Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.5780

Characterizing the Socioenvironmental and Behavioral Profile of Individuals with OCD Using the PNS 2019 Database

Anna Puga Campos Rodrigues [**Pontifícia Universidade Católica de Minas Gerais (PUC Minas)** | ***annapugac@gmail.com***]
Luis Enrique Zárate [**Pontifícia Universidade Católica de Minas Gerais (PUC Minas)** | ***zarate@pucminas.br***]

Instituto de Ciências Exatas e Informática, Pontifícia Universidade Católica de Minas Gerais (PUC Minas), Av. Dom José Gaspar 500, Belo Horizonte, Minas Gerais, 30535-610, Brazil.

Received: 30 March 2025 • **Published:** 13 March 2026

Abstract The objective of this study is to characterize the profile of individuals diagnosed with Obsessive-Compulsive Disorder (OCD) in the Brazilian population, considering socioenvironmental and behavioral aspects. For this purpose, the 2019 National Health Survey (PNS) database is considered. Based on a knowledge discovery process, including conceptual modeling of the domain for conceptual selection of attributes, the *Explainable Boosting Machine* (EBM) and Decision Tree algorithms are applied, aiming to identify relevant attributes for the classification of OCD. The results indicate that both aspects improve the model's performance, reaching an average *F1-score* of 63% (59% for OCD = yes, and 66% for OCD=No). Results consistent with the literature were also found, such as the relationship between OCD and poor sleep quality, diet quality, and mental disorders such as anxiety and depression, among other factors. This study has limitations, such as the use of data that may not accurately reflect socioeconomic and behavioral conditions during the development of OCD. Thus, this study serves as an exploratory guide, capable of identifying profiles more vulnerable to triggers of the disorder, but without the intention of replacing medical or psychological evaluation.

Keywords: Machine learning, Health informatics, Obsessive-Compulsive Disorder, Health database

1 Introduction

OCD (*Obsessive-Compulsive Disorder*), as established by the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) (American-Psychiatric-Association [2013]), published by the *American Psychiatric Association*, is characterized by the presence of obsessions, compulsions, or both in the daily life of those diagnosed with the disorder. According to the DSM-5 manual, obsessions are defined as recurrent thoughts and impulses that the individual consciously attempts to suppress or neutralize through other thoughts and actions. These attempts to mitigate impulses are characterized by compulsions, which are repetitive behaviors or mental acts performed in response to an obsession or a perceived obligation to act according to rigid and immutable rules. This state of mental distress impairs social, professional, and other aspects of one's life.

According to the World Health Organization (WHO), OCD has a prevalence of 1% to 3% in the global population (Baland *et al*. [2023]) and is classified among the 10 most disabling diseases worldwide. In Brazil, it is estimated that approximately four million people suffer from the disorder.

Based on a literature review in the clinical context, it was observed that efforts to understand OCD are directed toward three main areas: 1) understanding brain function from a chemical and biological perspective, 2) linking symptoms to other psychiatric disorders, and 3) analyzing the individual's behavioral aspects (Brander *et al*. [2016],Hiranandani *et al*. [2023]). The literature review revealed that most investigations on the topic focus on the first area, where classification approaches have been used to identify OCD patients based on brain imaging (Hu *et al*. [2016], Kirsten *et al*. [2021]). The second area typically involves analyzing overlapping OCD symptoms with other disorders to reduce misdiagnosis (Højgaard *et al*. [2023]). The third approach, widely considered by researchers, has primarily focused on a single behavioral aspect of individuals previously diagnosed with OCD (Segalàs *et al*. [2021], Boger *et al*. [2020]).

The literature review also revealed that there is still a limited number of studies focused on analyzing the social profile of individuals with OCD, including the socio-environmental characteristics of their lives. Some studies have provided evidence of the correlation between traumatic events and social factors. For example, Brander *et al*. [2016] identified potential risk factors such as perinatal complications, the reproductive cycle, and stressful life events associated with the onset of the disorder. These studies highlight the characterization of OCD as a complex and multifactorial problem.

In Data Science, studies employing Machine Learning (ML) techniques to investigate mental disorders and health conditions have become common. For example, Souza *et al*. [2020] explores *Deep Learning* techniques to develop a classifier for the automatic identification of depression, anxiety, and comorbidities. It was also observed that little attention had been given to OCD and its factors within the ML domain, where the main contributions are relatively more recent: Kirsten *et al*. [2021], Clemmensen *et al*. [2022], Patel *et al*. [2023], and Zaboski *et al*. [2024].

Acknowledging ML's limited contribution to understanding OCD, this study aims to expand the field by identifying socioenvironmental factors. Its goal is to determine whether these factors impact the diagnosis of OCD and, if the answer

is affirmative, which factors are the most relevant. Additionally, the study seeks to determine whether behavioral factors described in the literature as associated with OCD are present in the characterization of this profile within the Brazilian population.

For this study, we consider two interconnected profiles: a behavioral profile, which includes a set of characteristics such as food and beverage consumption habits, exercise habits, sleep quality, prescription medication use, and the socioenvironmental profile which includes housing and workplace vulnerability, reported violence, and mental disorders, such as distress, excessive fear, sadness, and anxiety.

To better understand the behavioral and socioenvironmental profile of individuals with OCD, this study applies a detailed knowledge discovery process KDD to identify the main factors affecting the disorder in the Brazilian population. As learning model will be considered supervised techniques to characterize that profiles. To this end, the most recent National Health Survey (PNS) conducted by IBGE (IBGE [2020]) is considered. This study collected data on the health status and lifestyle of the Brazilian population in 2019, comprising 1,087 attributes and 293,727 records. The first model considers only attributes in the behavioral category, while the second model includes attributes from both behavioral and socioenvironmental categories. The goal is to evaluate the performance of both models and primarily to understand whether the addition of socioenvironmental characteristics strengthens or weakens the evaluation metrics, thus contributing to the diagnosis of the disorder. A secondary goal is to know whether the behavioral characteristics, when used without the interference of socioenvironmental factors, align with results found in the literature. These research directions were established based on the observations and knowledge of interviewed mental health experts. Considering the Brazilian population, this study seeks to reveal behavioral patterns and socioenvironmental specificities that may be unique to this population and not necessarily affect individuals with OCD in other countries. Thus, it is expected to contribute to a deeper and more contextualized understanding of OCD in Brazil.

As a key step in the adopted methodology, this study introduces a preliminary phase focused on understanding the problem domain to construct a conceptual model (CM) that will guide the process of selecting relevant attributes. The model is developed based on explicit knowledge (from the literature) and tacit knowledge (through interviews with domain experts). This stage aims to identify various dimensions (perspectives) and socioenvironmental factors that may contribute to the development of OCD in an individual.

