# Prioritizing patients for stomach cancer screening programs: a machine learning approach

Presentado por:

María Carolina Poveda Amaya [a]

Asesores:

David Barrera Ferro [b], Diego Alejandro Patiño [c], Raúl Murillo [d]

[a] *Estudiante, Maestría en Ingeniería Industrial, Pontificia Universidad Javeriana*
[b] *Profesor, Departamento de Ingeniería Industrial, Pontificia Universidad Javeriana*
[c] *Profesor, Departamento de Ingeniería Electrónica, Pontificia Universidad Javeriana*
[d] *Director Científico, Centro Javeriano de Oncología*

---

**Abstract:** Stomach cancer ranks fifth in incidence and is the fourth cause of death by cancer in the world. Since usually this disease is asymptomatic or the symptoms are shared with other diseases, it is diagnosed when the probabilities of recovery are low or null. In this context, performing endoscopy screenings and biopsy follow-ups during early stages could allow the detection of stomach cancer when the patient has a higher probability of recovery. Hence, a proper prioritizing of patients can make feasible the implementation of endoscopy screening programs. This work presents a Decision Support System (DSS) to support the prioritization of patients for endoscopy screening programs. For this purpose, we use the information available in the national healthcare system of Colombia (Sistema General de Seguridad Social en Salud, SGSSS). Our contribution to literature is twofold. First, we identify variables that explain the probability of being diagnosed with stomach cancer, including clinical pathways modeled from a Process Mining approach. Second, we assess the effectiveness of two machine learning approaches for classifying patients and their performance in terms of coverage. Our results show a feasible way to design prevention programs for patient prioritization in a cost-effective approach.

*Keywords: Machine Learning, Process mining, Clinical pathways, early detection, prevention, stomach cancer*

---

## 1. Introduction

Cancer is a complex disease that causes uncontrolled, undesirable growth of abnormal cells (Mallavarapu et al., 2020). Consequently, a malignant tumor can appear in organs and body tissues (Camargo et al., 2004). According to the World Health Organization WHO (2020), this disease causes 9.6 million deaths every year, and it is the second cause of death worldwide. In 2020, more than 19 million new cases were detected around the world, and 71% of deaths were in lower and middle-income countries (LMICS), (World Health Organization.WHO, 2020b). For Latin America and the Caribbean region, the burden predictions forewarn of the need to plan for the provision of oncological care services and the human health force, since cancer incidence will rise from 1.5 million new cases in 2020 to over 2.4 by 2040, implying a close to 66% increase overall (Piñeros et al., 2022).

In 2020, stomach cancer ranked fifth in incidence rate with 5.6% of new cases and it was the fourth leading cause of death by cancer (The Global Cancer Observatory, 2020). Further, current incidence and mortality rates are high in East Asia, Latin America, and Eastern Europe and within specific subgroups in the USA (Balakrishnan et al., 2017). Worldwide, more than 90% of cases are detected in late stages of the disease, with a five-year survival rate lower than 10% (Gómez et al., 2015). In this context, early detection and timely access to treatment for stomach cancer are crucial (Broeders & Elfström, 2020). Pasechnikov et al. (2014) found that endoscopy screening and biopsy follow-up are still considered as the standard for early cancer detection for this disease.

The purpose of screening is to catch early stages of the disease and to delay or prevent further development (Ro et al., 2015). According to Oliveros et al. (2019) the strategy of screening and treating Helicobacter Pylori seems to be the best approach for reducing stomach cancer risk. Authors also found that this strategy is cost-effective, considering that the costs of testing and treating this infection represent less than 1% of the costs of stomach cancer treatment. Additionally, Ro et al. (2015) conclude that in most high-risk populations, such as Singapore, Japan, and Korea, endoscopy screening reported the most cost-effective screening method using incremental cost-effective ratio (ICER) analysis. However, screening programs face important implementation challenges in low-income and middle-income countries, since resources usually are invested in problems considered of greater urgency (Oliveros et al., 2019). Consequently, prioritizing patients can make feasible the implementation of endoscopy-based screening programs in limited resource settings.

In Colombia, stomach cancer is the fourth type of cancer in incidence, and the leading cancer death cause (World Health Organization.WHO, 2020b). On the one hand, since this is usually an asymptomatic disease, in Colombia, the diagnosis takes place in a stage with low or null probabilities of recovery (Barreto et al., 2019). On the other hand, the national healthcare system does not have monitoring and prevention programs for stomach cancer, since this issue is not prioritized as a public health issue (Oliveros et al., 2019). This have led to stomach cancer accounting for the 5.7% of new cases detected and the 14.9% of deaths caused by cancer in the country yearly (World Health Organization.WHO, 2020a). Therefore, the design of prevention and early detection strategies is crucial. One of the approaches of those programs could be screening.

In this context, the information registered by the healthcare systems is highly valuable to develop prediction models and improve the earliness for diagnosing (Villamil et al., 2017). Applying machine learning techniques to solve problems related to different stages of stomach cancer is very common. Nevertheless, recent work has shown some challenges to be tackled. First, most of the research answers questions related to the stages following detection: predicting prognosis, chemo-resistance, treatment benefit, treatment guidance, recurrent risk and survival (Jin et al., 2020). Additionally, the developments must consider the context's characteristics and realities. For instance, in low-income and middle-income countries, available information quality could be a limitation for recommender systems development (Strasser-Weippl et al., 2015). Consequently, it is essential to adapt international protocols and generate local models for the application of machine learning models (Cazap et al., 2019; Strasser-Weippl et al., 2015).

In this work we develop a Decision Support System (DSS) to prioritize patients, based on the risk level of being diagnosed with stomach cancer, using information available in the national healthcare system in Colombia. According to Chaudhuri & Bose (2020) the use of information systems, in contexts with multiple variables and fragmented data, has the potential to improve resource allocation strategies. We follow the Design Science Research (DSR) approach proposed by Peffers et al. (2007) to describe the development of the DSS and address the following questions:

- How reliably can routinely collected data on patient visits be used to predict probabilities of being diagnosed with stomach cancer?
- Which Machine Learning approach performs best, using the AUROC score?
- How might insights obtained from these classification models be used in practice to prioritize patients for endoscopy screening strategies?

The work is structured as follows. Section 2 presents recent work in machine learning applications for stomach

cancer early detection. Section 3 shows general and specific objectives. Section 4 presents our methodology approach, describing the problem definition, the proposed solution, the design, and the demonstration and evaluation phases. Section 5 presents a descriptive analysis of the available data. Section 6 discusses the results and how these classifications techniques could be used in practice to prioritize patients for endoscopy screenings. Finally, Section 7 presents some general conclusions and future work.

