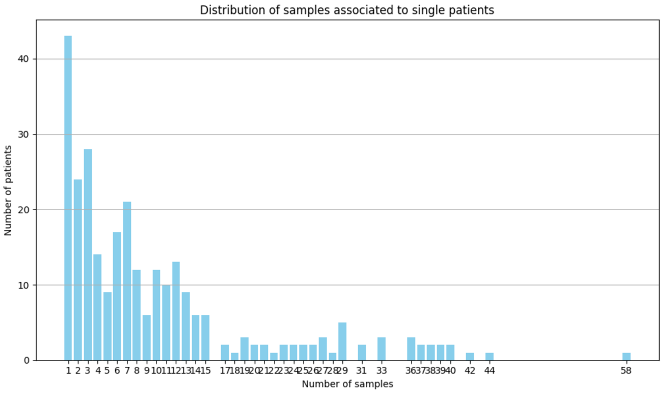


Figure 4.24: Distribution of the samples available per patient.

As Figure 4.24 and Table 4.11 tell, the vast majority of patients contribute only with one entry to the study. The total number of individual patients who registered at least one sample is of 277.

In order to analyze the changes registered in the output variable multiple samples need to be compared by a user, at least 2 of them, as a minimum, therefore, 43 users are discarded.

The dataset is organized in a sort of time series, where there are the entries associated to the user, ideally, there should be one sample per user monthly since their register; however, since this is a tool where the compliance by the user side is required, the temporality is not guaranteed between samples. An exploratory data analysis can give hints about the behavior of the evolution of bowel urgency.

Figures 4.25, 4.26 and 4.27 display the behavior of the output variable for three randomly chosen patients during the available timestamps. These plots are particularly useful whenever a graphical portrayal of the values of the variable of interest.

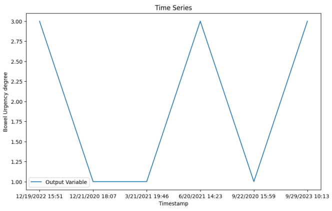


Figure 4.25: EDA for a single patient

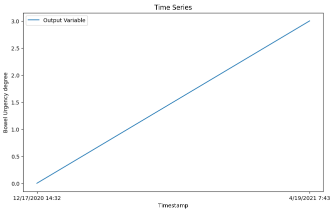


Figure 4.26: EDA for a single patient

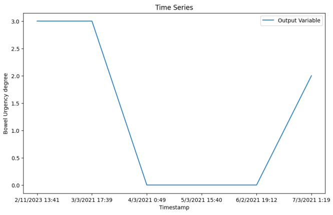


Figure 4.27: EDA for a single patient

The idea is to iterate through the entries associated with every patient and create a simple classification model, since the interest points toward the checking of the features and their importance, no mapping of the output classes is needed, therefore, the selected statistical method is random forest. An important remark on the construction of the random forest models is that, given the relatively small amount of available samples per patient, it is decided to evaluate the feature importance only on those who register at least 15 entries to the study, patients with lower contributions may not be suitable for this analysis and may introduce undesired biases.

The results found just confirm the previous hypothesis about the importance and weight of the questionnaires filled by the patients in the prediction of bowel urgency degree.

Feature	**Importance**
SCCAI	0.268
IBD-DISK	0.230
MIAH-UC	0.214
topical_treatm	0.0239
AntiTNF	0.0125
VDZ	0.0103

Table 4.10: Top feature importance scores

Number of patients	**Number of samples by patient**
43	1
28	3
24	2
21	7
17	6
14	4
13	12
12	8
12	10
10	11
9	5
9	13
6	15
6	14
6	9
5	29
3	27
3	33
3	36
3	19
2	17
2	20
2	21
2	23
2	24
2	25
2	26
2	31
2	37
2	38
2	39
2	40
1	18
1	44
1	22
1	28
1	42
1	58

Table 4.11: Distribution of the available samples per patient.

Chapter 5

Conclusion and future works

After the work of this thesis, a relationship between the telemonitoring functionality of the IBD Tool and the analysis of bowel urgency in patients diagnosed with UC has been explained. The underlying importance of working around the bowel urgency issue is critical significance to patients affected by the disease; it affects the quality of life from different points; objectives such as inflammation of the gastrointestinal tract, rectal bleeding, etc. but also from a psychological side, bowel urgency tends to decrease confidence and self-perception, it also affects the normal development of daily activities such as social interactions, study and work.

The clear causes of IBD are still a topic of research by the scientific community, and efforts from multiple disciplines gather to improve the quality of life of patients who suffer from these diseases. The telemedicine field is not an exception, one clear case is the IBD Tool, a web application built by engineers, data analysts and physicians with the objective of implement a telemonitoring solution to keep track and keep an eye on the evolution of the IBD in the patients. The core functionality of the IBD Tool is the questionnaires sent to the patients to be filled, they are meant to collect information directly and indirectly related to the symptoms associated with IBD going from objective questions such as the absence or presence of clear medical indicators and questions open to the interpretation of the patient, for example about his social interactions. The scope of this thesis is under three particular questionnaires, the P-SCCAI, the MIAH-UC, and the IBD-DISK. They are aimed to collect information about the disease activity index, the colonic activity, and about the quality of life, respectively.

