BACHELOR' S THESIS / BIOMEDICAL ENGINEERING 2022

Use of Text Mining techniques for the selection of cohorts in a clinical trial:

Classifying compliance with the selection criteria for a clinical trial by analyzing patient medical records.

Ricard Cambray Alvarez

upf. Universitat Pompeu Fabra Barcelona

Use of Text Mining techniques for the selection of cohorts in a clinical trial:

Classifying compliance with the selection criteria for a clinical trial by analyzing patient medical records.

Ricard Cambray Alvarez

Bachelor's Thesis UPF 2020/2021

Thesis Supervisor(s):

Horacio Saggion, (Full Professor in Computer Science and Artificial Intelligence at Universitat Pompeu Fabra)

Universitat Pompeu Fabra Barcelona

Acknowledgments

I would like to express my sincere gratitude to all those people who have collaborated in any stage of the project. Also to those people who have supported me both academically and personally. First of all, I would like to appreciate the help, advice and collaboration of the project supervisor, Horacio Saggion, in all stages of the project. Also, I would like to thank the support I have received on a personal level from my environment, family and friends.

Summary/Abstract

.

Introduction Nowadays, most of the data with which clinicians work are unstructured such as texts. In this large volume of unstructured data lies the need to create analytical procedures in order to maximize the value extraction. As a combination of computational linguistics and Machine Learning techniques arises Text Mining and Natural Language Processing (NLP) techniques.

Objective: One of the most tedious and time-consuming tasks in clinical trials is the subject selection process. The information of the patients susceptible to inclusion in the study must be consulted manually to check if they meet the defined selection criteria. The project aimed to build an automatic subject selection system for clinical trials from longitudinal patient medical records.

Materials and Methods: Starting from a set of clinical histories annotated according to whether the patient meets or does not meet 13 selection criteria, several preprocessing tasks related to NLP techniques were applied. An hybrid approach combining Machine Learning (ML) models and Rule-Based models were studied for the classification task setting the majority classifier algorithm as the Baseline for comparison.

Results: 10 of the selection criteria achieved best results when applying ML models after a preprocessing stage, the remaining selection criteria achieved better classification performance when applying the Rule-Based approach. The overall micro F1 score of the proposed model achieved a 0,8574 value.

Conclusion: This study concludes that the proposed hybrid approach offers the possibility to develop a useful tool for the automatic selection of patients for a clinical trial cohort.

Keywords

Text Mining, NLP, Clinical Trial, Cohort Selection, Machine Learning, micro F1, Rule-based model, Hybrid approach.

Preface or prologue

The final degree project presented below is entitled "Use of Text Mining techniques for the selection of cohorts in a clinical trial: Classifying patients into potential subjects for a clinical trial according to selection criteria by analyzing medical records".

My personal interests regarding artificial intelligence and its numerous applications in the biomedical field have been increasing in the last years of my university career.

The implementation of artificial intelligence in biomedical data analysis has generally been studied in biomedical imaging. This has provided a significant number of studies and techniques that have improved the performance of many clinical medicine processes: diagnosis, choice of treatment, evolution of a disease, planning of surgical interventions, etc.

However, after consulting some experts, including Horacio Saggion (expert in computational linguistics and professor at UPF) as well as several professionals involved in preventive medicine, I decided to consider the possibility of applying artificial intelligence techniques for automatic text data analysis (Text Mining). In this way, to be able to study the great potential of these techniques and tools in the automation of processes related to clinical medicine.

This thesis has given me the opportunity to delve into a branch of artificial intelligence that I was completely unaware of until the beginning of the project. In addition, it has motivated me to continue learning about the concepts that I have learned during the development of the project.

Index

List of Figures

List of Tables

1 Introduction

1.1 Importance of data analysis in clinical medicine

Nowadays, data analysis and information extraction have an important role in the improvement of many professional fields. Data has led the development of companies, has allowed the optimization of industrial processes, has accelerated business decision-making and has improved the effectiveness of some studies and research in the biomedical field.

Data analysis have an influence on practically all the fields of biomedicine: genomics, hospital management, administrative field, epidemiology, clinical medicine [1], etc. There is a lot of data related to medicine, from personal data of each patient to those derived from clinical practice (images, reports, etc.). The creation of intelligent alert systems, prediction of hospital readmissions, profiling and detection of over-visited patients, prediction of healthcare spending and optimization of resources, personalization of treatment and control of epidemics are some of the most frequent applications of the analysis of data in hospitals.

All of the above indicates that data analysis has great potential to achieve a more effective medicine: a personalized, participatory, preventive, predictive and population-based medicine.

1.2 Patient's clinical data

As previously mentioned, there are different types of hospital data related to patients: administrative and general patient data and medical history data. The clinical history [2] of a patient comprises the set of documents related to the care processes of each patient. It is defined as the legal document that collects all data related to health and health services and procedures provided to each patient, in order to provide adequate and effective medical assistance. This information allows an appropriate communication of the patient's health status across the main stakeholders of the medical profession. Helping in the proper diagnosis, influencing the treatment decision and becoming the control tool of the patient's evolution. As it can be seen in Table 1, different descriptive data can be found in the medical records.

Data can be divided into two major categories. Structured and unstructured data. Structured data are organized into fields and columns in relational databases. Those type of data that allow a quick and efficient research, write and manipulation of data. On the contrary, unstructured data are those that cannot be processed or analyzed with the conventional methods. Currently, it is estimated that around 80% of the data generated in the context of clinical medicine is unstructured [3]. For instance, the health care system constantly generates a large volume of unstructured data. From data generated by biomedical machines (images, graphics, sensors logs...) to textual data generated by doctors themselves (conversation with the patient, hospital discharge etc.). Most of the data and information regarding patients to which doctors have access is in text format. The documents of the clinical history of a patient that are shown Table 1 are an example of this. Another type of data that health professionals usually work with are images, but even from the images, there

Data	Information it gives
Patient identifying data	Data referred to the identification of the patient (name, id, address, identifying image etc.)
Anamnesis, physical examination	Information gleaned through a conversation with the patient and the examinations maneuvers
Urgency reports	Information collected during the patient's emergency visit
Chronological clinical evolution	Chronological Information of the clinical evolution
Medical orders issued	Recipes, treatment and cares to be followed by the patient
Complementary examinations	Complementary examinations requested by the medical-health personnel (images, ECG, EEG...)
Consultation sheet	Clinical form used in communication between Primary Care and the second level of care
Informed consent	In which you give permission to carry out treatment or surgical interventions.
Surgical report	Information given by the surgical staff in an intervention
Nursing report	General nursing care information
Clinical discharge report	Summary report of the hospitalized patient when discharged

Table 1: Document and data provided by the clinical history of a patient and its description

is a textual report that summarizes and describes them. In addition, the texts of the medical records are unstructured at different levels as it can be seen in Figure 1.

In this large volume of unstructured data lies the need to create analytical procedures in order to maximize the value extraction of the information that can be hidden in them.

1.3 Subjects' selection in Clinical Trials

The increase in life expectancy due to the aging of the population, the chronicity of diseases and the consequent increase in the frequency of medical care needs and the health spending are going to be some of the most important challenges for health services in the near future. To deal with this problem, health institutions have begun to promote prevention health policies, control of risk factors and early detection of diseases. The medical specialty responsible for the prevention and early detection of diseases, identification of risk factors and epidemiological studies is Preventive

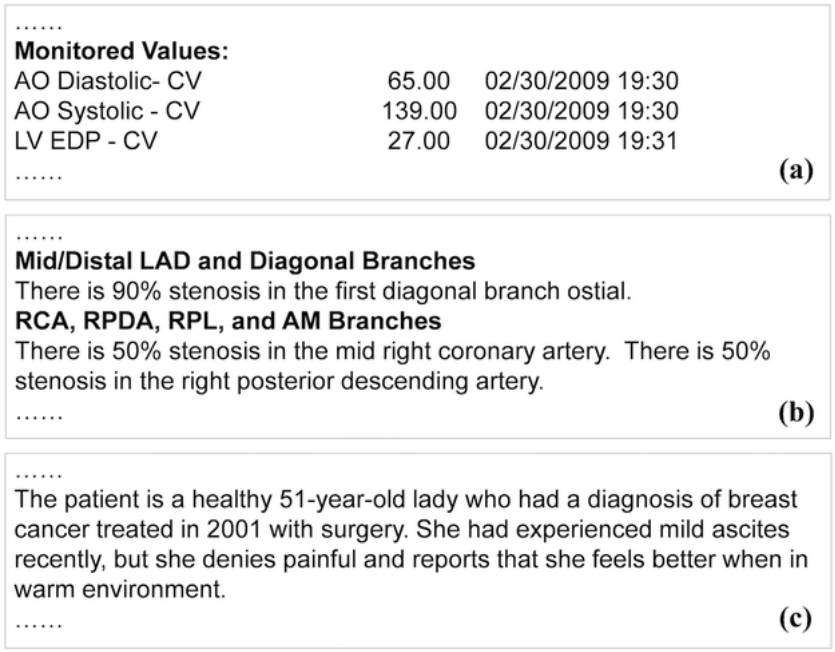
Medicine. It is applied at the care level, both in specialized care (hospital) and in primary care. The clinical trial [5] is an experimental study, framed in the specialty of preventive medicine, whose objective is the experimental evaluation of a product, drug, diagnostic technique or treatment in its application to human beings. From its results, it is intended to obtain information on the efficacy and safety of said product for its subsequent use in clinical medicine. One of the most important as well as controversial and ethically regulated steps in conducting a clinical trial is the selection of subjects [6] to participate in the study. In this stage, the general characteristics that the subjects must meet to be included in the trial are described:

1. Universe of Study: The source from which the participants will be included is established (hospital services, volunteers from the common population, primary care, etc.)