## 2. Related work

The use of machine learning models to support decision making in cancer management and control is a growing body of knowledge. Particularly, three recent reviews have analyzed work on stomach cancer (Kailin et al., 2021, Jin et al., 2020 and Niu et al., 2020). According with the authors, despite being a widely studied area, two challenges remain to be tackled. First, most of the work analyses data collected after the diagnosis. Jin et al. (2020) gathered 68 studies where only 14 aimed at supporting detection. Second, recent developments have been focused on the use of images and videos as source of information. Kailin et al. (2021) gathered retrospective studies done between 2013 and 2021. From that, 14 works used images and 2 used videos. Niu et al. (2020) included 31 papers that used images, videos and *in vivo* Raman spectra. Jin et al., (2020) present 57 studies based on images and videos from different types of endoscopies, while only 11 are based on text data. Table 1 presents a summary of relevant studies regarding Machine Learning applications for stomach cancer

*Table 1 - Machine Learning applications for stomach cancer literature review*

| Authors | Machine Learning Techniques | Variables | Stage | Data type |
|---|---|---|---|---|
| (Ali et al., 2018) | Gabor-based gray- level co-occurrence matrix (G2LCM) | Texture features (images) | Diagnosis | Image |
| (Gholami et al., 2021) | Combination of deep neural network, Support Vector Machine, and Deep Convolutional Neural Network (CNN) | Tongue colors<br>Lint features of the tongue | Diagnosis | Image |
| (Huang et al., 2018) | Compound covariate classifier<br>Diagonal linear discriminant analysis (DLDA) classifier<br>Support Vector Machine classifier | Micro RNAs biomarkers:<br>miR-21-5p<br>miR-22-3p<br>miR-29c-3p | Chemo resistance | Alphanumeric |
| (Ishioka et al., 2019) | Convolutional Neural Networks | Texture features to detect lesions. | Prognosis | Image |
| (Leung et al., 2021) | Support Vector Machine<br>Lasso Regression<br>SGB<br>Random Forest<br>XGBoost | Demographics<br>Concurrent Medical Illness<br>Medications | Helicobacter Pylori Detection | Alphanumeric |
| (Li et al., 2019) | Convolutional Neural Network (CNN) | Fluorescence spectral images | Diagnosis | Image |
| (Royel et al., 2020) | Designing of a risk prediction Algorithm based on five filtering techniques: Correlation, Information Gain, Gain Ratio, Relief, and Symmetrical Uncertainty | Pre-operative risk factors such as: Symptoms (Abdominal Pain, Nausea, Frequent Vomiting), Diagnostics: (Stomach Lymphoma), Diet and habit: (Spicy and Salted Food, Green Vegetables, Tobaco Status, Daily food) and socio-demographics: (Educational Level, Skin Color, Monthly income) | Survival | Alphanumeric |

| | | | | |
|---|---|---|---|---|
| (Sakai et al., 2018) | Convolutional Neural Network (CNN) | Texture features (images) | Prognosis | Image |
| (Song et al., 2020) | Convolutional Neural Network (CNN) | Pixel-level patches. | Prognosis | Image |
| (Taninaga et al., 2019) | XGBoost<br>Logistic Regression | H. Pylori serology testing<br>chronic atrophic gastritis<br>gastric or duodenal ulcers<br>Barrett's oesophagus and postgastrectomy<br>Sex<br>Age<br>Body Mass Index | Diagnosis | Image / Alphanumeric |
| (Xu et al., 2017) | Support Vector Machine | Five kinds of genes:<br>-Differential pathway genes (DPGs)<br>-Hub genes in differential pathways based on Mutual Information Network (MIN) analysis<br>-Differentially Expressed Genes (DEGs)<br>identified by Significance Analysis of Microarrays (SAM),<br>-Informative genes (DEGs in differential pathways),<br>-Key genes (hub DEGs) | Diagnosis | Alphanumeric |
| (Zhang et al., 2018) | Support Vector Machine | MicroRNAs biomarkers | Diagnosis | Alphanumeric |
| (Zhu et al., 2020) | Gradient-Boosted Decision Trees | Age, Gender, Paraclinical tests: Neutrophil count, lymphocyte count, neutrophil lymphocyte ratio (NLR), hemoglobin (Hb), red cell distribution width (RDW), platelet (Plt), albumin (Alb), alanine transaminase (ALT), total bilirubin (TB), creatinine (Cr), triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), lipoprotein (Lpa), carcinoembryonic antigen (CEA), carbohydrate antigen 125 (CA125), carbohydrate antigen 199 (CA199) and carbohydrate antigen 724 (CA724) | Survival | Alphanumeric |

From this review, we found that most of the authors implement and compare two or more Machine Learning techniques (Jin et al., 2020), and the most popular of them are Convolutional Networks (Gholami et al., 2021; Ishioka et al., 2019; Li et al., 2019; Sakai et al., 2018; Song et al., 2020) and Support Vector Machine (Leung et al., 2021; Xu et al., 2017; Zhang et al., 2018). On the other hand, the authors focused on variables such as demographics, paraclinicals, genes, concurrent medical illness, medications, biomarkers and specific clinical variables such as Helicobacter Pylori serology testing, Chronic Atrophic gastritis, Gastric or duodenal ulcers, that are strongly related to stomach cancer. Some of them include combinations of variables. For instance, (Zhu et al., 2020) analyze both demographics and paraclinical variables, or (Taninaga et al., 2019) include demographics, specific diagnostics related with stomach cancer and body mass index. Nevertheless, we did not find works that evaluate clinical variables such as diagnostics, medications and medical procedures in terms of their predictive capability or estimate the probabilities of being diagnosed with stomach cancer based on those

variables. Our contribution is a Machine Learning approach for prioritization of patients for endoscopy screenings that allow the stomach cancer early detection. For that, we implement a Design Science Research methodology by following the approach proposed by Peffers et al. (2007), explained in Section 4.

## 3. Objectives

To assess the performance of different models aimed at predicting individual probabilities of being diagnosed with stomach cancer, by leveraging data available in the national healthcare system in Colombia (Sistema General de Seguridad Social en Salud de Colombia, SGSSS).

We pursue the following specific objectives:

a. To quantify the relationship between the patient clinical pathway and the diagnosis probability.
b. To assess the predictive power of the variables included in the SGSSS.
c. To assess performance of two classification models in terms of their potential impact on the design of early diagnosis plans.

## 4. Methodology

This section presents a Design Science Research (DSR) approach to prioritize patients that will be recommended to undergo an upper gastrointestinal endoscopy. Table 2 provides an overview as five steps of the steps proposed by Peffers et al. (2007).

*Table 2 - Methodology for Design Science Research (Peffers et al., 2007)*

| Phase | Our Study |
|---|---|
| Problem definition and motivation | Stomach cancer is the main cause of cancer deaths in Colombia and since is usually an asymptomatic disease, the diagnosis takes place in a stage with low or null probabilities of recovery. Early detection and access to treatment are crucial for reduce the risk of death, particularly in low-income and middle-income countries. |
| Proposed solution | To prioritize patients to undergo an upper gastrointestinal endoscopy in order to increase early detection. |
| | We compare two ML techniques to address the following questions:<br>a. How reliably can routinely collected data on patient visits be used to predict probabilities of being diagnosed with stomach cancer?<br>b. Which ML approach performs best, using the AUROC and coverage metrics?<br>c. How might insights obtained from these classification models be used in practice to prioritize patients for endoscopy screening strategies? |
| Demonstration | Performance assessment using the average AUROC score of a 10-by-10 cross-validation |
| Evaluation | Impact on the coverage of an intervention program when a classification algorithm is used |

### 4.1. Problem definition and motivation

Colombia faces important challenges regarding stomach cancer detection and treatment. On one hand, in 2020 stomach cancer was the eight general death cause and the leading cause of cancer death, in Colombia (Departamento Administrativo Nacional de Estadística DANE, 2022). As it is usually an asymptomatic disease, the diagnosis takes place in stages with low or null probabilities of recovery (Barreto et al., 2019). In fact, 73% of stomach cancer cases are diagnosed in a late stage in the country, with a 5-years survival rate of 21%, while

the survival rate for the 27% of cases detected during an early stage increases to 46%(Fondo Colombiano de Enfermedades de Alto Costo, 2022). According this source, the mean elapsed times between the suspicion of stomach cancer and diagnosis is 43.6 days and the first treatment started 55 days after the stomach cancer was diagnosed. Registered cases received not-exclusive treatments: 41% had systemic therapies such as chemotherapy or immunotherapy, 34% had a surgery and 11% had radiotherapy (Cuenta de Alto Costo, 2018).