To gather experiential knowledge, systematic interviews with mental health professionals were conducted. During these interviews, direct questions were asked about identifying the main dimensions and aspects related to the disorder to gain a deeper understanding of the clinical approach to diagnosis. These interviews provided diverse insights into the problem and how each expert perceives the symptoms and triggers of OCD. Based on this, a conceptual model was created to integrate all perspectives into a single visualization to guide the knowledge discovery process, considering the disorder's specific aspects. This procedure introduced a new step to the pipeline of a typical knowledge discovery process: the conceptual attribute selection. This step was necessary to address the high dimensionality of the PNS 2019 database.

To describe the socioenvironmental profile of individuals with OCD, a supervised learning model was constructed. The study considers individuals who indicated during the PNS survey whether they had been diagnosed with OCD or not. Using the PNS data, transformations were applied to derive new attributes with greater informational power related to the aspects highlighted by the conceptual model, thus creating a new dataset for the study. Next, the *Explainable Boosting Machine* (EBM) algorithm (The InterpretML Contributors [2023]) was used to interpret the importance of each attribute in the classification. Finally, the rules generated by the Decision Tree via the *KNIME* platform (KNIME [2023]) were evaluated to check for the presence of behavioral patterns mentioned in the literature and to determine whether socioenvironmental factors might influence the identification of the disorder.

Therefore, this study encompasses both the second and third research approaches mentioned earlier, as it correlates the presence of other mental disorders with OCD and uses socioenvironmental and behavioral characteristics for the classification of the disorder. This work is an extension of the article published in the 39th Brazilian Symposium on Data Bases Rodrigues and Zárate [2024].

In addition to this introduction, this paper is structured into four main sections: the literature review, which discusses the main research lines on OCD; the methodology, which describes the knowledge discovery process adopted to identify the profile of individuals suffering from the disorder; and the experiments and results analysis section, which considers the EBM and Decision Tree models. Finally, the conclusions and future work based on this study's findings are presented.

2 Literature Review

Based on the literature, and as mentioned earlier, three main directions of study regarding OCD were observed. Focusing on neuroanatomy, where the primary objective is to analyze the causes of OCD, Hu *et al*. [2016] applies a multivariate analysis (MVPA) on high-dimensional structural images to discriminate between OCD patients and healthy subjects. Using *Support Vector Machine* (SVM) and *Gaussian Process Classifiers* (GPC), the authors analyzed differences in gray matter and white matter volume in the brain. The results demonstrated an accuracy of over 75% for both classifiers, showing that these anatomical characteristics help differentiate OCD patients.

Considering the relationship between OCD and other psychiatric disorders, it is possible to highlight studies that focus on autism spectrum disorder (ASD), generalized anxiety disorder (GAD), and eating disorders (EDs). Højgaard *et al*. [2023] seeks to differentiate children with OCD and autistic traits from those with OCD without these traits. The presence of attention deficit hyperactivity disorder (ADHD) and tic disorders was significantly associated with OCD with autistic traits. In Bang *et al*. [2020], the authors assessed the presence of ED symptoms in OCD patients. The results sug-

gest that individuals with OCD may present clinical EDs or be at high risk of developing them.

Taking into account studies on behavioral characteristics, specific symptoms and their impact on quality of life are typically evaluated. In Segalàs *et al*. [2021], the authors measured and analyzed the sleep quality of OCD patients and the control group using multiple linear regression. The results indicated that OCD patients exhibited lower sleep quality and more disturbances compared to the control group. The study also found that symptoms of depression and anxiety traits generated by OCD were correlated with poor sleep quality.

Socioenvironmental characteristics related to OCD have also been studied. Boger *et al*. [2020] analyzed the impact of childhood maltreatment on the presence and severity of OCD in adulthood and concluded that childhood abuse is associated with more severe OCD symptoms in adulthood. Brander *et al*. [2016], Singh *et al*. [2023] and Endres *et al*. [2025] conduct an in-depth review of the OCD literature from a clinical perspective. They conclude that the review identified a range of potential environmental risk factors for OCD, particularly in perinatal complications, reproductive cycle events, and stressful or traumatic life events, but none of these factors can be conclusively said. All conclude that the quality of life is significantly impaired by OCD, and highlight the difficulties of its diagnosis.

Within the field of machine learning (ML), OCD has gained attention from the scientific community in recent years. For example, Kirsten *et al*. [2021] addresses the diagnosis of the disorder through the analysis of brain activity images obtaining AUC of 0.954. Clemmensen *et al*. [2022] and Kalmady *et al*. [2022] use ML to develop frameworks and intelligent devices capable of detecting the presence of OCD in individuals under monitoring. The results obtained Accuracy of 89% (sample with 47 individuals) and 80.3% (sample with 350 instances) respectively. Xuanyi *et al*. [2023] conducted a review of works, based on neuroimaging, to explore the relationship between changes in patient neurological function and OCD. Patel *et al*. [2023] proposes strategies to enhance the diagnosis of OCD by incorporating neurobiological information from individuals. The results obtained Accuracy of 86%. Segalàs *et al*. [2021] aimed to examine the long-term course of OCD in patients treated with different approaches (drugs, psychotherapy, and psychosurgery) and to identify predictors of clinical outcome by machine learning. The best machine learning model achieved a correlation of 0.63 for predicting the long-term YBOCS Scale score. The main limitations was the sample size of 60 patients.

Recently, in Zaboski *et al*. [2024], the authors incorporate socioenvironmental aspects to predict the severity of OCD within social contexts influenced by individuals' personalities and religiosity.

There are several studies on the causes of OCD and the unique characteristics that interfere with its diagnosis. However, few works offer a global and systemic view of OCD symptoms, especially within the Brazilian population. This study aims to contribute to this scope by exploring socioenvironmental characteristics related to OCD through machine learning techniques.

3 Methodology

3.1 Dataset description

The most recent National Health Survey (PNS) from 2019 (IBGE [2020]), a study conducted by the Brazilian Institute of Geography and Statistics (IBGE) in partnership with the Ministry of Health, was used as the data source.

The survey includes not only demographic data of the respondents but also characteristics related to labor activity, household income, the presence of individuals with disabilities in the household, use of health insurance plans, and utilization of healthcare services. General health information is collected for individuals aged 60 years or older, and for women aged 50 or older, information on women's health conditions is gathered. The study also considers self-perception of health status, accidents and violence experienced, lifestyle, chronic diseases, prenatal care, oral health, and medical care. It encompasses data on the Brazilian population's overall health situation and lifestyle in 2019, with 293,727 respondents and 1,087 questions to be answered.

