UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

MÁSTER EN BIOLOGÍA COMPUTACIONAL

DEPARTAMENTO DE INTELIGENCIA ARTIFICIAL

DEPARTAMENTO DE TECNOLOGÍA FOTÓNICA Y BIOINGENIERÍA

Characterization of risk factors for depression using machine learning techniques to generate new knowledge in brain health domain.

TRABAJO FIN DE MÁSTER

Autor: Laura García Martínez

Tutor Externo: Paloma Chausa Fernández

Tutor Académico: Esteban García Cuesta

July, 2023

1. ABSTRACT

Depression is a mental disorder that is characterised by feelings of sadness and guilt, low self-worth, and loss of interest or pleasure. It affects almost four percent of the world's population. Although many factors related to depression are known, the prevalence of this disorder makes it necessary to extract new knowledge to generate new clinical hypotheses and to have more resources for both prevention and treatment if the causality is validated. In this work, self-reported health, socio-economic and lifestyle data collected in the context of the Barcelona Brain Health Initiative have been used to generate machine learning models with several classifiers (Logistic regression, Random Forest, K-Nearest Neighbours, Gradient boosting, and Multilayer perceptron). These data include depression risk factors already reported in the scientific literature and newly variables: life purpose, personal growth, number of psychiatric, neurodegenerative, and total diagnoses, and information about childhood, and social interaction. As a result, three models with similar metrics were obtained, with accuracies near seventy percent, Logistic regression, Random Forest, and Gradient boosting. Models were also generated with these same classifiers for the prediction of gender-specific depression, improving the metrics of the models trained with the full dataset. As a result of the calculation of the Shapley values (relevance metric) of the factors, life purpose, social engagement, psychiatric diagnoses, total diagnoses, and the father's job during childhood are related to the development of depression disorder.

2. INTRODUCTION

2.1 DEPRESSION

Depression is a mental disorder that, according to the World Health Organisation, is characterised by persistent sadness, feelings of guilt, low self-esteem and loss of interest or pleasure (WHO, 2023). It affects all ages and is 50% more prevalent in women than in men. The most common symptoms include lack of concentration, suicidal thoughts, sleep disturbance, continuous fatigue, and changes in appetite. It affects about 3.8% of people worldwide and is the third leading cause of disability in the world. In addition, 50% of suicide cases are related to a depressive diagnosis (Park, 2019).

Nowadays, apart from observing the symptoms, there are several psychological tests used in depression diagnosis whose efficacy has been tested in clinical trials, such as the Depression Anxiety Stress Scale (DASS) (Henry, 2005) or the anxiety and depression module of the Patient Heath Questionnaire (PHQ-4). The last one exhibits a high sensitivity rate of 0.90 but a lower specificity rate of 0.61 (Muñoz-Navarro, 2017).

Treatment for depression includes a psychological approach with psychologists and a pharmacological approach with antidepressants such as selective serotonin reuptake inhibitors or serotonin and norepinephrine reuptake inhibitors including examples such as fluoxetine and venlafaxine, respectively (Brent, 2016 and Furukawa, 2019).

Determining factors in the development of this disease include socio-economic status (education level, job, salary); conditions related to the place of residence (pollution, violence, crowding); demographic factors such as age or gender; habits related to the general health of the individual such as smoking, alcohol consumption, physical exercise; genetic predisposition; and the individual's health status (Kwang-Sig, 2022).

2.2 MACHINE LEARNING IN HEALTHCARE

With the digitisation of many areas including healthcare, there are different computational techniques that have started to be applied to make diagnoses by processing big volumes of data in different formats such as images and text. Particularly, in the field of mental health, Machine Learning (ML) algorithms are being tested to predict possible diseases or disorders at an early stage and apply a treatment to prevent their development or improve the patient's conditions (Aleem, 2022).

Machine Learning is a growing field that focus in making computers learn from data to predict or classify new input data. It can be supervised if the instances are already labelled with a specific condition, which allows the classification of new observations according to that label, or unsupervised if there are no labels, so its main objective is to find patterns in data (Deo, 2015).

Machine Learning-generated models provide a better understanding of patients' mental health conditions and assist professionals in decision-making related to the treatment. With the assessment of some biomarkers, a ML model can predict the possible diagnosis of the patients facilitating the work of healthcare professionals and saving treatment costs due to late diagnosis of these diseases (Aleem, 2022).

There are articles published that already describe Machine Learning models to predict depression, achieving accuracies above 60%. These models use different types of data, such as: demographic data (age and gender), socio-economic status, pain, access to therapy, gut microbiota, disabilities, and other genetic and physiological markers, among others. Depending on the format and/or resource of this data, different ML techniques would be appropriate for the diagnosis. Logistic regression, Random Forest, Support vector machine and Artificial Neural Networks are the methods that achieved good predictions in the case of numeric data; Random Forest is useful with genomic data and social-networks data; and Random Forest and Support vector machine for radiomic data (Kwang-Sig, 2022).