2. Diagnostic criteria: If the study proceeds in subjects with a certain disease, the criteria that confirm the diagnosis of the disease must be marked.

3. Inclusion Criteria: Those characteristics that the subject must meet to be eligible for inclusion in the study must be defined and specified. These requirements can be the age, weight or height of the subject and the quantitative limits that indicate the intensity of the disease. A fundamental requirement for the inclusion of the subject in the study is informed consent.

4. Exclusion Criteria: The characteristics that limit the entry of a subject to


.....
Monitored Values:

AO Diastolic- CV	65.00	02/30/2009 19:30
AO Systolic - CV	139.00	02/30/2009 19:30
LV EDP - CV	27.00	02/30/2009 19:31

..... **(a)**

.....
Mid/Distal LAD and Diagonal Branches
There is 90% stenosis in the first diagonal branch ostial.
RCA, RPDA, RPL, and AM Branches
There is 50% stenosis in the mid right coronary artery. There is 50% stenosis in the right posterior descending artery.
..... **(b)**

.....
The patient is a healthy 51-year-old lady who had a diagnosis of breast cancer treated in 2001 with surgery. She had experienced mild ascites recently, but she denies painful and reports that she feels better when in warm environment.
..... **(c)**

Figure 1: Documents with different levels of structure: (a) Semi-structured Report, (b) Template based narration and (c) Complex narration [4]

the trial are defined. Either due to cognitive disability of the subject himself, due to limitations due to the presence of other diseases or due to medical contraindications.

5. Exit criteria: The criteria that exclude a subject from the statistical analysis of the trial when the study has already started.

Patients who meet all criteria of inclusion (or those inclusion criteria that need to be met) and none of the exclusion ones will be included in the trial.

The main goals of the Selection and Recruitment of individuals as a clinical trial subjects are: burdens and benefits fairly distribution, social value insurance, scientific validity enhancement, harm minimization, benefit maximization and vulnerable individuals protection. For this reason, the steps of the participant selection process described above must be carefully followed to meet the ethical and scientific requirements necessary for the development of the trial.

Making an appropriate selection of subjects is crucial for the development of the study as well as for the veracity and extrapolation of the results. A poor selection of participants can lead to unintended results and consequences. It can cause bias in the results and harm the health of some of the participants.

1.4 Text Mining

More and more data is being generated at greater speed and variety. In the context of hospitals, these data lack structure, so alternative tools or methods to the conventional ones are needed to proceed with their analysis and extraction of informative value. As said, one of the types of unstructured data with which clinicians interact every day is text data, and the method of extracting information from textual data is Text Mining [7]. Text Mining extracts information [8] that can be found explicitly or not within the text by identifying patterns or correlations between words or terms. This method can be effective for texts from web pages, books, emails, reviews, articles or papers, among others. The most powerful applications of this analysis procedure are [9]: Information and concept extraction: allows the acquisition of information that would otherwise be impossible to extract. Is a Natural Language Processing (NLP) task that consists of detecting and structuring information from text data.

Text Classification and Clustering [10]: it allows to identify common characteristics between different texts to classify or group them in different categories. Its development is carried out using Machine Learning and Deep learning techniques with supervised algorithms (in the case of classification) and unsupervised algorithms (in the case of clustering).

As mentioned before, as a combination of computational linguistics and Machine Learning techniques arises what is known as Natural Language Processing (NLP) [11], allows computers to process human language in the form of text (or voice) data learning its full meaning. Its objective is to learn and replicate the human language faculty by handling all the complexities of human language, such as grammar and semantics, structure, sentiment analysis, etc. Its well-known applications are: the translation of texts into hundreds of languages, the automatic transcription of oral language, information retrieval or the creation of interactive dialogues.

1.5 Objectives

One of the most tedious and time-consuming tasks in clinical trials is the subject selection process. The information of the patients susceptible to inclusion in the study must be consulted manually to check if they meet the defined selection criteria.

The objective of this project was to build an automatic subject selection system for clinical trials from longitudinal medical records of patients using Natural Language Processing (NLP) techniques and a hybrid approach combining Machine Learning (ML) models and Rule-based models.

In this way, the time spent in subject selection stage as well as the selection errors would be reduced. In short, the aim of this project was to speed up the process of selecting the participants of a clinical trial, in addition to reducing possible human errors that may arise due to the subjectivity of some of the selection criteria.

In the context of the project, the feasibility and performance of two different methods for determining each patient's compliance with the different selection criteria were studied: A system based on a probabilistic approach using ML algorithms and a deterministic approach using Rule-Based models.

This work followed the goals of the 2018 National NLP Clinical Challenges (n2c2) Cohort Selection for Clinical Trials Shared Task and Workshop [12].

1.6 Related Works

Before elucidating the projects aimed at automating the process of selecting participants for clinical trials, it is necessary to know how this process is currently carried out.

As mentioned above, the process of selecting subjects for clinical trials is a time-consuming phase, involves many health personnel, must comply with certain ethical rules, and requires communication between different departments of a hospital and even between different hospitals.

A team from the clinical trials department is in charge of compiling the information of each patient eligible for inclusion in the clinical trial. The clinical history is consulted manually by different health professionals to determine compliance with each selection criterion. On some occasions, due to the complexity of some selection criteria, the inclusion decision falls subjectively on the expert who consults the clinical history. Other times, also due to the complexity of some of the selection criteria, the professionals in charge of selecting the participants must consult the doctor who referred the patient to the trial for additional information.

Due, therefore, to the fact that the process is not automated, several problems may arise in the development of this step of the clinical trial: human errors due to the subjectivity of some of the criteria, errors in communication between health professionals, errors due to fatigue of the expert who determines compliance with the selection criteria or bias errors.

In addition, the non-automation of the procedure entails a delay in the development of the clinical trial and consequently in obtaining clinical results that can be of vital importance.

For this reason, some hospitals are trying to develop automatic subject selection systems for clinical trials in collaboration with engineers, data scientists and com-

puter scientists. They do it from the automatic analysis of medical records, using Text Mining and NLP techniques together with AI algorithms.

Some teams that participated in the challenge used rule-based classification methods, others used machine learning classification methods, and still others tried hybrid approaches of the two models above.

A team from the Medical University of Graz [13] used a rule-based classifier, orthogonal machine learning strategies, such as support vector machines, logistic regression, and deep learning algorithms such as long short-term memory neural networks. They evaluated these algorithms with vector input representation schemes of the texts using the BioWordVec word embedding algorithm. Using the rule-based classifier, they obtained an overall micro F1 score of 0.9100, the highest of all challenge participants. Using machine learning strategies, they achieved higher scores than with deep learning strategies.

As an example of a hybrid approach to the problem, a Univeristy of Michigan team [14] combined pattern-based, knowledge-intensive, and feature-weighting techniques. They used NLP techniques to preprocess the data, developed individual criteria-specific components by collecting relevant knowledge resources for each criterion, and applied pattern-based and weighting approaches to classify. In this way, they achieved an overall micro averaged F1 of 0.9075, placing second in the challenge ranking.

The team of Harbin Institute of Technology [15] designed a hierarchical neural network composed of 5 components: They used a convolutional neural network (CNN) for the representation of sentences, a highway network to adjust the flow of information, a self-attention neural network to respond sentences, a representation of documents using LSTM, to take into account the chronological order and finally a fully connected neural network to determine the classification of each criterion. Using this procedure, they obtained a maximum micro F1 score of 0.9021.

Outside the context of the challenge, a group made up of experts from various French institutions and universities developed a prototype of a computerized subject recruitment support system (CRSS) based on a semantic web approach [16]. The system was based on data collected from the Urology clinical area and 4 prostate cancer clinical trials. It was designed to be scalable to other clinical domains.

2 Methods

In this section, the materials and methods used to achieve the objectives of the project are described

2.1 Materials

The definition of the dataset and its characteristics, the separation of the data into the training and test subsets, as well as the programs, libraries and database resources used for the classification of the selection criteria of each patient to be part of the clinical trial are described below.

2.1.1 Dataset

The data set given by the Challenge organizers consists of 3 subsets: Training Set, unlabeled Test Set and the same Test Set with the standard gold. In total, it consists of 288 longitudinal records of patients annotated to determine if patients meet ("met" or "not met") a list of 13 selection criteria to be selected as a subject in a clinical trial that studies the risk of heart disease in diabetes patients. All the patients in this dataset had diabetes, and most were at risk for heart disease. Of those records, 70% (202) were made available as the training set and the remaining 30% (86) were kept as the test set. The medical records for this corpus came from a larger corpus of the Partners HealthCare Electronic Medical Records(EMR). EMR comprises a platform shared by two large academic tertiary hospitals: Massachusetts General Hospital (MGH) and Brigham and Woman's Hospital (BWH).

Between 2 and 5 records were added to each patient record in chronological order. In the training set, the annotations "met" and "not met" were added at the end of the record for each of the 13 selection criteria. The corpus contains 781 006 tokens, with an average of 2711 tokens per set of patient records. This provides a reasonably robust set of information about each patient. This corpus was previously de-identified using a "risk averse" interpretation of the Health Insurance Portability Accountability Act guidelines. All information linked to a patient was removed and replaced with realistic surrogates, and dates were time-shifted a random amount for each patient.

For the classification of a patient according to the possibility of being selected as a subject of a cohort of a clinical trial, it was determined if said patient met a series of requirements. The selection criteria are described below. Table 2 also shows the class distribution of each selection criterion in "met" or "not met" of the whole data set, as well as the distribution of this classes between the two subsets of training and test.