The PNS dataset is primarily structured around four key areas: the performance of the national health system, the health conditions of the population, the surveillance of diseases and health-related issues, and associated risk factors.

3.2 Understanding the problem and conceptual attribute selection

It is important to note that the database considered for this study originally contains 1081 variables, and any feature selection method such as Filters, Wrapper, or Embedded applied to this number of variables would be computationally infeasible. These methods would be more efficient and effective if a conceptual feature selection process could be applied beforehand. This conceptual selection should begin with an understanding of the problem domain, based on both explicit and tacit knowledge. The importance of domain understanding had already been emphasized by Fayyad *et al*. [1996] and later by Cao [2010], but it has been largely neglected in data mining and machine learning projects. In Guyon *et al*. [2019], an analysis of the solutions presented by participants during AutoML challenges was conducted. The authors identified that preprocessing was not a target among the participants. According to the analyses, the top-ranked participants did not apply a feature selection process, and two-thirds of the participants ignored irrelevant features. Currently, the quest to improve the accuracy of Large Language Models (LLM) has increased interest in modeling domains expressed through domain-specific knowledge graphs, which correspond to semantic representations, or ontology, about a domain. Recently, Yuan *et al*. [2020] has proposed the construction of knowledge graphs with minimal human supervision. Brady *et al*. [2020] present a guide for public health researchers, who are interested in building conceptual models to convey their ideas to diverse audiences and purposes.

For domain understanding and subsequent conceptual feature selection, the CAPTO method, proposed in Gonçalves *et al*. [2024], was applied. This method is based on the Spiral

Model of Knowledge, one of the most influential theories in knowledge management today Teece [2000]. CAPTO aims to capture tacit knowledge extracted from domain experts and combine it with explicit knowledge obtained from various sources such as scientific literature, technical reports, and database dictionaries. It then proposes constructing a conceptual model to support the selection of attributes in data science projects. These conceptual models are structured into dimensions (different perspectives on the problem domain), aspects (relevant directions within each dimension), and potential attributes (variables linked to each aspect of the domain). These attributes should then be aligned with the variables from the available data sources for the study. Below, we briefly outline the stages of the CAPTO method.

As a first step, systematic interviews were conducted with domain experts in psychological disorders, including professionals from both psychology and psychiatry. During the interviews, the major behavioral characteristics of individuals with OCD were discussed, along with how these characteristics can be used for diagnosing the disorder. In addition to domain expert interviews with psychologists and psychiatrists, we also examined research on behavioral components associated with OCD. Prior studies have shown correlations between poor sleep quality and increased depression and anxiety symptoms in OCD patients Segalàs *et al*. [2021]. Nagata *et al*. [2023] analyzed screen time in children with the higher probability of developing OCD , and Boger *et al*. [2020] verified the impact of childhood maltreatment on greater symptom severity in adulthood . Additionally, the socioenvironmental factors that might influence the presentation of the symptoms were addressed. These insights, combined with expert perspectives, were consolidated into the unified conceptual model (CM), see Figure 1. These interviews helped identify various dimensions (perspectives) of the domain. The dimensions and aspects of the CM, along with the attributes linked to the 2019 PNS dataset, through the conceptual attribute selection, are shown in Table 1. A more detailed description of the selected attributes can be found in the PNS data dictionary IBGE [2020].

Table 1. Dimensions and Aspects for "Obsessive-Compulsive Disorder (OCD) selected by the conceptual model"

Problem Domain: OCD	
Dimension: DSM-5 Diagnostic Criteria	
Aspects	**Mapped Attributes**
Chronic Diseases	Q00201, Q03001, Q060, Q06306, Q068, Q074, Q079, Q088, Q11604, Q120, Q124
Mental Suffering (distress, excessive fear, sadness, and anxiety)	Q092, Q11006, Q11007, Q11008, Q11010
Controlled medications (substances with potential for psychological or physical dependence)	Module Q - Chronic Diseases
Alcohol and Tobacco	ALCOHOL: P027, P02801 TOBACCO: P050 to P053
Dimension: Psychological Feelings and Triggers	
Aspects	**Mapped Attributes**
Trauma and Violence	Module V - Violence (For people aged 18 or older)
Work	Module E - Work Characteristics of people aged 14 or older
Housing	Module A - Household Information
Sanitary Habits	U00204, U00101
Individual Characteristics	C006, C008, C009, C011, VDD004A
Dimension: Health Control and Lifestyle	
Aspects	**Mapped Attributes**
Use of Health Services	J012
Nutrition	P00601 to P02602
Physical Exercise	P034 to P03702
Health Plan	I00102, I001021
Screen Time	P04501 and P04502
Sleep	Q132 to Q134, N010

3.3 Dataset assembly

The next step is to compose the new dataset after the conceptual attribute selection step, guided by the conceptual model.

For the instance selection, the Q11009 attribute from the 2019 PNS (OCD diagnosis, 1: Yes, 2: No) was considered. Instances without classification for OCD and all instances corresponding to individuals under 18 years of age were removed. The latter removal was necessary because several attributes of this group that were used in the subsequent preparation and transformation steps contained missing values. The dataset consists of 270 instances for the *class OCD = Yes* and 370 instances for the *class OCD = No*.

3.4 Data transformation and preprocessing

The selected attributes were divided into two categories to form two distinct datasets. The first dataset includes only behavioral attributes, while the second one contains socioenvironmental attributes. This separation was made to evaluate how these two categories contribute to describing the profile of individuals with OCD. The behavioral characteristics include *Alcohol consumption, Tobacco use, Health services utilization, Frequency of oral hygiene, Physical exercise, Sleep quality, Screen time, Dietary quality*, and *Medication use*. The socioenvironmental characteristics include *Sex, Age, Race/ethnicity, Marital status, Education level, Indicators of violence (psychological, physical, and sexual), Work vulnerability, Housing vulnerability, Health insurance, Chronic diseases*, and *Mental distress*.

The transformation and attribute merging rules are shown in Tables 2, 3, and 4. The attribute nomenclature can be found in the PNS microdata dictionary IBGE [2020].

The attribute values derived from numerical calculations were discretized by dividing them into percentiles. Each group of values within these percentiles was transformed into a category for a new discretized attribute. The socioenvironmental characteristics, especially those related to *Work vulnerability* and *Housing vulnerability*, were based on the social determinants of health outlined by Buss and Pellegrini [2007], and the studies of Azeredo *et al*. [2007] and Pasternak [2016] for defining irregular housing conditions.