On the other hand, the national healthcare system does not have monitoring and prevention programs for stomach cancer, since this disease is not prioritized as a public health issue (Oliveros et al., 2019). According to Ministerio De Salud y Protección Social & Instituto Nacional de Cancerología, (2012), the most recent version of the national plan for cancer control in Colombia published in 2012 defines some goals related screening tests for cervix uteri, breast, colorectal, and prostate cancer. However, there are not definitions regarding a similar strategy for stomach cancer.

### 4.2. Proposed solution

In this work, a Machine Learning-enabled DSS to prioritize patients for endoscopy screening is designed. We adopt a DSR approach and aim at predicting individual probabilities of being diagnosed with stomach cancer. Table 3 presents the resulting design principle following the structure provided by Gregor et al. (2020). Within this strategy, the success of endoscopy-based screening programs relies on the quality of the classification. Then, a performance measure that describes the accuracy is defined: coverage. This metric is calculated as the proportion of stomach cancer cases that would be selected for two cut-off points applied to the global population included in the study. These cut-off points are determined by clinical experts in order to assess and compare the performance of the implemented solution techniques.

*Table 3 - Components of the design principle*

| Design principle | Decision Support System proposed |
|---|---|
| Aim, Implementer and user | To prioritize patients for endoscopy screening programs, aimed at prevention and early detection of stomach cancer |
| Context | The national healthcare insurance system in a middle-income country where stomach cancer is the leading cause of cancer mortality |
| Mechanism | Predict individual probabilities of being diagnosed with stomach cancer using Machine Learning techniques |
| Rationale | Implementation of prevention and early detection programs is efficient and timely when patients are classified according to their risk of being diagnosed with stomach cancer. |

### 4.3. Design and development

The first step is the representation of clinical pathways as model process, by implementing Process Mining techniques. We aim at quantifying the relationship between a given care pathway and the diagnosis probability. For the purpose of this study, a patient is considered to have been diagnosed with stomach cancer if there is a record of the disease and a test used for detection. Table 4 shows the information of codes used to tag patient as stomach cancer cases.

*Table 4 - Codes for stomach cancer tagging*

| Code | Type of variable | Name |
|---|---|---|
| **C15** | Diagnostic | Malignant neoplasm of esophagus |
| **C16** | Diagnostic | Malignant neoplasm of stomach |
| **D00** | Diagnostic | In situ neoplasms |
| **89.8.1.01** | Procedure | Study of basic coloration in biopsy |

Hutchison & Mitchell, (n.d.) algorithm was implemented to build up event graphs resulting from Markov Chains. This algorithm identifies and represents a set of clusters using the corresponding Markov Chain with two states: input and output state. Therefore, the assignments of sequences to clusters is based on the probability of each cluster producing the given sequence (Veiga, 2009). We implemented this technique by using the plugin *Sequence Clustering* of *Prom v.5.2* with the parameters shown in Table 5. For this implementation, we kept the default values for event occurrence (percentage) in 0% and 100%, then we assigned as maximum number of events in a sequence the total diagnostics of the clinical pathways analyzed, and the total cases included was set as the maximum sequence occurrence (Veiga, 2009). The total number of cluster is obtained as the root square of the total cases (Veiga, 2009). As a result of this phase, a set of typical care pathways emerging from the data were obtained.

*Table 5 - Parameters for Sequence Clustering plugin*

| Parameter | Value |
|---|---|
| Min event occurrence (percentage) | 0 |
| Max event occurrence (percentage) | 100 |
| Min number of events in a sequence | 1 |
| Max number of events in a sequence | 408 |
| Min sequence occurrence | 1 |
| Max sequence occurrence | 1342 |
| Number of clusters | 36 |

Then, for each patient considered in the study, we obtained a continuous variable that indicates how their clinical pathway meets the process model of each cluster. First, we applied Process Discovery technique that takes an event log and produces a model without using any a-priori information (Van der Aalst, 2012). This technique enabled us to represent each cluster and the clinical pathway of each patient included in the study as process models. We used *Mine Petrinet with inductive Miner* of *Prom v.1.2* and obtained a model process to represent each cluster. Second, a Conformance Checking was applied. This approach is used to compare two process models and determine whether the actual process corresponds to a target process (Jans et al., 2021). We compared each cluster with each clinical pathway to get the *TraceFitness*, a metric with continuous values from 0 to 1, explains how each patient's clinical pathway and each cluster's process model fits, being 0 no fitness and 1 perfect fitness. This output allowed to create a set of variables that were included in the Machine Learning models.

The second step is the quantification of linear relationships between the variables and the individual probabilities of being diagnosed with stomach cancer. To do so, a LASSO regression was implemented. For classification problems, according to Dreiseitl & Ohno-Machado (2002) one of the most popular models in health research is the logistic regression due to its high interpretability. However, Ordinary least squares (OLS) could increase the risk of overfitting, reducing the accuracy of the results (Dreiseitl & Ohno-Machado, 2002). To tackle this issue, LASSO regression minimizes the residual sum of squares and ensures that the sum of the absolute value of the coefficients is less than some chosen value (Tibshirani, 1996). This approach is used frequently to model the relationship between multiple independent variables and a categorical dependent variable, with emphasis on medical research (Boateng & Abaye, 2019) **¡Error! No se encuentra el origen de la referencia.** We performed a parametric analysis on the penalty constant and a 10-by-10 cross validation process. Table 6 provides information of the parameters for the LASSO Regression Model using PySpark Machine Learning Classification LogisticRegression function. Annex 1 presents a flowchart of this algorithm.

*Table 6 - Logistic Regression model parameters*

| Parameter | Value |
|---|---|
| Maximum iterations | 200 |
| Tol | 1e-10 |
| Lambda | 0.005 |
| Train Ratio | 0.5 |

To model non-linear relationships, we implemented Random Forest (RF). RF is a method based on Bagging, that aims to reduce the variance of a statistical model, as well as simulates the variability of data through the random extraction of bootstrap samples from a single training set and aggregates predictions on a new record (Aria et al., 2021). One of the main benefits of implementing random forest models for prediction, is the ability to handle datasets with a large number of predictor variables (Speiser et al., 2019). In addition, random forest models have a high predictive accuracy obtained through non-parametric approaches based on iterative algorithms that generate so-called "black-box" models, which means that their interpretability can be difficult (Aria et al., 2021). For the implementation of the Random Forest, we used PySpark Machine Learning Classification RandomForest Classifier function with different values of the parameters were tested as shown in Table 7. Annex 2 presents a flowchart of this algorithm.

*Table 7 - Random Forest model parameters*

| Parameter | Values tested |
|---|---|
| Number of trees | From 1500 to 2300 (step length = 100) |
| Maximum Depth | 8, 10 |
| Impurity | Gini, Entropy |
| Minimum Instances per node | 1, 2 |

### 4.4. Demonstration and evaluation

Model performance was assessed using the AUROC score. From the database, training (70%) and test (30%) sets were randomly generated. In the demonstration phase, a 10-fold cross-validation process repeated 10 times (10-by-10 CV) was carried out, using the training set. In the evaluation phase, we used the test set to assess the quality of the results. Different patients may be selected to undergo an endoscopy, according to which classification algorithm is used. As discussed in Section 4.2, we use the coverage of an intervention to assess the quality of a given classification. In this context, coverage is the percentage of actual cancer cases that are included in the intervention group, using each of the two possible cut-off points. To quantify the impact of variability on the performance assessment, a simulation was carried out. For each iteration of the simulation, random training and test sets were generated. Then, we trained the algorithms and computed the coverage for each cut-off point.