As it can be seen in Table 2, some of the criteria showed an unbalanced class distribution. Even one of the criteria (Keto-1YR) had only one patient classified as 'met' in the entire training set and none in the test set. Also, as explained in the challenge description, the only 'met' classification for this criterion was due to a data entry error. Therefore, all the patients in the dataset were actually classified as 'not

	Criterion Description	*Met* (train/test)	*Not Met* (train/test)
Abdominal	History of intra-abdominal surgery.	107 (77/30)	181 (125/56)
Advanced-CAD	Advanced cardiovascular disease (CAD).	170 (125/45)	118 (77/41)
Alcohol-Abuse	Current alcohol use over weekly recommended limits.	10 (7/3)	278 (195/83)
Asp-For-Mi	Use of aspirin to prevent Myocardial Infraction.	230 (162/68)	58 (40/18)
Creatinine	Serum creatinine higher than the upper limit of normal.	106 (82/24)	182 (120/62)
DietSupp-2MOS	Taken a dietary supplement in the past 2 months.	149 (105/44)	139 (97/42)
Drug-Abuse	Drug Abuse: current or past	15 (12/3)	273 (190/83)
English	Patient must speak English.	265 (192/73)	23 (10/13)
HBA1C	Any hemoglobin A1c (HbA1c) value between 6.5% and 9.5%.	102 (67/35)	186 (135/51)
Keto-1YR	Diagnosis of ketoacidosis in the past year.	1 (1/0)	287 (201/86)
Major-Diabetes	Major diabetes-related complication.	156 (113/43)	132 (89/43)
Makes-Decisions	Patient must make their own medical decisions.	277 (194/83)	11 (8/3)
MI-6MOS	Myocardial Infraction in the past 6 months.	26 (18/8)	262 (184/78)

Table 2: 13 selection criteria description and "met" and "not met" distribution through training set and test set for each criterion

met' for that criterion. This made the ketoacidosis diagnosis criterion statistically irrelevant and difficult to study, so it was removed from the analysis.

The different selection criteria could be grouped into 5 categories in relation to the type of answers they received:

1. Evidence of substance abuse: DRUG-ABUSE and ALCOHOL-ABUSE.

2. Numerical inference from laboratory tests: CREATININE and HBA1C.

3. Clinical complications: ABDOMINAL, MAJOR-DIABETES, ADAVNCED-CAD and MI-6MOS.

4. Treatments or medication: ASP-FOR-MI AND DIETSUPP-2MOS.

5. Patient independence: ENGLISH and MAKES-DECISIONS.

2.1.2 Programs, libraries and resources

This project was carried out using Spyder. An integrated development environment for the Python programming language. This software is used in data science and machine learning. This includes high-volume data processing, predictive analytics, and scientific computing. In collaboration, libraries were defined. A set of functional implementations, coded in the Python programming language, that provide a well-defined interface to the functionality for which they are used. In this project, in addition to the basic libraries for data import and manipulation, some specific libraries designed to solve NLP problems were invoked:

NLTK (Natural Language Toolkit) [17]: is used for tokenization, stemming, stemming, parsing, POS tagging, etc. This library has tools for almost all NLP tasks.

Scikit-learn: [18] offers a large library for machine learning, as well as tools for text preprocessing.

2.2 Methodology

This section describes the steps and methodologies that were carried out using the materials described above to achieve the objective of the project.

To identify the diverse set of selection criteria, this project followed a hybrid approach that combined NLP with ML models and rule-based models. All the selection criteria were classified using ML models, while two of them (HBA1C and CREATININE) were also classified using Rule-Based models.

The workflow of the project is defined in Figure 2, where the different steps and stages for the development of the project are described.

2.2.1 Preprocessing

The patient records that made up the data set had a complex narrative structure.The processing stage was the first stage of the methodology of this project and aimed to normalize those texts. This section summarizes the techniques and methodologies used to convert long patient records into a group of words and meaning entities. This process was intended to reduce the complexity of the data and transform the texts into important entities to enhance and facilitate the extraction of concepts and information. The different transformation that were applied over the texts are described below:

Lowercasing: It is one of the most common tasks in text preprocessing and consists of converting all characters to lowercase. It was used to maintain the consistency of the texts in the data set.

Noise removal: Denoising is defined as the text-specific normalization task that often take place before tokenization. Denoising tasks include: removing headers and footers, removing HTML or XML markup and metadata, removing punctuation, and more.

Tokenization: Also defined as text segmentation, it is the process by which

long text strings are divided into shorter meaningful entities (tokens). This process can break the text into words or phrases. In the context of the project, the texts were tokenized by words.

Stop Wrods Removal: Stop words are a set of words commonly used in a given language, such as: conjunctions, prepositions or articles. These words do not add much meaning to a sentence. They can be safely ignored without sacrificing the meaning of the text. In this phase, English Stop Words removed.

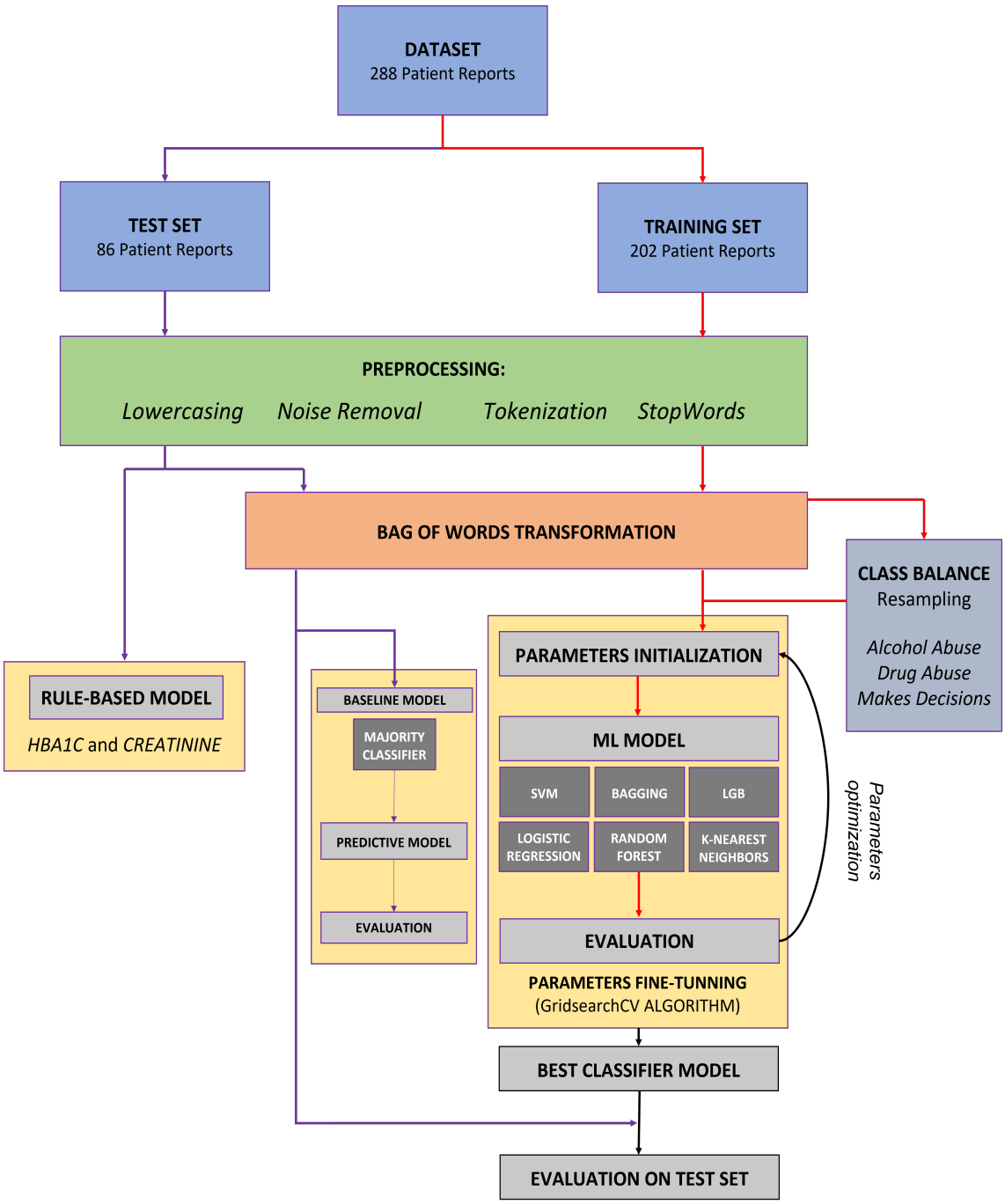


Figure 2: Overall methodology Workflow

Figure 3 shows the frequency of the 25 most abundant tokens in the texts of the data set after preprocessing.