Continuous attributes were transformed into discrete categories through a supervised discretization process. To achieve this, the statistical distribution of the attributes, specifically quartiles, was used as the initial criterion to define the limits of the intervals. Furthermore, the distribu-

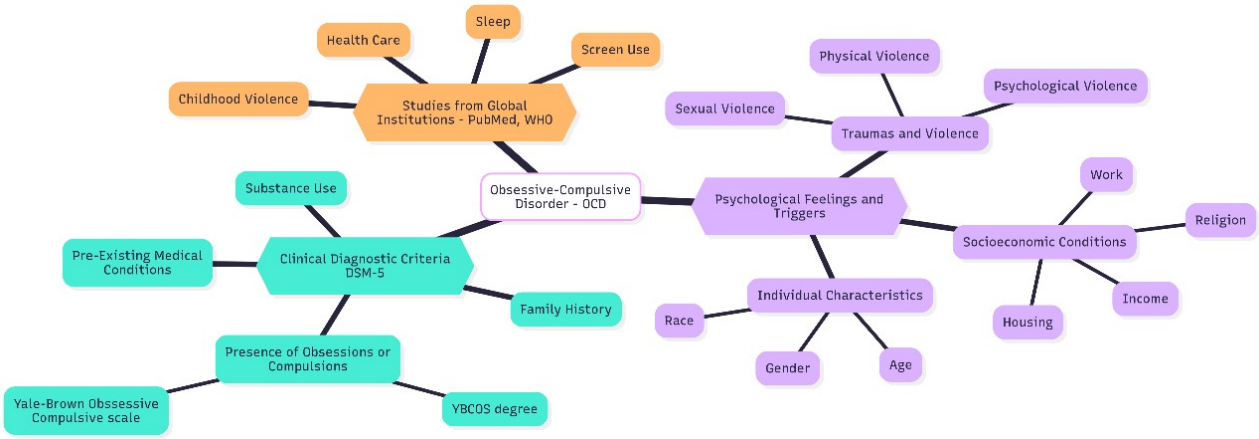

Figure 1. Conceptual model for Obssessive-Compulsive Disorder (OCD)"

Table 2. Data transformation - Behavioral Aspects

Behavioral Aspects	
Aspect	**Transformation rules**
Alcohol	If "P027" = 1, then "Never drunk" If "P027" = 2 and "P02801" = Null, then " 1x per week" If "P027" = {2,3} and "P02801" = 0, then " 1x per week" If "P027" = 3 and "P02801" = {1,2,3}, then " 1 to 3x per week" If "P027" = 3 and "P02801" = {4,5,6}, then " 4 to 6x per week"
Tobacco	If "P052" = 3, then "Never smoked" If "P050" = 3 and "P052" = {1,2}, then "Smoked, but not anymore" If "P050" = 2, then "Smokes occasionally" If "P050" = 1, then "Smokes daily"
Teeth brushing	If "U0024" = 2, then " 1x per day" If "U0024" = 1 and "U00101" = 4, then " 1x per day" If "U0024" = 1 and "U00101" = 3, then "1x per day" If "U0024" = 1 and "U00101" = 2, then "2x per day" If "U0024" = 1 and "U00101" = 1, then "3x or more per day"
Physical exercises	**Time = Work out time in hours = (P025 * P03701) + (P03702 / 60)**
	If "P034" = 2 and "P035" = 0, then "Doesn't exercise" If "P034" = 2 and Time = 2, then " 2h per week" If "P034" = 2 and 2 < Time <= 4, then "2 to 4h per week" If "P034" = 2 and 4 < Time <= 6, then "4 to 6h per week" If "P034" = 2 and Time > 6, then "> 6h per week"
Sleep Quality	**R = Poin system** **"Q132": Yes = -1, No = 0 +** **"Q133": 1 to 3 days = -1;** **4 to 7 days = -2; 8 to 14 days = -3 +** **"Q134": Yes = -1, No = 0**
	If -1 <= R <= 1, then "Bad" If 2 <= R <= 4, then "Moderate" If R = 5, then "Good"
Screen Time	If "P04501" = 6 and "P04502" = 6, "Doesn't use screens" If "P04501" + "P04502" < 3, "Low" If 3 <= "P04501" + "P04502" < 5, "Moderate" If "P04501" + "P04502" = 5, "High"
Food Quality	**R = Amount of days * Point Sytem:** **"P02501", "P02602" = - 8** **"P02002" = - 5, "P02001" = - 4** **"P023" = 2, "P01601" = 3** **"P01101", "P013", "P015", "P018" = 4"P006" = 5, "P00901" = 8**
	If -11 <= R < 6, then "Very bad" If 6 <= R < 10, then "Bad" If 10 <= R < 14, then "Moderate" If 14 <= R <= 24, then "Good"
Controlled medications	**X = Attributes from module "Q"**
	If any X = 1, then "Uses controlled medications" If all X != 1, then "Doesn't use controlled medications"

Table 3. Data transformation - Socioenvironmental Aspects

Socioenvironmental Aspects	
Aspect	**Transformation rules**
Violence	**Psychological = V00201 to V00205** **Physical = V01401 to V01405** **Sexual = V02701 to V02802**
	A = If all attributes = 2 then "No" A = If any attribute = 1 then "Yes"
Offender	**If A = Yes, then:**
	If attribute = 1, 2 or 3 then "Romantic Relationship" If attribute = 4,5,6 or 7 then "Family" If attribute = 8 then "Friend or Neighbor" If attribute = 9 or 10 then "Work colleague" If attribute =11, 12 or 13 then "Other"
Vulnerab. at Work	**Auxiliary Variables**
	Has a job: If "E001","E002","E003","E004" or "E02601" = 1, then "Yes"
	Has a paid job: "E001", "E01601" or "E01801" = 1 then "Yes"
	Is retired: If "F001011" = 1, then "Yes"
	Has multiple jobs: If "E011" = 2 or 3, then "Yes"
	Total Income: "E01602" + "E01802" + "F001021"
	Rules
	If Has a job = "No" and Is retired = "No", then "Yes" If Has a job = "Yes" and Has a paid job = "No", then "Yes" If Has a job = "Yes" and Has multiple jobs = "Yes", then "Yes" If Has a job = "Yes" and Has multiple jobs = "No" and Income Minimim Wage, then "Yes" If any other case, then "No"

tion of the classification attribute (presence or absence of OCD) within each generated interval was considered. This approach aims to maximize the discriminative capacity of the intervals, that is, to create cutoffs that best represent the separation between the classes of the target variable. After the transformation and initial selection of the attributes, an analysis of the distribution of the attributes in relation to the target variable was performed. This step involved comparing the relative frequency of each attribute category in both classes (with and without OCD). Attributes such as *Sex*, *Use of Health Services*, *Sexual Violence*, and *Chronic Diseases* were excluded from the analysis because they presented highly balanced distributions between the classes, that is, they did not demonstrate a significant discrimination with the variable of interest.