The main objective was to establish relationships among a set of variables available in the IBD Tool regarding medical information of the patients diagnosed with UC such as therapies, diagnosis, disease location, score of the questionnaires, etc. and the registered degree of bowel urgency. To study the dependencies and relationships between the gathered information. The approach from the machine learning part consisted of the creation of classification models to properly predict the urgency level based on the information about the questionnaires and pharmaceutical therapies; precisely this work used logistic regression and random forest.

The best model obtained implemented random forest in a detection fashion, in essence, mapping the original classes [0, 1, 2, 3] into a total absence of bowl urgency and presence. The obtained classifier predicts the bowel urgency based mostly on the following features: P-SCCAI score, MIAH-UC score, IBD-DISK score, disease extent and the following pharmaceutical therapies: VDZ, topical treatment and AntiTNF. Nevertheless, the most informative features are the questionnaires. The model has an overall accuracy of 92%.

The following are some comments about future works and the relationship between telemedicine and ICT.

- In the scenario of IBD Tool, the compliance of the patients when filling out questionnaires is crucial, and it is one of the aspects that has to be stressed in the telemedicine field. The awareness and encouragement of patients to use and trust telemedicine services is vital to improve the quality of care and the new developments.

- Telemedicine is a fast-growing field, with new updates and research ongoing, the future works associated with the work carried out during this thesis are the enhancement of telemonitoring solutions oriented into IBD.

- Since the results demonstrate an existing relationship between the scores of the questionnaires and bowel urgency, a future feature of the IBD Tool could be the automatic prediction of a possible remission on relapse based on the methods exposed in this thesis, which will alert the physicians about the state of the patient.

Bibliography

[1] World Health Organization. *A Health Telematics Policy.* WHO, 1998 (cit. on p. 1).

[2] M. Waller and C. Stotler. «Telemedicine: a Primer». In: *Curr Allergy Asthma Rep* 18.10 (2018). DOI: 10.1007/s11882-018-0808-4 (cit. on p. 1).

[3] L. Rosta et al. «Telemedicine for diabetes management during COVID-19». In: *Front Endocrinol* 14 (2023). DOI: 10.3389/fendo.2023.1129793 (cit. on p. 1).

[4] B.C. Gates, A.V. Venegas-Vera, and E.V. Lerma. «Utility of telemedicine in the COVID-19 era». In: *RCM* 21.4 (2020). DOI: 10.31083/j.rcm.2020.04.188 (cit. on p. 2).

[5] W. Strober, I. Fuss, and P. Mannon. «The fundamental basis of inflammatory bowel disease». In: *J Clin Invest* 117.3 (2007). DOI: 10.1172/jci30587 (cit. on pp. 2, 3).

[6] IBD Relief. *IBD Relief: What is IBD?* https://www.ibdrelief.com/learn/what-is-ibd. 2023 (cit. on p. 3).

[7] J. Torres, S. Mehandru, J.F. Colombel, and L. Peyrin-Biroulet. «Crohn's disease». In: *Lancet* 389.10080 (2017). DOI: 10.1016/s0140-6736(16)31711-1 (cit. on p. 4).

[8] Harvard Health Publishing. *Crohn's Disease: A to Z.* 2023. URL: https://www.health.harvard.edu/a_to_z/crohns-disease-a-to-z (cit. on p. 4).

[9] D. Turner, C.M. Walsh, A.H. Steinhart, and A.M. Griffiths. «Response to Corticosteroids in Severe Ulcerative Colitis: A Systematic Review». In: *Clin Gastroenterol Hepatol* 5.1 (2007). DOI: 10.1016/j.cgh.2006.09.033 (cit. on p. 4).

[10] M. Gajendran, P. Loganathan, G. Jimenez, A.P. Catinella, N. Ng, C. Umapathy, N. Ziade, and J.G. Hashash. «A comprehensive review and update on ulcerative colitis». In: *Dis Mon* 65.12 (2019). DOI: 10.1016/j.disamonth.2019.02.004 (cit. on p. 5).

[11] J.D. Feuerstein and A.S. Cheifetz. «Ulcerative Colitis». In: *Mayo Clin Proc* 89.11 (2014). DOI: 10.1016/j.mayocp.2014.07.002 (cit. on p. 5).

[12] AboutKidsHealth. *Ulcerative colitis in children and teens.* https://www.aboutkidshealth.ca/Article?contentid=924&language=English. 2023 (cit. on p. 5).

[13] M.C. Dubinsky et al. «Impact of Bowel Urgency on Quality of Life in UC». In: *Crohn's Colitis 360* 4.3 (2022). DOI: 10.1093/crocol/otac016 (cit. on p. 6).