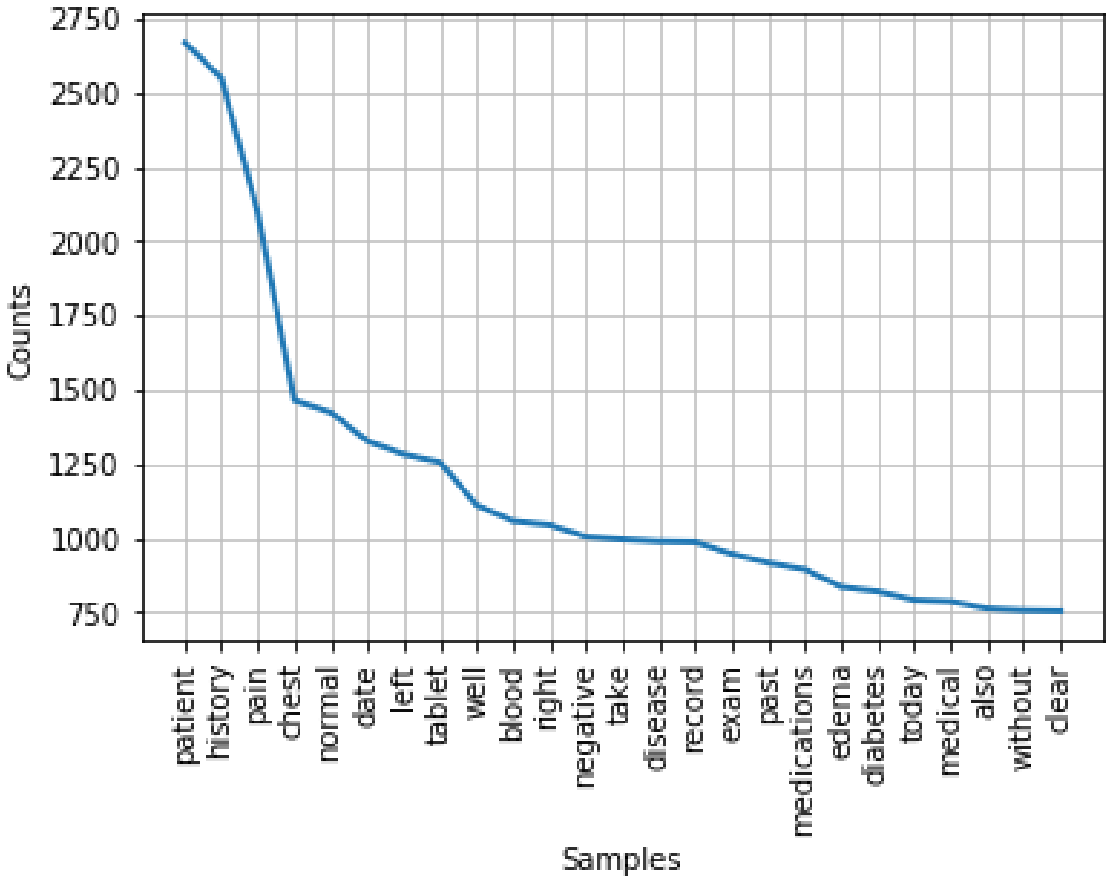


Figure 3: Frequency graph of the 25 most repeated words in the texts of the data set

2.2.2 Bag-of-Words

Once the texts were preprocessed and a representation of the most important words and meaning entities was obtained for each of them, they were transformed into a bag of words. ML algorithms do not have the ability to learn about plain text or strings. Therefore, it was necessary to transform the preprocessed texts into its numerical representation. The Bag-of-Words model is commonly used in document classification methods where the frequency of each word is used as a feature to train a classifier (see Figure 4). The Bag-of-words model is an example of a Vector space model. In the context of the project, the texts were transformed into a matrix with n rows (where n is the number of documents in the data set) and m columns (where m is the number of words that make up the vocabulary that appears in all the data set's texts). Then, each position of the Bag-of-words matrix was filled with the frequency of each vocabulary word in each of the texts of the data set. In the

context of the project, the Bag-of-Words matrix had 21340 columns, which means 21340 different words in the texts.

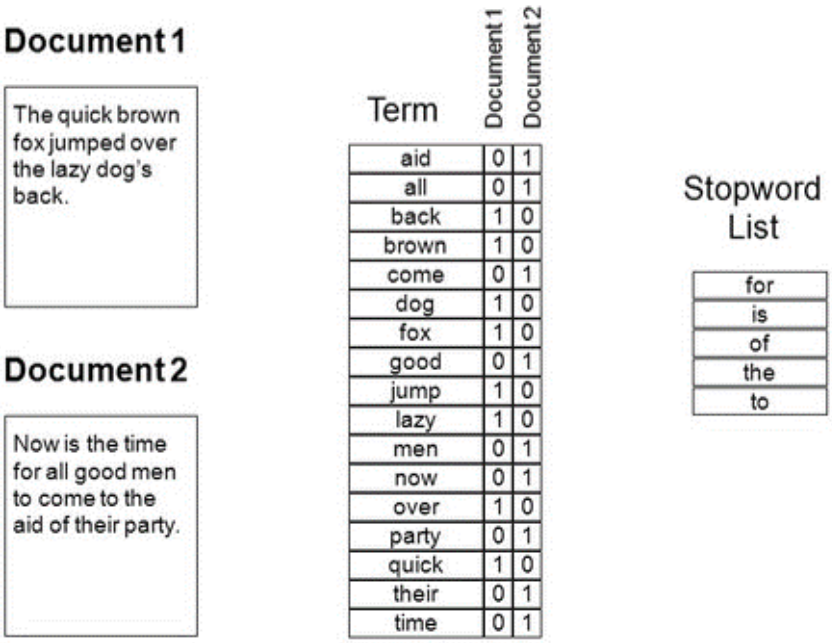


Figure 4: Bag-Of-Words example

2.2.3 Class Balance

As seen in the data set description in Table 2 the distribution of classes between the 'met' and 'not met' categories were not the same for all the selection criteria. Most of the selection criteria had a balanced distribution, which means a similar number of patients classified as each binary class. It was the case of DietSupp-2MOS selection criterion that had 105 patients labeled as 'met' and 97 labeled as 'not met' in the training set and 44 classified as 'met' and 42 classified as 'not met' in the test set. On the other hand there were three criteria that were unbalanced in a ratio approximately of 1:28, which means 1 instance classified as the minority category for every 28 instances classified as the majority category. The three unbalanced selection criteria were: Alcohol-Abuse, Drug-Abuse and Makes-Decisions. An inclusion criteria it was consider as an unbalanced one when the proportion of instances classified as the minority class respect to the total of the instances was lower than 5%. This is considered as an extreme degree of class unbalance.

Class imbalance can lead to errors in the training of ML models that learn a lot about one class (the majority class) and little about the minority class. They can also lead to confusion in the evaluation. This is because most of the time ML models behave as majority classifiers for unbalanced variables. This means that they classify all the instances as the majority class and since there are so few instances classified as the minority, the accuracy of the model can be very high, giving the false impression that the model has a good performance.

There are several techniques designed to solve the problem of class imbalance in classification. In the context of the project, 2 different methods were used:

Resampling technique: It consists of removing samples from the majority class (under-sampling) and/or adding more samples from the minority class (over-sampling) (see Figure 5). In the context of the thesis, two resampling techniques were applied by duplicating random records from the minority class and removing random records from the majority class, until class balance was achieved.

In the context of the thesis, both resampling techniques were applied: duplicating random records of the minority class and eliminating random records of the majority class, until class balance was achieved. Due to the small size of the training set, it was decided to do the resampling in such a way that the number of records would not be reduced, achieving a proportion of classes of 40% of the minority class and the remaining 60% of the majority class.

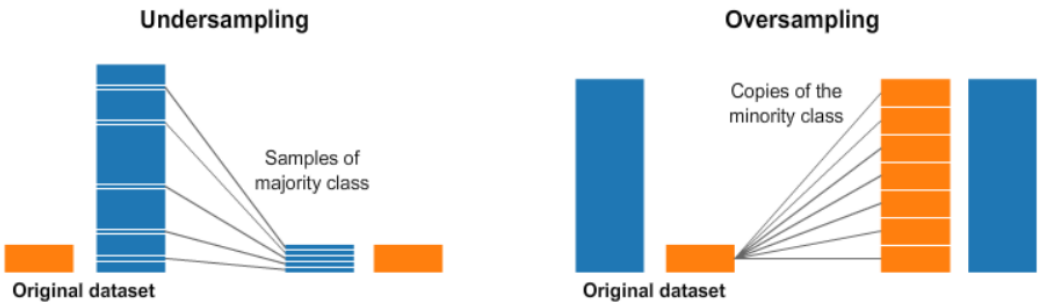


Figure 5: Resampling technique scheme [19]

met / not met	*AlcoholAbuse*	*DrugAbuse*	*MakesDecisions*
Imbalanced	7/195	12/190	194/8
Balanced CAD	83/120	90/119	114/84

Table 3: Comparison between training set class distribution before and after resampling implementation

Penalize Algorithms (Cost Sensitive Training): To increase the cost of miss-classification in the minority class, by penalizing minor class errors by an amount proportional to their under-representation.

Table 3 shows the class distribution resulting from the resampling technique implementation to the training set. Three new balanced training sets where the proportion of the minority class were the 40% were obtained.

2.2.4 Machine Learning algorithms

In order to classify each patient in the categories 'met' or 'not met' for the different selection criteria, the performance of different supervised machine learning classifier algorithms were evaluated, in order to find which was the best classifier for each

selection criterion. Simple models and ensemble methods were implemented using the Python Scikit-Learn library. The different algorithms used during the project are defined below:

K-Nearest Neighbors (K-NN): It is a non-parametric and lazy learning algorithm. Non-parametric means that the model structure is determined from the dataset. K-NN is an algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is the most common amongst its k-nearest neighbors measured by a distance function.

Support Vector Machines (SVM): Given a set of training examples, each marked as belonging to one of the two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training examples to points in space so as to maximise the width of the gap between the two categories.

Bagging: A bagging classifier is an ensemble meta-estimator that fits the basic classifiers to each of the random subsets of the original dataset and then aggregates their individual predictions by voting to form a final classification.

Random Forest: It combines random decision trees with bagging to achieve very high classification accuracy. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses voting to improve the predictive classification accuracy and control over-fitting.