Many of these predictor variables for Machine Learning models in depression are based on medical data that usually require tests or measurements to be performed on the patient, which can be time-consuming and take resources, in addition to the risk of patient abandonment (Kwang-Sig, 2022 and Yuan, 2020). Variables like sociodemographic data can be collected by filling in a test, saving time for early response to a possible diagnosis of depression.

To extract knew clinical knowledge, it is proposed to search for new variables related to mental health that can be obtained through simple tests. To obtain these factors, a search through the multiple tests carried out in the Barcelona Brain Health Initiative is going to be performed. And, in order to compare the relevance in depression development of this new variables to the already studied ones, several ML models are going to be generated using the algorithms described in previous articles (e.g. Logistic regression and Random Forest) (Kwang-Sig, 2022). Also it is added the assessment of models trained with datasets split by gender to study whether men and women are affected differently by these factors.

Between women and men some factors may be more relevant than others and therefore the preventive measures to be taken would be different. On the contrary case, measures could be taken at the population level without the need to distinguish between women and men.

Therefore, the aims of this study are to generate new Machine Learning models to predict depression, to analyse the relevance of factors related to mental health in the development of depression and to study the effect of these factors in gender-specific diagnosis.

3. MATERIAL AND METHODS

3.1 DATA SOURCE: INITIATIVE BBHI

The data used in this study come from a dataset from the first phase of Barcelona Brain Health Initiative (BBHI), conducted by the Institut Guttmann. This initiative is a study that started in 2017 with the objective of identifying factors related to brain health (Delgado-Gallén, 2023). 5933 participants from both sexes, between 40 and 65 years and free from any self-reported neurological or psychiatric diagnosis at the time of recruitment, fulfilled a series of questionnaires related to seven “pillars of brain health” which are nutrition, physical activity, general health, sleep, socialization, cognitive activity, and vital plan (Cattaneo, 2018).

The dataset is presented in different tables in MySQL format. There are two tables for each questionnaire, one with the answers codified and one with the scores of the questions decodified. The first year, there were seven questionnaires for all the brain health pillars. There is also a table with some sociodemographic information as well as the dates of each questionnaire’s fulfilment and, there is a more complete separated table with these dates.

Complementing these SQL tables, there were some excel files with complete statements of the questions and information about the questionnaires used for each pillar.

Furthermore, they were added to the analysis some excel tables with the information necessary to add the new depression diagnostics of the last questionnaire and the variables related to vital plan, childhood, social information, and other diagnostics.

In the Table 1 it is shown the available tests used in each questionnaire from the BBHI dataset (Cattaneo, 2020).

Table 1. Official questionnaires for pillars of brain health and the studied factors.

BBHI Questionnaire	Acronym	Description
Patient-Reported Outcomes Measurement Information System, General health	PROMIS	Assessment of an individual's physical, mental, and social health
Patient Health Questionnaire-4	PHQ-4	Self-percieved Mental Health
Physical Activity Questionnaire	PAQ	Physical activity measure
Godin Leisure Time Physical Activity Questionnaire	GLTPAQ	Measurement of leisure-time physical activity
Football Players Health Study	FPHS	Physical activity measure
Mediterranean Diet Adherence screener	MeDas	Estimation of the adherence to the
Personal growth and Purpose in life subscales of Ryff's scale of Psychological well-being		Test about the six items of psychological well-being: autonomy, environmental mastery, personal growth, positive relations with others, purpose in life, and self-acceptance
The engaged Living Scale	ELS	Assessment of an individual's level of engagement in their life activities and experiences
Subjective Self Health Horizon Questionnaire	SHH-Q	Assessment of individuals' future time perspectives related to 4 dimensions: Novelty, Body, Work and Life Goals
Sense of Coherence scale	SOC	Assessment of life vision
Alcohol Use Disorders Identification Test	AUDIT	Alcohol consumption
Jenkins Sleep Evaluation Questionnaire	JSEQ	Measurement of the sleep quality
Pittsburg Sleep Quality index	PSQI	
Patient-Reported Outcomes Measurement Information System, Cognitive Abilities and Cognitive	PROMIS Cognitive Abilities	Assessment of patient-perceived cognitive deficits
Lubben Social Network Scale	LSNS	Measurement of perceived social support received by family and friends
Pack-year index		Smoking habits

3.2 PROGRAMMING ENVIRONMENT: MYSQL, JUPYTER NOTEBOOK, ANACONDA

The script to extract the table of variables to create the models was written in Structured Query Language (SQL), a programming language that allows you to manipulate and select data from relational databases using MySQL Workbench version 8.0.32 (Reichardt, 2021).