## 5. Data collection and processing

The data analyzed contains information of all patients who used the health services system, during September and December of 2013, in Colombia. To model the clinical pathways, we used information of a 12-month time window, before the diagnosis. Namely from October 2012 to September 2013 for patients seen during September 2013, and January to December 2013 for patients seen during December 2013 that were not diagnosed during September 2013. The database includes two types of variables. The first group is socio-demographic information: Date of Birth, Gender and Municipality where the service was provided. The second group contains information about the healthcare service. This information is recorded each time a patient access

the health system in Colombia, for billing purposes. We retrieved information of the date of visit, Diagnostic ICD-10 code, Medical Procedure unique code *(Clasificación Única de Procedimientos en Salud, CUPS)*, Medications Unique Code *(Código Único Nacional de Medicamentos, CUM)*, Healthcare Provider Institutions (IPS) and Healthcare Insurer Companies (EPS).

A total of 8,234,601 patients used the health services, during September and December of 2013. As expected, each patient can have more than one visit. Additionally, for each visit, it is possible to have more than one log, since a patient can have more than one diagnostic, medical procedure or medication and each register only can have one code of each mentioned variable. Therefore, the database has 41,152,616 logs. From this population, 305 patients were diagnosed with stomach cancer. As shown in Table 8 the distribution of cancer cases is 131 women (43%) and 174 men (57%). They represent the 0.0079% of the population of the study. Also, we can see that the prevalence of stomach cancer cases is the highest in the patients of the category >60 years old, and is the lowest for the category <30 years old.

*Table 8 - Descriptive statistics*

| Variable | Category (years) | Stomach Cancer Cases | | Controls |
|---|---|---|---|---|
| Age | <30 | 2 | 0.0001% | 3,275,995 |
| | 30-49 | 53 | 0.0023% | 2,301,198 |
| | 50-60 | 70 | 0.0060% | 1,158,101 |
| | >60 | 180 | 0.0120% | 1,499,002 |
| Gender | Female | 131 | 0.0027% | 4,937,987 |
| | Male | 174 | 0.0053% | 3,296,309 |

During the time window, 1,809 diagnostics, 5172 medical procedures and 703 medications were detected for all the population included. It is important to take into account that the diagnostics, medical procedures and medications considered for the evaluation of predictive capabilities of variables exclude some values because they correspond or are used to diagnose stomach cancer. This review was done by the clinical team of the project. Table 9 shows the amount of input and output variables for each stage.

*Table 9 - List of variables included in the Logistic Regression Models*

| Set of variables | Variables included | | Significant variables (Correlation) | | |
|---|---|---|---|---|---|
| | Initial input | First LASSO model | Final LASSO model | Positive | Negative |
| Diagnostics | 1798 | 59 | 10 | 8 | 2 |
| Medical procedures | 5162 | 108 | 40 | 37 | 3 |
| Medications | 691 | 62 | 21 | 11 | 10 |
| Age (Bins) | N/A | 4 | 2 | 0 | 2 |
| Gender | N/A | 2 | 0 | 0 | 0 |
| Municipality | N/A | 80 | 8 | 7 | 1 |
| Process Model (Bins) | N/A | 36 | 12 | 5 | 7 |

In addition, the ICD-10 codes that are integrated by 4 digits were grouped by using the 3 first digits. This decision was made after a computational experimentation in which we implemented a Logistic Regression for variables selection and found that creating groups of diagnostics following the coding logic added value in terms of the significance of the variables, compared with running the same algorithm with the codes of 4.

## 6. Results

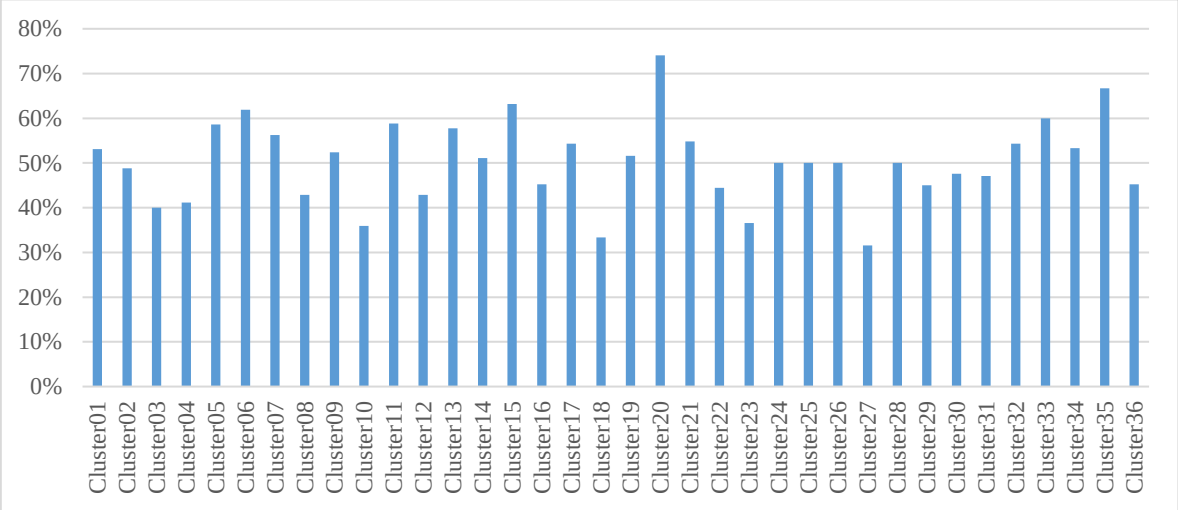
Results are organized in three sections. First, the process modeling and variables obtained from the Process

Mining approach are shown. Second, we quantify the impact of each variable in the LASSO regression model on the probability of being diagnosed with stomach cancer. Finally, we present a comparison of Machine Learning techniques used for classification of patients in terms of coverage.

### 6.1. Process Mining approach

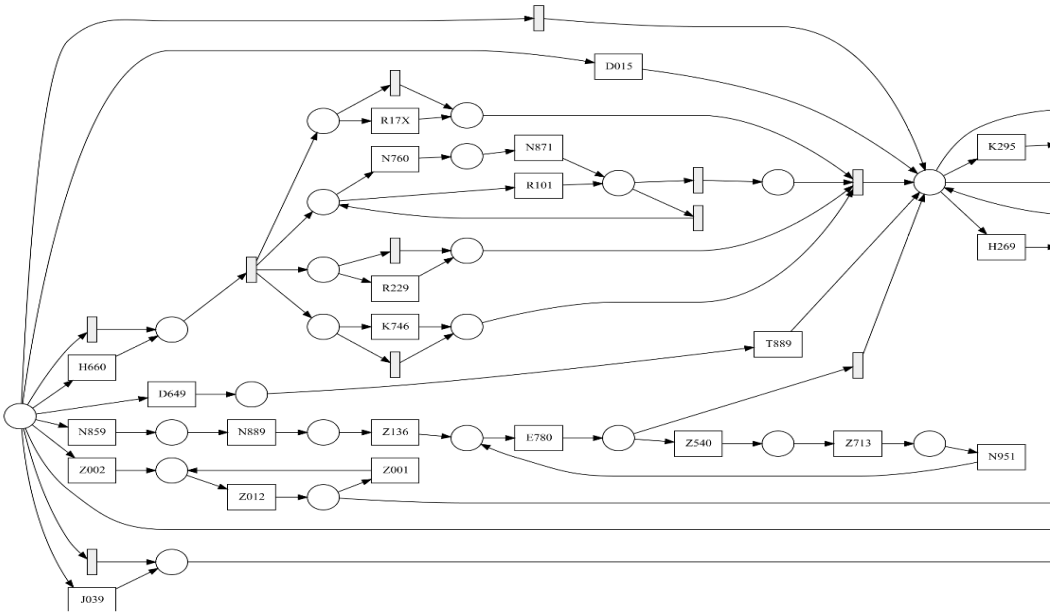
This work implemented Process Mining techniques to analyze the clinical pathways of diagnostics from a model process approach. For this purpose, 1342 cases were obtained and analyzed (671 stomach cancer cases detected in patients seen not only in September and December but during 2013, and 671 no stomach cancer cases selected randomly from the same year). These clinical pathways were grouped in 36 clusters according the similarity between them. Each cluster groups stomach cancer cases and no stomach cancer cases. Figure 1 shows the percentage of stomach cancer cases assigned to each cluster.