Logistic Regression: Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. So, in its basic form uses a logistic function to model a binary dependent variable.

LightGBM: LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for classification and many other machine learning tasks. A boosting framework involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models misclassified.

Baseline Classifier: It is a basic and simple machine learning model. It is used in a comparative way to check the real performance of the previous classifiers. This algorithm classifies each instance as the majority class of the data set.

2.2.5 Parameter Fine-Tuning

Once the classifiers were chosen, the hyperparameters involved in every algorithm were need to be optimized in order to obtain the best metrics. This was achieved with the GridSearchCV algorithm of the Python Scikit-learn library [20].

This algorithm executed and evaluated on the training set the different classifiers described in the previous section with different combinations of values -previously indicated- for each parameter. GridSearchCV evaluated the performance of the models based on a chosen scoring metric—in the context of the project, the micro F1— through cross-validation.

Table 4 shows the changed parameters, its description and the different values that were tested for each classifier.

2.2.6 Rule-based models

As discussed in the data set description, the selection criteria were grouped into different categories. One of the categories contained those selection criteria that described a numerical inference from a laboratory test. That is, those criteria whose response was a numerical value:

HBA1C: referring to a hemoglobin A1c value between 6.5% and 9.5%.

Creatinine: referring to the concentration of serum creatinine greater than the upper limit of the normal range. A normal creatinine concentration is between 0.7 and 1.3 mg/dL [21].

A rule-based system is a model that applies rules made by experts to solve a specific task [22]. In the context of the project, the rules describing the numerical selection criteria were applied to classify compliance with said criteria for each patient. Among the advantages of rule-based systems is the ability to represent the explicit knowledge of experts, that is, the ability to accurately describe the rules for classification. While the main disadvantage is the inability of the system to learn, so having extracted new knowledge from the system does not provide methods to be able to learn more things and more quickly later.

In the context of the project, two rule-based automatic models were created. One for each of the numerical criteria.

The input data for both models were the preprocessed test set texts. In both cases, the tokens or sets of tokens that defined the parameters of each selection criterion as well as their synonyms were searched in the preprocessed texts. For example, in the case of the HBA1C selection criterion, words such as 'hba1c', 'a1c', 'hemoglobin a1c', 'heme a1c', among others, were searched for.

Subsequently, the numerical values that accompanied, both in previous positions and in later positions, the previously defined tokens were extracted. Finally, these extracted values were compared with the limit values defined by each selection criterion to determine compliance with the criterion. In the case of HBA1C, it was verified if the extracted values were between 6.5 and 9.5 as indicated in the description of the inclusion criteria.

2.2.7 Metrics

Cross-validation was used as the evaluation technique. This procedure consists of repeating and calculating the arithmetic mean obtained from the evaluation measures

ALGORITHM	VARIED PARAMETERS	DESCRIPTION OF THE PARAMETER	TIRED VALUES
K-Nearest Neighbors (K-NN)	N_neighbors	Number of neighbors to use	Range (1,20)
	Weights	Weight function used in prediction	Uniform, Distance
	Metric	Distance metric to use	Euclidean, Manhattan, Minkowski
Support Vector Machines (SVM)	C	Regularization parameter	50, 10, 1.0, 0.1, 0.01
	Kernel	Kernel function	Linear, Rbf, Sigmoid, Poly
	Gamma	Kernel coefficient for 'rbf', 'poly' and 'sigmoid	Scale, Auto
	Class_weight	Weights associated with classes (for unbalanced classes)	None, Balanced
Bagging	N_estimators	Number of base estimators in the ensemble	10, 100, 1000
Random Forest	N_estimators	Number of trees in the forest	10, 100, 1000,10000
	Max_features	Number of features to consider when looking for the best split	Sqrt, Log2
	Class_weight	Weights associated with classes (for unbalanced classes)	None, Balanced, Balanced_subsamples
Logistic Regression	Solvers	Algorithm to use in the optimization problem	Newton, Lbfgs, Liblinear
	Penalty	Norm of the penalty	l1, l2, elasticnet
	C	Inverse of regularization strength	100, 10, 1.0, 0.1, 0.01
	Class_weight	Weights associated with classes (for unbalanced classes)	None, Balanced
LightGB	Boosting_type	Type of boosting	Gbdt, Goss, Dart
	Num_leaves	Maximum number of leaves for tree for base learners	30, 50, 100, 150
	Learning_rate	Boosting learning rate	$e^{\ln(a)}$ where a = range(0.005,0.2)
	Class_weight	Weights associated with classes (for unbalanced classes)	None, Balanced

Table 4: Fine-tuned parameters

(metrics) on different partitions of de dataset. Ensuring, in this way, that all data has been used, at least once, as evaluation data. This guarantees the independence of the model performance from the partition between training and test data.

Accuracy, Precision, Recall and measured F (F1) [23] were defined as the evaluation metrics for this project. For each criterion, Precision, Recall, and F1 were calculated for "met" and "not met" responses, then averages were calculated to obtain the microscore for each criterion. The average of all selection criteria together was then evaluated to obtain the overall micro-averaged F1 score. Besides, the accuracy of every selection criterion was also calculated.

$$\textbf{Accuracy} = \frac{True\ Postitive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

$$\textbf{Precision} = \frac{True\ Postitive}{True\ Positive + False\ Positive}$$

$$\textbf{Recall} = \frac{True\ Postitive}{True\ Positive + False\ Negative}$$

$$\boldsymbol{F_{measure}} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 6: Evaluation Metrics

The precision, measures the correctness of the classification prediction between "met" and "not met" for each selection criterion. It measures the proportion of well-labeled criteria for a given class with respect to the total number of criteria labeled with that class. It is a metric that should be enhanced to minimize False Positives.

The recall (or sensitivity), is the proportion of criteria correctly labeled with a certain class with respect to those that should have been labeled with that class. It is a metric that should be enhanced to minimize False Negatives

Measure F1 is defined as the harmonic median of the metrics described above (Precision and Recall). It is a metric that allows control of False Positives and False Negatives.

Finally, Accuracy is the fraction of well classified instances respect to total.

In this project the average F1 (macro F1) metric was calculated as the simple binary average between F1 measure obtained for the two class ('met' and 'not met') for each criterion. Then, the micro F1 score (short for micro-averaged F1 score) was also calculated for every criterion. It measures the F1-score of the aggregated contributions of all classes. Finally, the overall macro F1 score and the overall micro F1 score were calculated as the average of these metrics of all the criteria.

The final metric used for ranking the model's results was overall micro F1 score evaluated over the test set.

3 Results

3.1 Machine Learning models and parameter Fine-tuning

Once the preprocessing and bag of words transformations were applied over the longitudinal clinical history of every patient of the data set, the different ML classifiers were evaluated with the different parameters combinations as described in Parameter Fine-Tuning section. The performances of every model over every criterion is shown in Table 5. The ML models performances of the numerical criteria (HBA1C and Creatinine) are also shown as well as the ML models performances of the balanced and imbalanced Alcohol-Abuse, Drug-Abuse and Makes-Decisions criteria.

Classifiers/Criteria	*Ridge*	$K-NN$	*SVM*	*Bagging*	*RandForest*	*LogReg*	*LGB*	*Baseline*
Abdominal	0.698	0,651	0,723	0,814	0,640	0,709	0,791	0,651
Advanced CAD	0.709	0,581	0,698	0,826	0,581	0,709	0,814	0,523
Alcohol Abuse	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965
Alcohol Abuse balanced	0.907	0.942	0.895	0.965	0.860	0.953	0.965	0.965
Asp For MI	0.756	0.791	0.791	0.814	0.779	0.779	0.826	0.790
Creatinine	0.791	0.709	0.791	0.767	0.756	0.791	0.779	0.710
DietSupp 2MOS	0.605	0.523	0.616	0.663	0.640	0.593	0.732	0.512
Drug Abuse	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965
Drug Abuse balanced	0.779	0.919	0.919	0.884	0.860	0.930	0.988	0.965
English	0.849	0.872	0.849	0.919	0.849	0.849	0.860	0.849
HBA1C	0.604	0.604	0.628	0.674	0.605	0.593	0.721	0.593
Major Diabetes	0.663	0.570	0.697	0.802	0.700	0.709	0.849	0.500
Makes Decisions	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965
Makes Decisions balanced	0.860	0.870	0.884	0.756	0.951	0.0.953	0.978	0.965
MI 6MOS	0.907	0.895	0.907	0.907	0.907	0.919	0.919	0.512
Average performance	0.802	0.788	0.771	0.846	0.802	0.825	0.845	0.71

Table 5: Best micro F1 score of each criterion for every classifier after applying Fine-tuning

On the one hand, best results were achieved by the Bagging algorithm for 3 of the 12 studied selection criteria: Abdominal, CAD, and English. On the other hand, the best results for DietSupp 2MOS, HBA1C, Major Diabetes, MI 6MOS and ASP For MI criteria were achieved when applying the LightGBM classifier. For the Creatinine criterion the model that achieved the best results was the Logistic Regression algorithm. However, three of the creterion (imbalanced Alcohol Abuse, Drug Abuse and Makes Decisions) achieved the same performance score for all the classifiers. In the case of the balanced Alcohol Abuse, Drug Abuse and Makes decisions criteria, the highest micro F1 score was achieved when applying the LGB model.

There was a selection criterion whose classification using ML models failed with respect to the Baseline model. This criterion was MI 6MOS and its optimized model obtained a micro F1 performance of 0.79 compared to 0.91 achieved by the Baseline.