After completing the transformations, the attributes *Sex, Healthcare Utilization, Sexual Violence*, and *Chronic Diseases* were removed. Preliminary experiments showed that including these attributes decreased the model's performance. By the end of the data preparation process, the resulting dataset is significantly smaller compared to the original, as shown in Table 5.

Table 4. Data transformation - Socioenvironmental Aspects

Socioenvironmental Aspects	
Aspect	**Transformation rules**
Vulnerability Housing	**Auxiliary Variables** **Aux 1. (V0022) / (A01401)** **Aux 2. (V0022) / (A011)**
	Inadequate Residence: If "A001" = 3, or "A00210" = {2,3} or "A004010" = 4, then "Yes"
	No Basic Sanitation: If "A01501" = {3,4,5,6}, then "Yes"
	Inadequate Water Source: If "A005010" = {3,5}, then "Yes"
	Inadequate Water Storage: If "A00601" = 3 or "A009010" = {2,6}, then "Yes"
	Inadequate Sanitary Installation : If "A01401" or Aux 1 = 3, then "Yes"
	Inadequate Waste Disposal : If "A016010" = {3,4,5}, then "Yes"
	More than 3 People per Bedroom: If Aux 2 = 3 then "Yes"
	Lack of Basic Household Appliances: If "A018013" = 2 or ("A018019 " and "A018017") = 2 or "A01901" = 2, then "Yes"
	Rules
	If any auxiliary = "Yes" then "Unhealthy Housing" If all auxiliaries = "No" then "Healthy Housing"
Health Insurance	If "I00102" = 1 and "I001021" = X, then X If "I00102" = 2, then 0
Chronic Diseases	**Q00201, Q03001, Q060, Q06306, Q068, Q074, Q079, Q088, Q11604, Q120, Q124**
	If any attribute = 1, then "Yes"
	If all attributes = 2, then "No"
Mental Disorders	**Q092 and Q11007 to Q11010**
	If any attribute = 1, then "Yes"
	If all attributes = 2, then "No"

Table 5. Comparison Original x Final Dataset

	Original	**Final**
Attributes	1087	19
Instances	293726	640

3.5 Machine Learning models

To describe the profile of individuals with OCD were constructed two classification models: one based on *Decision Tree* to obtain the interpretability rules, and another based on *Explainable Boosting Machine* (EBM) (Yin *et al*. [2013]) to rank the attributes. The models were built using the KNIME platform.

EBM is a machine learning technique designed to provide accurate predictions while being highly interpretable. The model is an evolution of *Generalized Additive Models* (GAMs), linear models that allow for non-linear interactions between attributes. The algorithm builds an initial model for each predictor attribute independently. It then applies a *boosting* procedure, an iterative method where errors from the initial model are used to adjust and improve the model. This procedure is carried out additively, leading to an interpretable model.

The parameterization of the algorithms is described as follows, and it took the results of preliminary experiments: The hyperparameters for the Decision Tree were Gini as the quality measure; simple pruning; the minimum number of instances required to split a tree node was set to 10; and binary splits for nominal values were applied. The dataset was divided into 70% for training and 30% for testing (*hold-out* technique). Stratified sampling by class was not applied to avoid the risk of bias, as the dataset size can be considered small, which already affects the learning process. In our work, we considered random sampling. A cross-validation procedure with *k-fold* = 10 was used for training. Fine-tuning of the algorithm was not necessary, given the results achieved.

The EBM based model was used to explain the relevance of each attribute and obtain the ranking of each about its presence and importance in the decision tree rules. The training and testing sets were also defined by the 70-30% non-stratified proportion. The training time for both models was insignificant due to the dataset's small number of instances.

The first model considers only attributes in the behavioral category (results in Table 6 and 8), while the second model includes attributes from both behavioral and socioenvironmental categories (results in Table 7 and 9). As mentioned, the aim is to know whether the behavioral characteristics, when used without the interference of socioenvironmental factors, align with results found in the literature.

4 Experiments and results analysis

For the Decision Tree-based models, Tables 6 and 7 show the results of the cross-validation training process, considering the partial (behavioral attributes) and complete (behavioral and socioenvironmental attributes) datasets. The average values for the *F1-score* are 53% and 59% for the partial and complete datasets, respectively. The best result corresponds to the complete dataset. For predicting *class OCD = Yes*, the *F1-score* reached an acceptable confidence interval of [0.62, 0.63] with a 95% confidence level.

Tables 8 and 9 show the results of the testing procedure for the partial and complete datasets, respectively. The complete dataset again achieved the best result, with an average *F1-score* value of 63%.

It is also noticeable that the model performs better for the *class OCD = No*. This is due to the slight imbalance in the number of instances processed, with 270 instances for the *class OCD = Yes* and 370 instances for the *class OCD = No*. For the experiments conducted, and given the reduced number of instances, it was decided not to apply under-sampling balancing, which would further reduce the number of available instances. Choosing over-sampling balancing could introduce distortions in the results, considering the difficulty in separating the classes.

Based on these results, it was possible to conclude that behavioral and socioenvironmental attributes can improve the model's performance and contribute to a more accurate diagnosis.

The rules generated by the decision tree for the complete dataset are shown in Table 12. It is important to emphasize that the rules considered for analysis correspond to those that represent at least 10% of the population of the predicted classification. Only those with the highest prediction accuracy for each class were explored.

Rule 1 represents 67% of cases that meet the rule and are diagnosed with OCD. The rule highlights the presence of mental suffering, poor sleep quality, and having suffered psychological violence, whether from a friend, romantic relationship, or work colleague. *Rule 2* represents 87% of the cases that meet the rule and are diagnosed with OCD. The rule emphasizes that, although there is no mental suffering, marital status as <*Single*> or <*Divorced*> may make the individual vulnerable to the onset of the disorder

Rule 3 indicates that an individual, although not presenting mental suffering or psychological violence, may develop OCD if they have poor sleep quality and feel vulnerable at work.