*Figure 1 - Percentage of stomach cancer cases per cluster*




Then, a Process Discovery (Van der Aalst, 2012) was implemented to create a process model based on the clinical information (logs) of the patients assigned to each cluster. This technique allows to represent the diagnostic logs of the cases grouped in the cluster as a Petrinet model process. An example of a typical Petrinet Chart is shown in Figure 2. From this chart, we can see that the clinical pathway can follow different sequences of diagnostics that have a specific order. The model shown presents paths of ICD-10 diagnostic codes, with different complexity levels. Some are easy to interpret and follow: if the first diagnostic of a patient is D015, the second will be K295 or K269, or if the first diagnostic is D649, the following in their clinical pawthay will be T889. Some others are more complex: if a patient starts with the diagnostic H660, it will be followed by R17X, N760, R229 or K746, but also a pathway can start in some of these diagnostics.

*Figure 2 - Process Model*




Next, the clinical pathway of each patient seen during September and December of 2013 was used to perform a Conformance Checking (Jans et al., 2021) to find how the clinical pathway of each of them fitted with those general clinical pathways. For each patient, a continuous variable was obtained. The values of this variable are between 0 and 1, where 0 means that the clinical pathway of the patient has no fitness with the general model, and 1 means a perfect fitness between the two clinical pathways. Finally, bins for each variable were created to turn those continuous variables in binary variables that could be included in the Machine Learning models. Annex 3 presents a sample of 10 patients and the values of their *TraceFitness* with the process models generated.

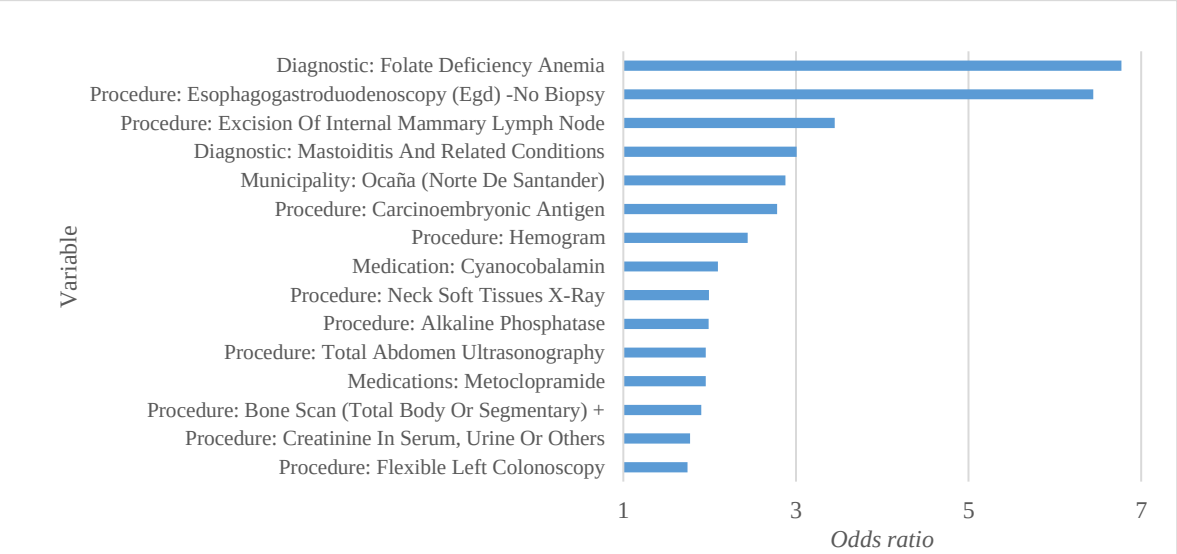
### 6.2. LASSO results

According to the expert validation done by the clinical team of the project, the significant variables can be grouped into four categories: infections, anemia, malnutrition and thromboembolic events. Patients that are in initial stages of the cancer usually present immunosuppression signs (Matsueda & Graham, 2014). As a consequence, those patients are susceptible to develop infections. About anemia, (American Cancer Society, 2022) has set this disease as an early manifestation of cancer, which means that is coherent to have this group of variables. Regarding malnutrition, (Heneghan et al., 2015) malabsorption is related with gastrointestinal disorders in general. Finally, thromboembolic events are strongly related with cancer since malignant cells are capable of activating the coagulation cascade and other prothrombotic properties of host cells (Razak et al., 2018). Regarding clinical variables with negative correlation, we found medications such as Beclomethasone, Diclofenac, Ibuprofen, Thiamine (Vitamine B1), Chloroquine, Levothyroxine Sodium, Methocarbamol, Naproxen, Nifedipine, Loratadine that are used to treat other general diseases. Also, we found that Bronchitis as well as procedures such as Vaginal Cytology or Nuclear Antibodies tests.

Regarding demographics, it was found that the gender is not a significant aspect compared with the impact of other variables, since none of the values of this variable (Male, Female) has correlation positive or negative with the fact of being diagnosed with stomach cancer. In addition, 7 municipalities have a positive correlation. Ocaña (Norte de Santander), Pasto (Nariño), La Ceja (Antioquia), Santander de Quilichao (Cauca), Calima (Valle del Cauca) and Copacabana (Antioquia) are located departments of higher risk according to (Pardo et al., 2017), due to there are mountain regions where the strains of *Helicobacter pylori* represent higher risk of developing stomach cancer. On the other hand, Chocó (Quibdó) also reported a correlation positive in spite of belonging to a coast zone. This is an interesting fact to explore from the socio-economic conditions of the

region, since they can impact the accessibility and timeliness of the healthcare system. Negative correlation applies for ages groups of 0 to 37 years and 37 to 54 years, which make sense with the risk factor of age: in spite of stomach cancer can occur in young people, risk goes up as the person gets older (American Cancer Society, 2022). Regarding the municipality, Barranquilla (Atlántico) was the only city that present a negative correlation, which is related to the findings of (Pardo et al., 2017) who state that coast regions report lower rates of incidence of this disease. Figure 3 shows the 15 variables with the greatest impact in the probability of being diagnosed with cancer.