The criteria that achieved the lowest micro F1 score were HBA1C and DietSupp 2MOS rising to near 0,6 (0,59 and 0,62 respectively). Contrary, the criteria that showed best model performance were imbalanced Alcohol Abuse, Drug Abuse and Makes Decisions with a micro F1 score of 0,97, which means that the 97% of the patients of the test set were correctly classified for these criteria. However, the micro F1 of these three criteria coincided with the micro F1 of the Baseline classifier (0.97). Studying the confusion matrix of these criteria, it was confirmed that the ML models had acted as classifiers for the majority, just like the Baseline model. In contrast, the ML models of these same criteria with the class balance showed a different performance than the Baseline. For this reason, it was decided to make a comparative study between the micro F1 score and the macro F1 score of the model with the best performance of the imbalanced criteria compared to the same balanced criteria.

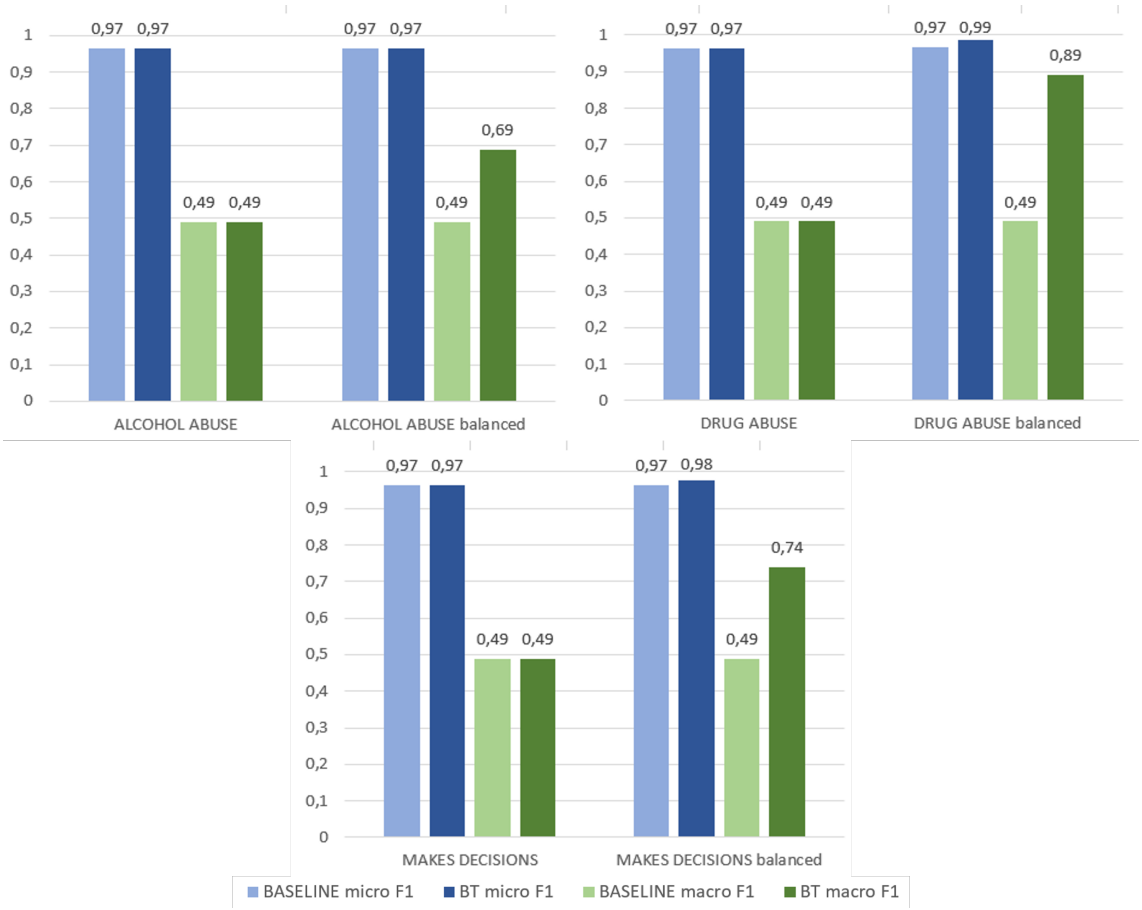


Figure 7: Micro F1 and macro F1 score of the imbalanced and balanced criteria in comparison with the Baseline performance

In Figure 7 the comparison between the micro F1 and macro F1 scores of the best model performances for the imbalanced and balanced Alcohol Abuse, Drug Abuse and Makes Decision criteria with the Baseline performance is observed. It can be

seen that for the balanced criteria, the macro F1 score were relatively different and higher than the Baseline and the imbalanced criteria models. For Drug Abuse and Makes decisions the micro F1 score of the balanced models were also higher, whilst the micro F1 of the balanced Alcohol Abuse was the same as the Baseline and the imbalanced model. This confirmed that the ML models for the balanced criteria no longer acted as majority classifiers and that they were sensitive to the minority class unlike the models for the unbalanced criteria. Therefore, it was decided that the best models for these selection criteria were those of the balanced criteria.

3.2 Rule-based model

When applying the Rules-based model to the numerical selection criteria (HBA1C and Creatinine), the micro F1 scores achieved were 0.87 and 0.82, respectively. As can be seen in the comparative graph in Figure 8, these micro F1 scores were higher than when the ML model and the Baseline classifier were applied to the same inclusion criteria. Confirming that the Rule-Based model achieved a better performance than the ML model for both criteria.

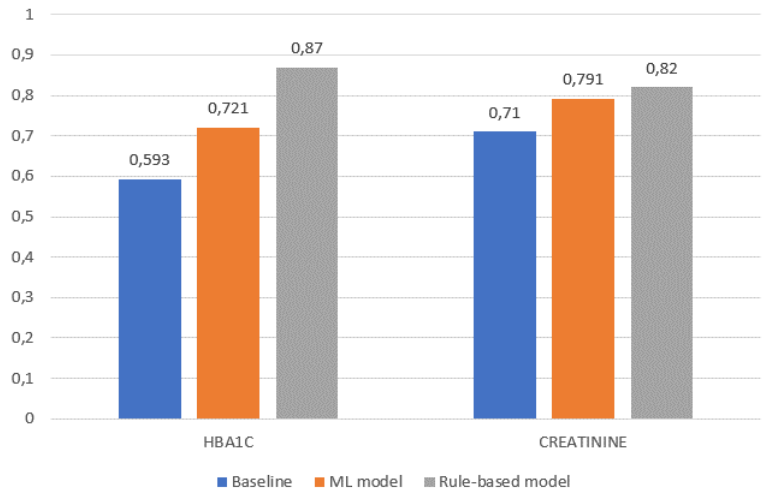


Figure 8: Performance (micro F1) comparison of Baseline model, ML model, and Rule-base model on HBA1C and Creatinine selection criteria

3.3 Overall model performance

In Figure 9 the comparison between the micro F1 and the macro F1 scores between the best model and the Baseline of each criterion is shown. For most of the selection criteria, an improvement was observed in both the macro F1 and the micro F1 scores of the proposed model with respect to those of the Baseline. Except, as mentioned before, for the MI 6MOS criterion, whose micro F1 was lower than Baseline even though macro F1 was higher.

When comparing the performance of the whole project's model (BT model), calculating the overall macro F1 and the overall micro F1 as the arithmetic average of each score of every selection criteria, with the performance (overall macro F1 and