[14] L. Peyrin-Biroulet et al. «Selecting Therapeutic Targets in IBD (STRIDE): Determining Therapeutic Goals for Treat-to-Target». In: *Am J Gastroenterol* 110.9 (2015). DOI: 10.1038/ajg.2015.233 (cit. on p. 6).

[15] F. Ferretti, R. Cannatelli, M.C. Monico, G. Maconi, and S. Ardizzone. «Pharmacotherapeutic Options for UC». In: *J Clin Med* 11.9 (2022). DOI: 10.3390/jcm11092302 (cit. on p. 6).

[16] E. Troncone and G. Monteleone. «Safety of non-biological treatments in UC». In: *Expert Opin Drug Saf* 16.7 (2017). DOI: 10.1080/14740338.2017.1340936 (cit. on p. 6).

[17] H. Akiho. «Promising biological therapies for UC». In: *World J Gastrointest Pathophysiol* 6.4 (2015). DOI: 10.4291/wjgp.v6.i4.219 (cit. on p. 6).

[18] P. Andersson and J.D. Söderholm. «Surgery in UC: Indication and Timing». In: *Dig Dis* 27.3 (2009). DOI: 10.1159/000228570 (cit. on p. 7).

[19] J. Stavsky and R. Maitra. «Diet and Exercise in UC: Metabolic Mechanism». In: *Nutr Metab Insights* 12 (2019). DOI: 10.1177/1178638819834526 (cit. on p. 7).

[20] myIBDcoach Project Team. *MyIBDcoach Project: Value Based Care for IBD with Telemedicine.* https://vbhcprize.com/myibdcoach-project/. Accessed: 2024-06-14. 2024 (cit. on p. 8).

[21] MongoDB. *MongoDB Document Databases.* https://www.mongodb.com/document-databases. Accessed: 13 11, 2023 (cit. on pp. 11, 12).

[22] Programiz. *Database introduction.* https://www.programiz.com/sql/database-introduction. Accessed: 13 11, 2023 (cit. on p. 12).

[23] C. Lia. «Functionalities development for IBD Tool, an italian web-app for Inflammatory Bowel Disease monitoring». Rel. Carla Fabiana Chiasserini, Guido Pagana. Master's Thesis. Turin, Italy: Politecnico di Torino, 2021 (cit. on pp. 13, 20).

[24] F. Bennebroek Evertsz', P.T. Nieuwkerk, P.C.F. Stokkers, C.Y. Ponsioen, C.L.H. Bockting, R. Sanderman, and M.A.G. Sprangers. «The Patient Simple Clinical Colitis Activity Index (P-SCCAI) can detect UC disease activity in remission». In: *J Crohn Colitis* 7.11 (2013). DOI: 10.1016/j.crohns.2012.11.007 (cit. on pp. 13, 15, 16).

[25] F. Boschi, V. Figini, R. Baroughi, and G. Pagana. «Trasferimento dell' architettura del servizio di telemonitoraggio precedentemente sviluppato (IBD Tool)». In: (2023) (cit. on p. 17).

[26] D. Damino. «IBD Tool: Enabling Continuous Data Export and Improvement of User Experience». Rel. Carla Fabiana Chiasserini, Guido Pagana. Master's Thesis. Turin, Italy: Politecnico di Torino, 2022 (cit. on pp. 18–20).

[27] N. Ishida et al. «Predicting UC Relapse in Clinical Remission With Fecal Immunochemical Occult Blood Test or Prostaglandin E-Major Urinary Metabolite». In: *Clin Transl Gastroenterol* 13.7 (2022). DOI: 10.14309/ctg.0000000000000501 (cit. on p. 35).

[28] K. Murphy. *Machine Learning: A Probabilistic Perspective.* 1st. ISBN-13: 978-0262018029. MIT Press, 2012 (cit. on p. 36).

[29] E. Safroneeva et al. «Systematic analysis of factors associated with progression and regression of ulcerative colitis in 918 patients». In: *Alimentary Pharmacology; Therapeutics* 42.5 (July 2015), pp. 540–548. ISSN: 1365-2036. DOI: 10.1111/apt.13307 (cit. on p. 37).

[30] S. Rigatti. «Random Forest». In: *Journal of Insurance Medicine* 47.1 (Jan. 2017), pp. 31–39. DOI: 10.17849/insm-47-01-31-39.1 (cit. on p. 37).

[31] Y. Li, J. Pan, N. Zhou, D. Fu, G. Lian, J. Yi, Y. Peng, and X. Liu. «A random forest model predicts responses to infliximab in Crohn's disease based on clinical and serological parameters». In: *Scandinavian Journal of Gastroenterology* 56.9 (July 2021), pp. 1030–1039. DOI: 10.1080/00365521.2021.1939411 (cit. on p. 37).

[32] A.K. Waljee, B. Liu, K. Sauder, J. Zhu, S.M. Govani, R.W. Stidham, and P.D.R. Higgins. «Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis». In: *Alimentary Pharmacology; Therapeutics* 6 (2018), pp. 763–772. DOI: 10.1111/apt.14510 (cit. on p. 37).