Table 6. Cross-Validation: DecisionTree

Behavioral Dataset			
	Precision	**Recall**	**F1-score**
OCD = Yes	0.483	0.383	0.42
OCD = No	0.604	0.686	0.639

Table 7. Cross Validation: Decision Tree

Behavioral & Socio-environmental Dataset			
	Precision	**Recall**	**F1-score**
OCD = Yes	0.534	0.549	0.533
OCD = No	0.666	0.655	0.655

EBM showed low performance for the attribute ranking using the Partial dataset, as presented in Table 13. The average Precision was 0.733, the average *Recall* was 0.162, and the average *F1-score* was 0.265. With the Complete dataset, the model performed better, as expected when complete data is considered, achieving an average Precision of 0.741, an average *Recall* of 0.421, and an average *F1-score* of 0.573. In this case, cross-validation was not applied because the objective was not to assess the model's performance but to leverage its interpretative ability to understand the relevance of each attribute for classification. The ranking of predictor variables for the complete dataset is shown in Table 10. At the first level (with an average absolute score up to 0.15), mental suffering (1) and sleep quality (2) emerge as key factors in individuals with OCD. At the second level (with an average score up to 0.08), food quality (3), work vulnerability (4), work vulnerability combined with alcohol consumption (5), psychological violence (6), and alcohol consumption (7) are identified as significant factors.

Additionally, it can be observed from Table 10 that, despite socio-environmental attributes such as Mental Suffering (1) and Work Vulnerability (4) having strong decision-making influence in classification, behavioral characteristics such as Sleep Quality (2) and Food Quality (3) still rank highly even after the inclusion of socioenvironmental factors.

This pattern aligns with findings in the literature, such as the study by Segalàs *et al*. [2021], where the authors identified a correlation between depression and anxiety symptoms in individuals with OCD and poor sleep quality. The decision tree rules for the complete dataset also support this conclusion, as seen in Table 12, where the characteristic Sleep Quality = Poor had a strong ability to separate tree nodes for the *class OCD = Yes*, as did Sleep Quality = [Good, Moderate] for the *class OCD = No*. Regarding food quality, the study by Bang *et al*. [2020] states that a considerable subset of OCD patients may have a clinical eating disorder or be at high risk of developing one. This correlation is visible in Table 11, which associates Food Quality = [Poor, Very Poor] with the *class OCD = Yes* classification.

Table 8. Test Results: Decision Tree

Behavioral Dataset			
	Precisão	**Recall**	**F1-score**
OCD = Yes	0.397	0.387	0.392
OCD = No	0.613	0.624	0.619

Table 9. Test Results: Decision Tree

Behavioral & Socio-environmental Dataset			
	Precision	**Recall**	**F1-score**
OCD = Yes	0.614	0.567	0.59
OCD = No	0.642	0.686	0.664

Table 10. Complete Rank: EBM

Rank	**Attribute**	**Value**
1	mental_sufferings	**0.18**
2	sleep_quality	**0.15**
3	food_quality	**0.1**
4	work_vulnerability	**0.08**
5	work_vulnerability & alcohol_consumption	**0.08**
6	psychological_violence	**0.08**
7	alcohol_consumption	**0.08**
8	psychological_violence & screen_time	**0.07**
9	instruction_level & exercise	**0.06**
10	instruction_level & exercise	**0.06**
11	mental_sufferings & sleep_quality	**0.06**
12	psychological_violence & sleep_quality	**0.05**
13	instruction_level	**0.05**
14	age & sleep_quality	**0.05**

Table 11. Decision Tree Rules: Behavioral Dataset

Behavioral Dataset		
Rule 1	**Total**	**Correct**
If exercise in {"Doesn't exercise", "< 2h a week", ">= 6h a week"} **and** sleep_quality = "Bad" **and** food_quality in {"Bad", "Very Bad"} **Then OCD = "Yes"**	59	41
Rule 2	**Total**	**Correct**
If exercise in {"Doesn't exercise", "< 2h a week", ">= 6h a week"} **and** sleep_quality = "Bad" **and** food_quality in {"Good", "Moderate"} **and** tobacco in {"Smoked, but not anymore", "Smokes occasionally", "Smokes daily"} **Then OCD = "Yes"**	30	17
Rule 3	**Total**	**Correct**
If sleep_quality = "Bad" **and** food_quality in {"Good", "Moderate", "Bad"} screen_time in {"Moderate", "High"} **and** tobacco in {"Smoked, but not anymore", "Smokes occasionally"} **Then OCD = "No"**	48	36

Finally, when analyzing the results generated by the EBM for socioenvironmental attributes in the complete dataset, we can identify Mental Suffering and Work Vulnerability as the most significant. It is important to note that Mental Suffering includes the presence of other mental conditions in addition to OCD, meaning pre-existing conditions or those diagnosed after the OCD diagnosis. In the case of individuals without OCD, this refers to mental conditions such as depression, anxiety, schizophrenia, among others.

According to the rules generated by the tree, Mental Suffering = No shows little correlation with OCD, whereas Mental Suffering = Yes is correlated with *class OCD = No*. This rule aligns with the clinical diagnosis of the DSM-5, which states that the symptoms presented by the patient for an OCD diagnosis must not be better explained by diagnostic criteria for another mental disorder. This means that if a person is diagnosed with OCD, symptoms of anxiety and depression, for example, presented by the patient, are considered symptoms of OCD and not separate diagnoses of Anxiety or Depression. Therefore, individuals diagnosed with OCD are also unlikely to have other diagnoses.

Regarding the attribute Work Vulnerability, results con-

Table 12. Decision Tree Rules: Complete Dataset

Behavioral & Socio-environmental Dataset		
Rule 1	**Total**	**Correct**
If mental_suffering = ”Yes” **and** sleep_quality = ”Bad” **and** psychological_violence in {”Friend or Neighbour”, ”Romantic Relationship”, ”Work Colleague”} **Then OCD = ”Yes”**	30	20
Rule 2	**Total**	**Correct**
If mental_suffering = ”No” **and** legal_status in {”Single”, ”Divorced”} **then OCD = ”Yes”**	30	26
Rule 3	**Total**	**Correct**
If mental_suffering = ”Yes” **and** sleep_quality = ”Bad” **and** psychological_violence = ”No” **and** work_vulnerability in {”Adult”, ”Middle-aged”, ”Elderly”} **Then OCD = ”Yes”**	29	19
Rule 4	**Total**	**Correct**
If mental_suffering = ”Yes” **and** sleep_quality in{”Bad”, ”Moderate”} **and** psychological = ”No” **Then OCD = ”No”**	75	71
Rule 5	**Total**	**Correct**
If mental_suffering = ”Yes” **and** sleep_quality = ”Bad” **and** psychological = ”No” **and** work_vulnerability = ”Yes” **and** skin_color in {”White”, ”Black”, ”Yellow”, ”Indigenous”} **Then OCD = ”No”**	37	28

Table 13. EBM training results

Model	**Precision**	**Recall**	**F1-score**
Partial Dataset	0.733	0.162	0.265
Complete Dataset	0.741	0.421	0.573

trary to those discussed with health specialists were found. During the interviews, it was reported that vulnerabilities such as stress and unpleasant work environments would have high relevance in the presentation of OCD symptoms. However, the rules identified by the Decision Tree showed the opposite, indicating that when this aspect is present, it correlates with the absence of OCD in the patient.