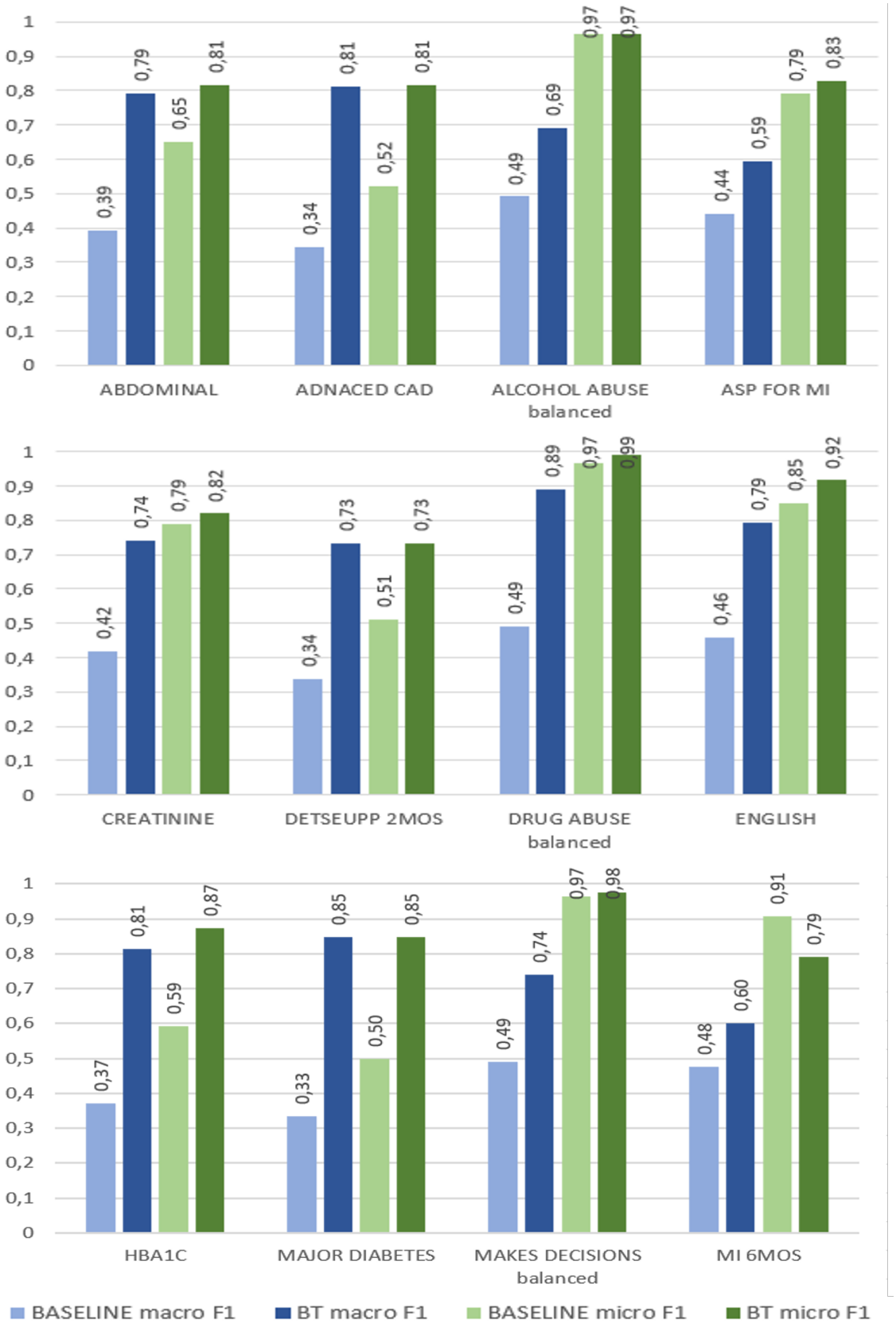


Figure 9: Micro F1 and macro F1 of the best model for each selection criterion in comparison with the Baseline performance

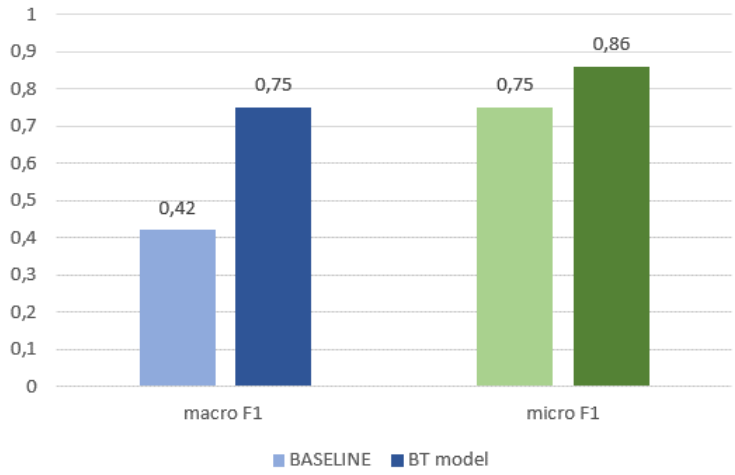


Figure 10: Overall micro F1 and macro F1 scores of the proposed model with respect to the Baseline

overall micro F1) achieved by the Baseline classifier, relatively high differences were observed. Relating the overall macro F1 score the baseline achieved a 0,42 value, whilst the project's model reached a 0,75 value. This is an absolute difference of 0,33 points, which corresponds to a relative difference of almost the 78,6%. Relating the overall micro F1 score the project's model reached a value of 0,86, whilst the Baseline classifier showed an overall micro F1 of 0,75. This means, an absolute difference of 0,11 and a relative difference of 14,67% as it can be seen in Figure 10.

3.4 Challenge Ranking

When comparing the average micro F1 of the proposed project's model (BT model) with that of the participants in the challenge, in Figure 11, it could be seen that the results were all very similar moving around an overall micro F1 of 0,9. However, the overall micro F1 of this project was the lowest of the challenge ranking only 0.0138 points below the next (National Taitung and Taipei Medical team). There was also no substantial difference between the proposed model performance and the one proposed by the best team as the overall micro F1 of the best participant (Medical University of Graz team) achieved 0.9100, only 0.0473 points above the BT's model performance.

4 Discussion

As described, Text Mining is a hot topic in the analysis of unstructured data and its application to medicine has begun to be studied in recent years, providing very satisfactory results in different fields. In the context of the thesis, the efficacy of an hybrid system for the classification of patient medical records for compliance with 12 selection criteria was studied, combining Machine Learning techniques and Rule-Based techniques.

At the beginning of the project, one of the selection criteria was discarded from

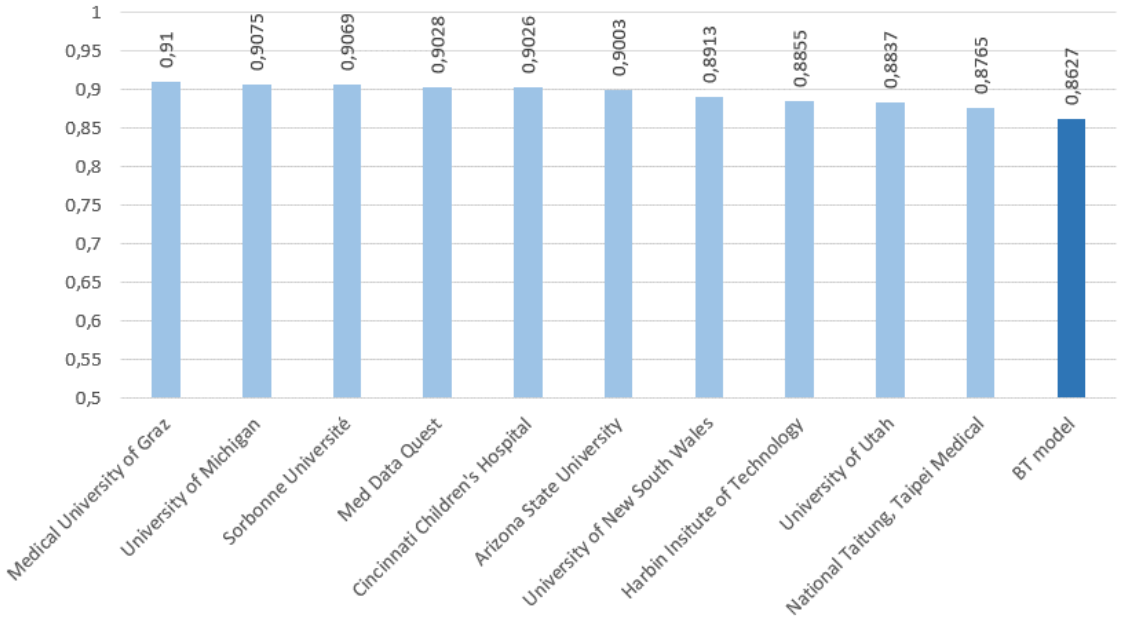


Figure 11: Top 10 challenge participant teams ranked by micro F1 score compared with micro F1 score of the project

the analysis (Keto-1YR referring to a diagnosis of ketoacidosis the year before the trial). The class distribution of this criterion made its analysis impossible since in the data set all the records were classified as the same class. With such an extreme class imbalance as that, no class balancing technique could have reversed the situation.

In fact, it was not the only selection criteria with an extreme class imbalance. Three selection criteria (Alcohol Abuse, Drug Abuse and Makes Decisions) showed a proportion of the minority class of less than 5% with respect to the total records. Then, a resampling procedure was applied over the training set. The performance of the different models described with both versions of the training set with respect to these criteria was studied. The original criteria (with the unbalanced classes) and the same criteria after resampling. This was done to improve the performance of the model and its sensitivity to the minority class, since with the original training set the models had behaved as majority classifiers. Balancing techniques were shown to positively influence the performance of the models, but there are certain aspects that should be taken into account. Implementing random under-sampling of records from the majority class could result in the loss of invaluable information for a model, while implementing random oversampling of the minority class' records could lead to overfitting for the models.

The method implemented for the numerical representation of the texts was the Bag of Words. This technique proved to be a good input for machine learning classification models considering the results obtained. However, Bag of Words discards any information about the order or structure of the words in the document. This could be one of the limitations of the project. Not analyzing the order in which the words were organized in the different sentences of the text could lead to the loss of information, for example, in the negations. If the text contained the words 'heart attack' preceded by negations such as 'no' or 'denied', these negations were

ignored by the Bag of Words rendering method. Besides, Bag of Words did not parse words semantically. Since semantic relationships were not studied, the different words with similar meanings were not represented as synonyms. This is why, for future work, the implementation of embedding methods could be proposed, such as Word2Vec, which represent individual words in a text, taking into account the context and other surrounding words with which that word appears, as well as the semantic relationships between words with similar meaning.

Regarding the combination of both classification approaches: Machine Learning and rule-based systems, it was confirmed that for the numerical selection criteria (HBA1C and Creatinine), whose inclusion rules were very well defined and their answers did not have a linguistic nature, the ML models they were not the best approach. With the rule-based systems it was possible to extract numerical values from the texts and analyze them comparatively with the exact definition of the selection criterion's rules. One of the advantages of rule-based systems is their great ability to obtain good results when working with small data sets, as was the case in this project. On the other hand, one of the most important limitations of rule-based systems is their inability to learn and to be able to reuse the information learned for future tasks.

Regarding classification using ML models, it was shown that, although all models performed better than Baseline, the Bagging, Random Forest and LGB assembly methods absolutely dominated against basic models such as SVM or K-NN. Regarding the basic models, the one that had a more satisfactory performance, approaching the performance of the assembly methods, was the Logistic Regression algorithm.