*Figure 3 - Odds ratio for the 15 variables with the greatest impact*

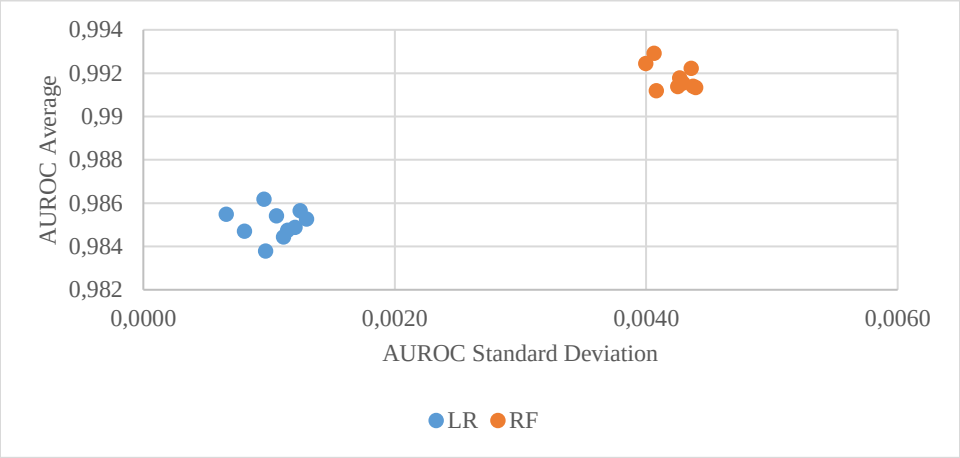



From the side of process models, we obtained as significant variables with positive correlation: Trace Fitness between 0.94 and 0.97 for Model 13 have a positive correlation with the fact of being diagnosed with cancer. This make sense since the most of patients included in this clusters are stomach cancer cases, as shown in **¡Error! No se encuentra el origen de la referencia.**. In contrast, Trace Fitness between 0 and 0.79 with Model 30 and Trace Fitness between 0 and 0.76 with Model 02. Since those models are mainly integrated for no cancer cases, which can be understood as low fitness with these clinical pathways increases the risk of having a clinical pathway that could lead to a stomach cancer diagnosis. Also, the model suggested a positive correlation for the Trace Fitness between 0 and 0.8 with Model 5, and Trace Fitness between 0 and 0.95 for Model 1. However, these models are mostly integrated by patients with stomach cancer. On the other hand, the model suggested a negative correlation for Trace Fitness between 0.92 and 0.98 with model 30. As mentioned before, this model is mostly composed by no cancer cases. This make sense since a high Trace Fitness with this no cancer clinical pathway suggest a low risk of being diagnosed with cancer. Annex 4 presents all the significant variables with their corresponding Odds Ratio value.

### 6.3. Machine Learning techniques comparison

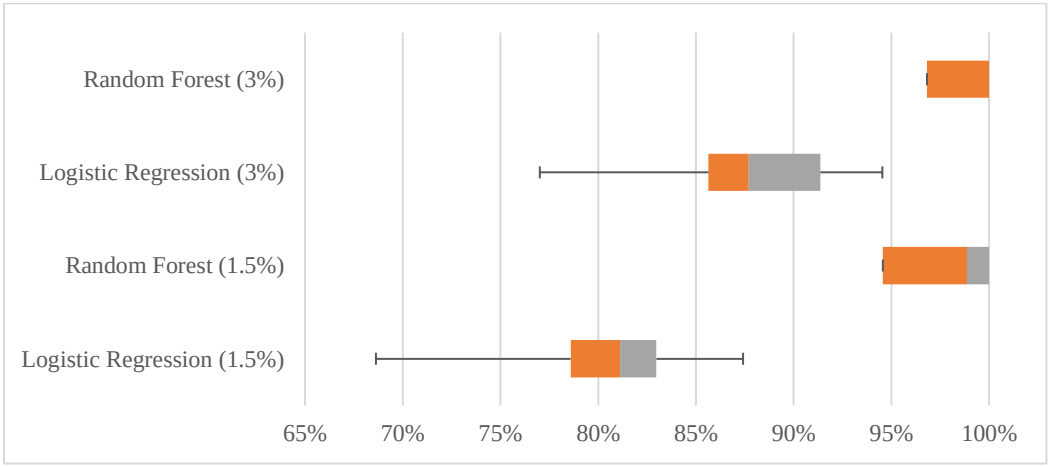
Models performance was evaluated by using the AUROC. The results are shown in Figure 4. Each point in the graph represents the average and standard deviation of the AUROC in a repetition of 10-by-10 cross-validation. In all the cases, the Random Forest reported a better performance than the Logistic Regression. However, for both models, all the AUROC obtained could be considered 'too good to be true' since all the values are included in the range between 0.982 and 0.994. This performance of the AUROC is explained by the fact that the database is highly unbalanced: the number of stomach cancer cases is 305 (0.0037%) of the total of patients that is 8,234,601.

*Figure 4 - Models performance*




In addition, a simulation was implemented with 40 training and test set generated randomly without cross-validating in order to get a proper statistical sample size for estimate an average coverage in Confidence Intervals of 95% and use them as an additional performance measure. For the results obtained for each dataset in each Machine Learning model, patients were sorted according their individual probability of being diagnosed with stomach cancer from the highest to the lowest value. Then, the coverage was estimated depending on two cut-off points: 1.5% and 3% of the population seen during September and December 2013. Figure 5 presents the results of the simulation in terms of coverage.

*Figure 5 - Coverage Simulation results*




According to the results of the simulation, the Random Forest reports a better performance. On one hand, the deviation of the coverage percentages is higher in the Logistic Regression for both cut-off points, which means more variation of the results depending on the dataset, reducing the confidence on the results obtained. On the other hand, the coverage of patients that actually were stomach cancer cases would be greater if we implement the Random Forest Model, comparing it with the Logistic Regression Model for both cut-off points. In addition, we found that the variation of coverage between the two cut-off points is higher for the Logistic Regression model than for Random Forest model. It means that if the intervention program is designed to perform endoscopy screenings to the 1.5% of the global patients according to the Random Forest Model, the coverage would be of 97.4% of the patients that finally would be stomach cancer cases, and if we duplicate the population

to be intervened, the program will just increment 1.2 percentage points. It means that using the cut-off point in 3% would duplicate the costs but the increases in coverage would not be proportional. From the perspective of the Logistic Regression, the variation between the two cut-off points would be of 7.7 percentage points in coverage by duplicating costs, which means that using the cut-off point in 3% is not cost-effective. Annex 5 presents the results of the computational experimentation implemented to get the coverage metric.

## 7. Conclusions and future work

This work is oriented to propose a Decision Support System for the prioritization of patients for endoscopy screening programs, considering the needs mentioned regarding this testing for timely detection of stomach cancer. In this context, the Machine Learning approach has demonstrated to be useful in terms of evaluation of predictive capability of the variables, as well as classification of patient according their individual probability of being diagnosed with stomach cancer.

First, Process Mining approach demonstrated a comprehensive way to face this challenge. From the understanding of the clinical pathways of the patients as a model process, this discipline offers techniques such as Process Discovery and Conformance Checking that helps to convert valuable logs in flowcharts that can provide an easy way to find relations between clinical variables and set general clinical pathways that generate alerts on the probability of diagnose a patient with stomach cancer. Eventually, this techniques could be applied to discover processes in order to create care models oriented to prevention actions for keep the low-risk population in this classification.