5 Conclusion and Future Work

The analysis of predictive models based on behavioral and socioenvironmental variables revealed valuable insights into the classification of OCD. The inclusion of socioenvironmental variables significantly increased the *F1-score* for *class OCD = Yes*, although the impact on *class OCD = No* was less pronounced.

Behavioral variables, such as sleep quality and food quality, were the most informative, as highlighted by both the EBM model and the decision tree rules in relation to existing literature. In particular, poor sleep quality was a significant constant in the decision tree rules, reinforcing its importance as a determining factor in behavioral analysis. By adding socioenvironmental variables, we observed an increase in the proportion of rules related to these characteristics, especially the absence of other mental sufferings, which is highly correlated with OCD. This aligns with the DSM-5, which indicates that individuals with OCD are unlikely to have prior diagnoses of other disorders due to symptom overlap.

In summary, although the dataset does not provide precise information due to several challenges, such as the low number of OCD cases, the need to exclude underage participants, and the removal of attributes too balanced between classes, it demonstrated that socioenvironmental characteristics are valuable for OCD classification, mainly when associated with behavioral traits. Therefore, the results support the usefulness of including various factors to enhance the accuracy and depth of predictive analyses on OCD.

It is essential to highlight a significant limitation of this study: the absence of a longitudinal study of the participants in the 2019 PNS. It is possible that the OCD diagnosis occurred many years before the interviews, which prevents asserting that the socioenvironmental characteristics presented in the survey were the same as those at the time of diagnosis. Thus, future work should include a longitudinal study to track changes in socioenvironmental characteristics over time and their relationship with the OCD diagnosis.

Moreover, there is a desire to replicate the process with a larger dataset, i.e., with a more significant number of OCD-labeled instances, and modify the data transformations of socioenvironmental characteristics, aiming for a more representative model. The algorithms’ fine-tuning process, which was not performed in this study, will also be executed. Additionally, the goal is to understand the reason behind the inverse relationship between Work Vulnerability and the diagnostic by merging new attributes or exploring characteristics correlated with this aspect.

Finally, it is important to emphasize that data mining searches for patterns within a data set, which are potential statistical hypotheses that must be confirmed through subsequent population studies. Given these discovered profiles for OCD, it is hoped that new research proposals will be developed in the fields of medicine and psychology to understand behavioral phenomena and contribute to improving the quality of life of individuals with OCD, in addition to potentially fostering public policy proposals.

This study has limitations, such as the use of data that may not accurately reflect socioeconomic and behavioral conditions during the development of OCD, as they represent only a snapshot of a specific moment in the interviewee’s life. It should be noted that the proposed analysis is not diagnostic in nature, since OCD should be diagnosed exclusively based on the clinical criteria of the DSM-5. Thus, this study serves as an exploratory guide, capable of identifying profiles more vulnerable to triggers of the disorder, but is not intended to replace medical or psychological evaluation.

Acknowledgements

The authors thank The National Council for Scientific and Technological Development of Brazil (CNPQ); The Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) (Grant PROAP 88887.842889/2023-00 – PUC/MG, Grant PDPG 88887.708960/2022-00 – PUC/MG - INFORMATICA and Finance Code 001); Minas Gerais State Research Support Foundation (FAPEMIG) under grant number APQ-01929-22, APQ-05058-23 and the Pontifical Catholic University of Minas Gerais, Brazil.

Funding

This research did not receive any specific financial funding.

Authors' Contributions

AR and LZ contributed to the conceptualization, methodology, formal analysis, investigation, Original Draft Preparation, and Review & Editing of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no conflicts of interest for the publication.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in `https://github.com/licapLaboratory/DataBase-PNS-TOC`.

References

American-Psychiatric-Association (2013). *Manual Diagnóstico e Estatístico de Transtornos Mentais*. Artmed, 5 edition.

Azeredo, C. M., Cotta, R. M. M., Schott, M., Maia, T. d. M., and Marques, E. S. (2007). Avaliação das condições de habitação e saneamento: a importância da visita domiciliar no contexto do programa de saúde da família. *Ciência e Saúde Coletiva*, 12(3):743–753. DOI: 10.1590/S1413-81232007000300025.

Baland, J., Chamberlain, S. R., and Sahakian, B. J. (2023). Obsessive-compulsive disorder: Etiology, neuropathology, and cognitive dysfunction. *Brain and Behavior*, 13. DOI: https://doi.org/10.1002/brb3.3000.

Bang, L., Kristensen, U. B., Wisting, L., Stedal, K., Garte, M., Åse Minde, and Øyvind Rø (2020). Presence of eating disorder symptoms in patients with obsessive-compulsive disorder. *BMC Psychiatry*, 20(1):36. DOI: 10.1186/s12888-020-2457-0.

Boger, S., Thomas Ehring, G. B., and Werner, G. G. (2020). Impact of childhood maltreatment on obsessive-compulsive disorder symptom severity and treatment outcome. *European Journal of Psychotraumatology*, 11(1):1753942. DOI: 10.1080/20008198.2020.1753942.

Brady, S., Brubaker, L., Fok, C., Gahagan, S., Lewis, C. E., Lewis, J., Lowder, J. L., Nodora, J., Stapleton, A., and Palmer, M. H. (2020). Development of conceptual models to guide public health research, practice, and policy: Synthesizing traditional and contemporary paradigms. *Health Promotion Practice.*, 21(4):510–524. DOI: https://doi:10.1177/1524839919890869.

Brander, G., Pérez-Vigil, A., Larsson, H., and Mataix-Cols, D. (2016). Systematic review of environmental risk factors for obsessive-compulsive disorder: A proposed roadmap from association to causation. *Neuroscience and Biobehavioral Reviews*, 65:36–62. Epub 2016 Mar 21. DOI: 10.1016/j.neubiorev.2016.03.011.

Buss, P. M. and Pellegrini, A. (2007). A saúde e seus determinantes sociais. *Physis: Revista de Saúde Coletiva*, 17(1):77–93. DOI: 10.1590/S0103-73312007000100006.

Cao, L. (2010). Domain-Driven Data Mining: Challenges and Prospects . *IEEE Transactions on Knowledge & Data Engineering*, 22(06):755–769.