The metrics used for the evaluation of the performance of the models were the micro F1 score and the macro F1 score. Micro F1 was chosen as the main metric because it was the one determined to establish the Challenge ranking. However, in this type of problem where the binary classification is combined with a class imbalance, the micro F1 achieves the same values as the accuracy metric. Accuracy can lead to confusion, since those models evaluated with the criteria whose classes were unbalanced achieved very high micro F1 values even when the minority class was not being detected. Therefore, it was argued that even if it were the metric determined to establish the ranking of the challenge, it would not be the only one to study. For this reason, the F1 macro was also determined as the metric to be analyzed, to take into account the sensitivity of the models to both classes, making the harmonic median between the F1 score of each class.

One of the most important limitations that had to be dealt with in the development of the project was the reduced size of the data set. It is difficult to specify the optimal data set size to maximize results when applying ML classifiers. But, given the complexity represented by the texts written in free narrative format, everything indicates that an increase in the size of the data set could have helped improve the performance of the proposed models. The possibility of accessing more data could have been considered, but in the context of textual data it is often difficult to get large amounts of data as well as label it. This process would entail the need for financial resources, personnel and time that would be difficult to achieve. Another option could have been the implementation of Data Augmentation techniques, this means, generating additional or synthetic text from the original data set.

For future works directed to the classification of biomedical text, the implementation of transfer learning could be considered, that is, using pretrained models with large amounts of data and fine-tuning the parameters for the specific classification task. An example of this could be the implementation of the pretrained BERT model (Bidirectional Encoder Representations from Transformers).

5 Conclusion

This study concludes that the use of a hybrid approach that combines supervised machine learning algorithms and rule-based models offers the possibility to develop a useful tool for the automatic selection of patients for a clinical trial cohort, classifying them according to whether they meet the selection criteria defined by analyzing their medical records.

However, due to the specificity of the selection criteria for the different clinical trials, it should be considered the need to create a large and balanced database with medical records classified according to their inclusion in the different existing clinical trials as well as according to the compliance with the different selection criteria defined for each of them. The possibility of adding other types of data to be analyzed, such as biomedical images, should also be considered.

In addition, both the choice of approach (machine learning or rule-based systems) and the chosen algorithm remain important factors that can limit the accuracy of the system.

All this could help in the optimal prediction of the classification, reducing the time consumed in this stage of clinical trials, as well as the personnel involved and the human errors that derive from the current manual consultation procedure.

6 Additional information

6.1 Parameters of the best models

Here, the parameters that were fine-tuned with the GridSearchCV algorithms to reach the best model for those selection criterion classified with the ML algorithms are described:

Andominal:
Best Model: Bagging
n_estimators: 1000

Advanced CAD:
Best Model: Bagging
n_estimators: 1000

Alcohol Abuse balanced:
Best Model: LGB
Boosting Type: Gradient Boosting Decision Tree
Number of Leaves: 50
Learning Rate: 0,0881
Class Weight: None

Asp For MI:
Best Model: LGB
Boosting Type: Gradient-based One-Side Sampling
Number of Leaves: 30
Learning Rate: 0,0881
Class Weight: Balanced

DietSupp 2MOS:
Best Model: LGB
Boosting Type: Gradient Boosting Decision Tree
Number of Leaves: 30
Learning Rate: 0,132
Class Weight: None

Drug Abuse Balanced:
Best Model: LGB
Boosting Type: Gradient Boosting Decision Tree
Number of Leaves: 50
Learning Rate: 0,132
Class Weight: None

English
Best Model: Bagging
n_estimators: 10

Major Diabetes:
Best Model: LGB
Boosting Type: Gradient Boosting Decision Tree
Number of Leaves: 50
Learning Rate: 0,132
Class Weight: None

Makes Decisions balanced:
Best Model: LGB
Boosting Type: Gradient Boosting Decision Tree
Number of Leaves: 50
Learning Rate: 0,132
Class Weight: None

MI 6MOS:
Best Model: LGB
Boosting Type: Gradient Boosting Decision Tree
Number of Leaves: 30
Learning Rate: 0,132
Class Weight: Balanced

Bibliography

[1] Emanuel EJ. Obermeyer Z. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." In: *Chromosoma* 13.374 (2016), pp. 1216–1219. DOI: https://doi:10.1056/NEJMp1606181.

[2] JOHN HARLEY WARNER GUENTER B. RISSE. "Reconstructing Clinical Activities: Patient Records in Medical History, Social History of Medicine". In: *Social History of Medicine* 5 (1992), pp. 183–205. DOI: https://doi.org/10.1093/shm/5.2.183.

[3] Kong Hyoun-Joong. "Managing Unstructured Big Data in Healthcare System". In: *hir* 25.1 (2019), pp. 1–2. URL: http://www.e-sciencecentral.org/articles/?scid=1119739.

[4] Shuai Zheng et al. "Effective Information Extraction Framework for Heterogeneous Clinical Reports Using Online Machine Learning and Controlled Vocabularies". In: *JMIR Medical Informatics* 5 (May 2017), e12. DOI: 10.2196/medinform.7235.

[5] Farrell B. Kenyon S. Shakur H. "Managing clinical trials." In: 11.178 (2010), pp. 183–205. DOI: https://doi.org/10.1186/1745-6215-11-78.

[6] Kennedy W. Laurier C. Malo J. Ghezzo H. L'archevêque J. "DOES CLINICAL TRIAL SUBJECT SELECTION RESTRICT THE ABILITY TO GENERALIZE USE AND COST OF HEALTH SERVICES TO "REAL LIFE" SUBJECTS?" In: *International Journal of Technology Assessment in Health Care* 1.19 (2003), pp. 8–16. DOI: https://doi.10.1017/S0266462303000023.

[7] A. Spasic I. Ananiadou S. McNaught J.Kumar. "Text mining and ontologies in biomedicine: making sense of raw text." In: *Briefing in Bioinformatics* 3.6 (2005), pp. 239–251. DOI: https://doi.org/10.1093/bib/6.3.239.

[8] Zheng S. Lu J. Ghasemzadeh N.Hayek S. Quyyumi A. Wang F. "Effective Information Extraction Framework for Heterogeneous Clinical Reports Using Online Machine Learning and Controlled Vocabularies". In: *JMIR Med Inform* 2.5 (2017). DOI: https://medinform.jmir.org/2017/2/e12.

[9] Luque C. Luna JM. Luque M. Ventura S. "An advanced review on text mining in medicine." In: *WIREs Data Mining Knowl Discov.* (2019). DOI: https://doi.org/10.1002/widm.1302.

[10] J.J. García Adeva J.M. Pikatza Atxa M. Ubeda Carrillo E. Ansuategi Zengotitabengoa. "Automatic text classification to support systematic reviews in medicine," in: *Expert Systems with Applications* 41 (2014), pp. 1498–1508. DOI: https://doi.org/10.1016/j.eswa.2013.08.047..

[11] Chowdhary K.R. "Natural Language Processing. In: Fundamentals of Artificial Intelligence." In: (2020). DOI: https://doi.org/10.1007/978-81-322-3972-7_19.

[12] Amber Stubbs Michele Filannino Ergin Soysal Samuel Henry Özlem Uzuner. "Cohort selection for clinical trials: n2c2 2018 shared task track 1". In: *Journal of the American Medical Informatics Association* 26.11 (2019), pp. 1163–1171. DOI: https://doi.org/10.1093/jamia/ocz163.

[13] Kreuzthaler M Oleynik M Kugic A. "Evaluating shallow and deep learning strategies for the 2018 n2c2 shared-task on clinical text classification". In: *J Am Med Inform* (2019). DOI: https://medinform.jmir.org/2017/2/e12.

[14] Kreuzthaler M Oleynik M Kugic A. "Hybrid bag of approaches to characterize selection criteria for cohort identification". In: *J Am Med Inform* (2019). DOI: https://medinform.jmir.org/2017/2/e12.

[15] Xiong Y. Shi X. Chen S. Jiang D. Tang B. Wang X. Chen Q. Yan J. "Cohort selection for clinical trials using hierarchical neural network". In: *Journal of the American Medical Informatics Association* 26.11 (2019), pp. 1203–1208. URL: https://doi.org/10.1093/jamia/ocz099.

[16] Cuggia M. Campillo-Gimenez B. Bouzille G. Besana P. Jouini W. Dufour J. C. "Automatic Selection of Clinical Trials Based on A Semantic Web Approach." In: *Studies in health technology and informatics* 216 (2015), pp. 564–568.

[17] Steven Bird Edward Loper. "NLTK: The Natural Language Toolkit, Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics". In: *Association for Computational Linguistics* (2002).

[18] Kramer O. "Machine Learning for Evolution Strategies. Studies in Big Data," in: *J Am Med Inform* 20 (2016). DOI: https://doi.org/10.1007/978-3-319-33383-0_5.

[19] Mrozek Petr; Panneerselvam John; Bagdasar Ovidiu. "Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets". In: *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)* (2022). URL: http://hdl.handle.net/10545/625574.

[20] Ghulab Nabi Ahmad et al. "Efficient Medical Diagnosis of Human Heart Diseases using Machine Learning Techniques with and without GridSearchCV". In: *IEEE Access* (2022), pp. 1–1. DOI: 10.1109/ACCESS.2022.3165792.

[21] S. Goya Wannamethee, A. Gerald Shaper, and Ivan J. Perry. "Serum Creatinine Concentration and Risk of Cardiovascular Disease". In: *Stroke* 28.3 (1997), pp. 557–563. URL: https://www.ahajournals.org/doi/abs/10.1161/01.STR.28.3.557.

[22] N. Indurkhya S. M. Weiss. "Rule-based Machine Learning Methods for Functional Prediction". In: 3 (1995), pp. 557–563. URL: https://doi.org/10.1613/jair.199.

[23] David M. W. Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *International Journal of Machine Learning Technology* (2011), pp. 37–63.