Second, the implementation of a Logistic Regression model for the evaluation of the predictive capability of the variables derived of the SGSSS showed consistent results from the clinical point of view. Considering that the time window is very close to the moment when the diagnostic of stomach cancer was confirmed, the model determined significant variables strongly related with characteristic conditions of the patients. It means that the model proved its capability to generate results aligned with aspects that have been found by other researches and already adopted in the daily context of clinical attention. An implementation of this model considering other time windows larger than a year and or more distant from the moment of the diagnostic is suggested

Third, the implementation of two Machine Learning techniques with different advantages and challenges allowed comparison. Logistic Regression models that offer a high interpretability but has some constraints regarding large and unbalanced datasets in terms of accuracy, and Random Forest models that are very appropriate for huge volumes of information with accuracy, but operates as a black box, which means that the explicability is not easy. The best performance was reported by the Random Forest. For the context of the problem, the accuracy and the capability to process large datasets are crucial. Hence, it could be interesting to implement and compare the performance of another Machine Learning methods aimed to meet these characteristics, even if they have constraints regarding interpretability. Other interesting scenarios to be studied in future works could be the implementation of other Machine Learning techniques to evaluate their performance

From the point of view of the implementation and evaluation of machine learning classification techniques, we found that it is possible to design cost-effective intervention programs. One of the main barriers in the implementation of endoscopy screening strategies is the high cost that it means for a healthcare system that results infeasible for lower-income and middle-income countries, where resources usually are invested in the solution of problems considered more urgent (Oliveros et al., 2019). The Decision Support System designed aims to take advantage on the Random Forest capabilities of classification to select the patients to be eligible for endoscopy screenings based on determine cut-off points that ensure cost-effective coverage. For the dataset used for this work and the conditions in which it was performed, the best solution is to select the 1.5% of the population with the highest probabilities to be diagnosed with stomach cancer because the program would reach an estimated coverage of 98.6% of patients that actually were diagnosed with this disease. Hence, our DSS proposes to implement this technique for datasets of interest in a real context of selection of patients to access to endoscopy screenings.
The research presented here contributes to the literature by assessing the effectiveness of machine learning

approaches using routine data to classify patients in terms of their probabilities of being diagnosed with stomach cancer, in the context of a middle-income country. The approach proposed in our work could help to other designers in healthcare matters, as well as in other fields where limited resources are challenging and prioritization is fundamental in terms of feasibility.

## 8. References

Ali, H., Yasmin, M., Sharif, M., & Rehmani, M. H. (2018). Computer assisted gastric abnormalities detection using hybrid texture descriptors for chromoendoscopy images. *Computer Methods and Programs in Biomedicine*, *157*, 39–47. https://doi.org/10.1016/j.cmpb.2018.01.013

American Cancer Society. (2022). *Stomach Cancer Causes , Risk Factors , and Prevention*. Cancer ORG. https://www.cancer.org/content/dam/CRC/PDF/Public/8839.00.pdf

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. *Machine Learning with Applications*, *6*(April), 100094. https://doi.org/10.1016/j.mlwa.2021.100094

Balakrishnan, M., George, R., Sharma, A., & Graham, D. Y. (2017). Changing Trends in Stomach Cancer Throughout the World. *Current Gastroenterology Reports*, *19*(8). https://doi.org/10.1007/s11894-017-0575-8

Barreto, C., Limas, L., Porras, A., & Rico, A. (2019). *Carga de enfermedad de cáncer gástrico durante los años 2010 y 2019 en Tunja, Boyacá, Colombia* (Vol. 2020, Issue 1) [Universidad del Bosque]. https://repositorio.unbosque.edu.co/bitstream/handle/20.500.12495/6881/Barreto_Noratto_Clara_Patricia_2022.pdf?sequence=4&isAllowed=y

Boateng, E. Y., & Abaye, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, *07*(04), 190–207. https://doi.org/10.4236/jdaip.2019.74012

Broeders, M., & Elfström, K. M. (2020). Importance of International Networking and Comparative Research in Screening to Meet the Global Challenge of Cancer Control. *JCO Global Oncology*, *6*, 180–181. https://doi.org/10.1200/JGO.19.00388

Camargo, M., Wiesner, C., Diaz, M., & Tovar, S. (2004). *El cancer. Aspectos Basicos sobre su Biología, prevencion, diagnostico y tratamiento*. 67.

Cazap, E., de Almeida, L. M., Arrossi, S., García, P. J., Garmendia, M. L., Gil, E., Hassel, T., Mayorga, R., Mohar, A., Murillo, R., Owen, G. O., Paonessa, D., Santamaría, J., Tortolero-Luna, G., Zoss, W., Herrero, R., Luciani, S., Schüz, J., & Espina, C. (2019). Latin America and the Caribbean Code Against Cancer: Developing Evidence-Based Recommendations to Reduce the Risk of Cancer in Latin America and the Caribbean. *Journal of Global Oncology*, *5*, 1–3. https://doi.org/10.1200/JGO.19.00032

Chaudhuri, N., & Bose, I. (2020). Exploring the role of deep neural networks for post-disaster decision support. *Decision Support Systems*, *130*(January), 113234. https://doi.org/10.1016/j.dss.2019.113234

Cuenta de Alto Costo. (2018). *Situación del Cancer en la población adulta atendida en el SGSSS de Colombia*.

Departamento Administrativo Nacional de Estadística DANE. (2022). *Estadísticas Vitales - Defunciones*. http://systema74.dane.gov.co/bincol/RpWebEngine.exe/Portal?BASE=DEFOC0820&lang=esp

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Fondo Colombiano de Enfermedades de Alto Costo. (2022). *El objetivo en cáncer en Colombia: igualdad en*

*la atención en salud priorizando la detección temprana*. Cuenta de Alto Costo. https://cuentadealtocosto.org/site/cancer/el-objetivo-en-cancer-en-colombia-igualdad-en-la-atencion-en-salud-priorizando-la-deteccion-temprana/

Gholami, E., Kamel Tabbakh, S. R., & kheirabadi, M. (2021). Increasing the accuracy in the diagnosis of stomach cancer based on color and lint features of tongue. *Biomedical Signal Processing and Control*, *69*(May 2020), 102782. https://doi.org/10.1016/j.bspc.2021.102782

Gómez, M. A., Riveros Vega, J. H., & Otero, W. (2015). Cáncer gástrico temprano vs. avanzado: ¿existen diferencias? *Revista de La Universidad Industrial de Santander. Salud*, *47*(1), 7–13.

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems*, *21*(6), 1622–1652. https://doi.org/10.17705/1jais.00649

Heneghan, H. M., Zaborowski, A., Fanning, M., McHugh, A., Doyle, S., Moore, J., Ravi, N., & Reynolds, J. V. (2015). Prospective Study of Malabsorption and Malnutrition After Esophageal and Gastric Cancer Surgery. *Annals of Surgery*, *262*(5), 803–808. https://doi.org/10.1097/SLA.0000000000001445

Huang, Y., Zhu, J., Li, W., Zhang, Z., Xiong, P., Wang, H., & Zhang, J. (2018). Serum microRNA panel excavated by machine learning as a potential biomarker for the detection of gastric cancer. *Oncology Reports*, *39*(3), 1338–1346. https://doi.org/10.3892/or.2017.6163

Hutchison, D., & Mitchell, J. C. (n.d.). *Lecture Notes in Computer Science*.