Clemmensen, L. K. H., Lønfeldt, N. N., Das, S., Lund, N. L., Uhre, V. F., Mora-Jensen, A.-R. C., Pretzmann, L., Uhre, C. F., Ritter, M., Korsbjerg, N. L. J., Hagstrøm, J., Thoustrup, C. L., Clemmesen, I. T., Plessen, K. J., and Pagsberg, A. K. (2022). Associations between the severity of obsessive-compulsive disorder and vocal features in children and adolescents: Protocol for a statistical and machine learning analysis. *JMIR Res Protoc*, 11(10):e39613. DOI: 10.2196/39613.

Endres, D., Schiele, M., von Zedtwitz, K., Dressle, R., Maier, A., Hohagen, F., Baldermann, J., Coenen, V., Jelinek, L., Domschke, K., and Voderholzer, U. (2025). Obsessive-compulsive disorder - a state-of-the-art review. *Neurosci Biobehav Rev.*, 177:106320. DOI: https://doi.org/10.1016/j.neubiorev.2025.106320.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34. DOI: 10.1145/240455.240464.

Gonçalves, L., Franca, D., and Zarate, L. (2024). Relevância do entendimento do domínio de problema na construção de modelos computacionais de aprendizado. In *Anais do XVIII Brazilian e-Science Workshop*, pages 135–142, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/bresci.2024.240233.

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H. J., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A., Tu, W.-W., and Viegas, E. (2019). *Analysis of the AutoML Challenge Series 2015–2018*, pages 177–219. Springer International Publishing, Cham.

Hiranandani, S., Ipek, S. I., Wilhelm, S., and Greenberg, J. L. (2023). Digital mental health interventions for obsessive compulsive and related disorders: A brief review of evidence-based interventions and future directions. *Journal of Obsessive-Compulsive and Related Disorders*, 36:100765. DOI: https://doi.org/10.1016/j.jocrd.2022.100765.

Hu, X., Liu, Q., Li, B., Tang, W., Sun, H., Li, F., Yang, Y., Gong, Q., and Huang, X. (2016). Multivariate pattern analysis of obsessive–compulsive disorder using structural neuroanatomy. *European Neuropsychopharmacology*, 26(2):246–254. DOI: https://doi.org/10.1016/j.euroneuro.2015.12.014.

Højgaard, D., Arildskov, T. W., Skarphedinsson, G., and et al. (2023). Do autistic traits predict outcome of cognitive behavioral therapy in pediatric obsessive-compulsive disorder? *Research on Child and Adolescent Psychopathology*, 51:1083–1095. DOI: 10.1007/s10802-023-01078-5.

IBGE (2020). Pesquisa nacional de saúde 2019 - instituto brasileiro de geografia e estatística. `https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=25921&t=resultados`. Acesso em: 2024-07-15.

Kalmady, S. V., Paul, A. K., Narayanaswamy, J. C., Agrawal,

R., Shivakumar, V., Greenshaw, A. J., Dursun, S. M., Greiner, R., Venkatasubramanian, G., and Reddy, Y. J. (2022). Prediction of obsessive-compulsive disorder: Importance of neurobiology-aided feature design and cross-diagnosis transfer learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 7(7):735–746. DOI: https://doi.org/10.1016/j.bpsc.2021.12.003.

Kirsten, K., Pfitzner, B., Löper, L., and Arnrich, B. (2021). Sensor-based obsessive-compulsive disorder detection with personalised federated learning. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 333–339. DOI: 10.1109/ICMLA52953.2021.00058.

KNIME (2023). Decision tree learner (3.6.0). `https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.decisiontree2.learner2.DecisionTreeLearnerNodeFactory3/`. Accessed: 2024-07-12.

Nagata, J. M., Chu, J., Zamora, G., Ganson, K. T., Testa, A., Jackson, D. B., Costello, C. R., Murray, S. B., and Baker, F. C. (2023). Screen time and obsessive-compulsive disorder among children 9–10 years old: A prospective cohort study. *Journal of Adolescent Health*, 72(3):390–396. DOI: https://doi.org/10.1016/j.jadohealth.2022.10.023.

Pasternak, S. (2016). Habitação e saúde. *Estudos Avançados*, 30(86):51–66.

Patel, K., Tripathy, A. K., Padhy, L. N., Kar, S. K., Padhy, S. K., and Mohanty, S. P. (2023). Accu-help: A machine-learning-based smart healthcare framework for accurate detection of obsessive compulsive disorder. *SN Comput. Sci.*, 5(1). DOI: 10.1007/s42979-023-02380-1.

Rodrigues, A. P. C. and Zárate, L. E. (2024). Análise dos fatores socioambientais e comportamentais an identificação do transtorno obsessivo compulsivo: Uma abordagem com dados da pesquisa nacional de saúde 2019. In *Proceedings of the 39th Brazilian Symposium on Data Bases*, pages 78–90, Porto Alegre, RS, Brasil. SBC.

Segalàs, C., Labad, J., Salvat-Pujol, N., Real, E., Pino, A., Bertolín, S., Jiménez-Murcia, S., Soriano-Mas, C., Monasterio, C., Menchón, J. M., and Soria, V. (2021). Sleep disturbances in obsessive-compulsive disorder: influence of depression symptoms and trait anxiety. *BMC Psychiatry*, 21(1):42. DOI: 10.1186/s12888-021-03038-z.

Singh, A., Anjankar, V., and Sapkale, B. (2023). Obsessive-compulsive disorder (ocd): A comprehensive review of diagnosis, comorbidities, and treatment approaches. *Cureus.*, 15:e48960. DOI: https://doi.org/10.7759/cureus.48960.

Souza, V., Nobre, J., and Becker, K. (2020). Characterization of anxiety, depression, and their comorbidity from texts of social networks. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 121–132, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2020.13630.

Teece, D. (2000). Strategies for managing knowledge assets: The role of firm structure and industrial context. *International Journal of Strategic Management*, 33:35–54. DOI: https://doi.org/10.1016/S0024-6301(99)00117-.

The InterpretML Contributors (2023). Explainable boosting machine. Accessed: 2024-05-23.

Xuanyi, L., Kang, Q., and Gu, H. (2023). A comprehensive review for machine learning on neuroimaging in obsessive-compulsive disorder. *Frontiers in Human Neuroscience*, 17:1280512. DOI: https://doi.org/10.3389/fnhum.2023.1280512.

Yin, L., Rich, C., Johannes, G., and Giles, H. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631.

Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., and Luo, J. (2020). Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowl Inf Syst*, 62:317–336. DOI: https://doi.org/10.1007/s10115-019-01351-4.

Zaboski, B. A., Wilens, A., McNamara, J. P., and Muller, G. N. (2024). Predicting ocd severity from religiosity and personality: A machine learning and neural network approach. *Journal of Mood and Anxiety Disorders*, 8:100089. DOI: https://doi.org/10.1016/j.xjmad.2024.100089.