Ishioka, M., Hirasawa, T., & Tada, T. (2019). Detecting gastric cancer from video images using convolutional neural networks. *Digestive Endoscopy*, *31*(2), e34–e35. https://doi.org/10.1111/den.13306

Jans, M., Weerdt, J. De, Depaire, B., Dumas, M., & Janssenswillen, G. (2021). Conformance Checking in Process Mining. *Information Systems*, *102*, 101851. https://doi.org/10.1016/j.is.2021.101851

Jin, P., Ji, X., Kang, W., Li, Y., Liu, H., Ma, F., Ma, S., Hu, H., Li, W., & Tian, Y. (2020). Artificial intelligence in gastric cancer: a systematic review. *Journal of Cancer Research and Clinical Oncology*, *146*(9), 2339–2350. https://doi.org/10.1007/s00432-020-03304-9

Kailin, J., Xiaotao, J., Jinglin, P., Yi, W., Yuanchen, H., Senhui, W., Shaoyang, L., Kechao, N., Zhihua, Z., Shuling, J., Peng, L., Peiwu, L., & Fengbin, L. (2021). Current Evidence and Future Perspective of Accuracy of Artificial Intelligence Application for Early Gastric Cancer Diagnosis With Endoscopy: A Systematic and Meta-Analysis. *Frontiers in Medicine*, *8*(March), 1–11. https://doi.org/10.3389/fmed.2021.629080

Leung, W. K., Shing, K., Bofei, C., Lui, T. K. L., & Law, S. Y. K. (2021). *Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication*. *November 2020*, 864–872. https://doi.org/10.1111/apt.16272

Li, Y., Deng, L., Yang, X., Liu, Z., Zhao, X., Huang, F., Zhu, S., Chen, X., Chen, Z., & Zhang, W. (2019). Early diagnosis of gastric cancer based on deep learning combined with the spectral-spatial classification method. *Biomedical Optics Express*, *10*(10), 4999. https://doi.org/10.1364/boe.10.004999

Mallavarapu, T., Hao, J., Kim, Y., Oh, J. H., & Kang, M. (2020). Pathway-based deep clustering for molecular subtyping of cancer. *Methods*, *173*, 24–31. https://doi.org/https://doi.org/10.1016/j.ymeth.2019.06.017

Matsueda, S., & Graham, D. Y. (2014). *Immunotherapy in gastric cancer*. *20*(7), 1657–1666. https://doi.org/10.3748/wjg.v20.i7.1657

Ministerio De Salud y Protección Social, & Instituto Nacional de Cancerología. (2012). Plan nacional para el control del cáncer en Colombia 2012-2020. *Ministerio De Salud Y Protección Social*, 1–85. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/IA/INCA/plan-nacional-control-

cancer-2012-2020.pdf

Niu, P., Zhao, L., Wu, H., Zhao, D., & Chen, Y. (2020). *Artificial intelligence in gastric cancer: Application and future perspectives*. *26*(36), 5408–5419. https://doi.org/10.3748/wjg.v26.i36.5408

Oliveros, R., Pinilla, R. E., Navia, H. F., & Oliveros, R. (2019). Gastric cancer is a preventable disease: Strategies for intervention in its natural history. *Revista Colombiana de Gastroenterologia*, *34*(2), 177–189. https://doi.org/10.22516/25007440.394

Pardo, C., De Vries, E., Buitrago, L., & Gamboa, Ó. (2017). Atlas de mortalidad por cáncer en Colombia. Cuarta edición. In *Instituto Nacional de Cancerología* (Vol. 1). https://www.cancer.gov.co/ATLAS_de_Mortalidad_por_cancer_en_Colombia.pdf

Pasechnikov, V., Chukov, S., Fedorov, E., Kikuste, I., & Leja, M. (2014). Gastric cancer: Prevention, screening and early diagnosis. *World Journal of Gastroenterology*, *20*(38), 13842–13862. https://doi.org/10.3748/wjg.v20.i38.13842

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Piñeros, M., Laversanne, M., Barrios, E., Cancela, M. de C., de Vries, E., Pardo, C., & Bray, F. (2022). An updated profile of the cancer burden, patterns and trends in Latin America and the Caribbean. *The Lancet Regional Health - Americas*, *13*, 100294. https://doi.org/10.1016/j.lana.2022.100294

Razak, N. B. A., Jones, G., Bhandari, M., Berndt, M. C., & Metharom, P. (2018). Cancer-Associated Thrombosis: An Overview of Mechanisms, Risk Factors, and Treatment. *Cancers*, 1–21. https://doi.org/10.3390/cancers10100380

Ro, T. H., Mathew, M. A., & Misra, S. (2015). Value of screening endoscopy in evaluation of esophageal, gastric and colon cancers 2015 Advances in Gastrointestinal Endoscopy. *World Journal of Gastroenterology*, *21*(33), 9693–9706. https://doi.org/10.3748/wjg.v21.i33.9693

Royel, R. I., Jaman, A., Masud, F. Al, Ahmed, A., & Ahmed, K. (2020). *Machine Learning and Data Mining Methods in Early Detection of Stomach Cancer Risk*. *24*(1), 1–8.

Sakai, Y., Takemoto, S., Hori, K., Nishimura, M., Ikematsu, H., Yano, T., & Yokota, H. (2018). *Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network*. 4138–4141.

Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., Ding, X., Liu, J., Yu, C., Ku, C., Liu, C., Sun, Z., Xu, G., Wang, Y., Zhang, X., … Shi, H. (2020). Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/s41467-020-18147-8

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Strasser-Weippl, K., Chavarri-Guerra, Y., Villarreal-Garza, C., Bychkovsky, B. L., Debiasi, M., Liedke, P. E. R., Soto-Perez-de-Celis, E., Dizon, D., Cazap, E., de Lima Lopes, G., Touya, D., Nunes, J. S., Louis, J. S., Vail, C., Bukowski, A., Ramos-Elias, P., Unger-Saldaña, K., Brandao, D. F., Ferreyra, M. E., … Goss, P. E. (2015). Progress and remaining challenges for cancer control in Latin America and the Caribbean. *The Lancet Oncology*, *16*(14), 1405–1438. https://doi.org/https://doi.org/10.1016/S1470-2045(15)00218-1

Taninaga, J., Nishiyama, Y., Fujibayashi, K., & Gunji, T. (2019). Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data : A case- control study.

*Scientific Reports*, *August*, 1–9. https://doi.org/10.1038/s41598-019-48769-y

The Global Cancer Observatory. (2020). *GLOBOCAN 2020: International Agency Research on Cancer*. *509*, 1–2.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van der Aalst, W. (2012). *Process Mining : Overview and Opportunities*. *99*(99). https://doi.org/10.1145/0000000.0000000

Veiga, G. M. (2009). *Developing Process Mining Tools: An Implementation of Sequence Clustering for ProM*. *September*.

Villamil, M. D. P., Barrera, D., Velasco, N., Bernal, O., Fajardo, E., Urango, C., & Buitrago, S. (2017). Strategies for the quality assessment of the health care service providers in the treatment of Gastric Cancer in Colombia. *BMC Health Services Research*, *17*(1), 1–16. https://doi.org/10.1186/s12913-017-2440-8

World Health Organization.WHO. (2020a). *Cancer*. Cancer.

World Health Organization.WHO. (2020b). *Cancer Today*. International Agency For Research Cancer. https://gco.iarc.fr/today/home

Xu, Y. G., Cheng, M., Zhang, X., Sun, S. H., & Bi, W. M. (2017). Mutual information network-based support vector machine strategy identifies salivary biomarkers in gastric cancer. *Journal of B.U.ON.*, *22*(1), 119–125.

Zhang, F., Xu, W., Liu, J., Liu, X., Huo, B., Li, B., & Wang, Z. (2018). Optimizing miRNA-module diagnostic biomarkers of gastric carcinoma via integrated network analysis. *PLoS ONE*, *13*(6), 1–12. https://doi.org/10.1371/journal.pone.0198445

Zhu, S., Dong, J., Zhang, C., Huang, Y., & Id, W. P. (2020). *Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics*. 1–13. https://doi.org/10.1371/journal.pone.0244869