The scripts to preprocess the data and to create the models were written in python, version 3.9.13, using the web interface Jupyter notebook in its version 6.4.12 installed through Anaconda version 2023.03-1.

3.3 DATA ANALYSIS METHODOLOGY

3.3.1 DATA SELECTION

For the creation of an input dataset, there was an initial selection of variables based on the existing evidence about machine learning and brain health, and known depression related factors:

- sum_sleep_quality: corresponds to the sum of the Jenkins sleep scale questions.

- sum_physical_health: based on the score obtained in FPHS questionnaire.

- sum_cognitive_function: corresponds to the score of the PROMIS Cognitive Abilities and Cognitive Concerns questions.

- sum_smoking: sum of the questions related to smoking habits.

- diabetes_bloodsugar: binary variable, 1 for the presence, 0 for the absence.

- marital_status: 1 for married, 2 for separated or divorced, 3 for widowed, 4 for single or 5 for a stable relationship but not married.

- age (at the moment of answering the first questionnaire).

- gender: 1 for male, 2 for female.

- study_level: 0 for no studies, 1 for primary, 2 for secondary or 3 for higher.

- sum_alcohol: sum of the questions related to alcohol consumption.

- sum_medas: was the score of the Mediterranean Diet Adherence Screener (MeDAS) questionnaire (dietary information).

- pain_medication: 1 for the intake, 0 otherwise.

- cancer: 1 for the presence and 0 for the absence.

- sum_global_health: coded as the resulting score of the PROMIS General health test.

All these variables were collected from the questionnaires of the first year, as it is the year with the greater number of participants (5933). Finally, it was included the depression column encoded as 1 if the person suffered from depression after the recruitment or 0 if they didn't.

After the creation of the first set of input variables in MySQL, it was imported to a Jupyter notebook for further data preprocessing.

The variables' table was merged to the tables of the later selected variables:

- Vital plan, separated in two: growth and purpose, based on the score of the Ryff Scales of Psychological Well-Being (Ryff, 1995).

- Childhood information: due to the lack of an official scale, the questions were added as individual variables and coded as Question n Childhood being n an identification number.

- Question 1 Childhood is "Were you born to term or were you premature?"
- Question 2 Childhood is "Did you have complications at birth (needed respirator, incubator or stayed in)?"
- Question 3 Childhood is "What was your birth weight?"
- Question 4 Childhood is "What was your father´s level of education?"
- Question 5 Childhood is "What was your mother´s level of education?"
- Question 6 Childhood is "What was your father´s usual job / job?"
- Question 7 Childhood is "What was your mother´s usual job / job?"
- Question 8 Childhood is "The house where you habitually lived during your childhood, how many rooms did you have?"
- Question 9 Childhood is "Did you live most of the time in a town or a city as a child?"

- social: coded as the score of the Lubben Social Network Scale.

- psychiatric_diagnoses: a sum of anxiety, attention deficit disorder and schizophrenia psychosis diagnoses.

-neurodegenerative_disorders: number of memory loss, dementia, Parkinson´s Disease amyotrophic lateral sclerosis and multiple sclerosis diagnoses.

- cardiovascular_diseases: sum of total heart problems, heart Attacks and cerebral infarctions.

- total_diagnoses: sum of all diagnoses.

3.3.2 DATA PREPROCESSING

Not all the participants fulfilled all the questions of the tests, so, in order to solve the missing data problem, it was decided to impute missing values before training instead of discarding the observations to avoid losing participants that suffered from depression. It was incorporated the sklearn.impute.SimpleImputer from the library scikit-learn, version 1.2.2 (Pedregosa, 2011).

Due to class imbalance between depression diagnosis (461 participants) and the absence of it (5137 participants), it was performed a resample to avoid biases in model training, since the model may predict all test subjects as the majority class. This resampling was performed with four techniques: two undersampling algorithms, random undersampling and Tomek links, which consists of randomly eliminating the necessary number of instances of the majority class to equal the minority class; and two oversampling techniques, random oversampling and SMOTE (Synthetic Minority Oversampling Technique), which consist of increasing the minority class, generating copies of the existing instances (Mohammed, 2020). The library used for these resampling was imbalanced-learn version 0.10.1.

3.3.3 MODELLING

Once the data was prepared, it was created the training (80%) and test set (20%). The validation set was omitted because cross-validation was used as the method for validating the models.

Several types of classifiers from scikit-learn library were used for modelling the data:

-Logistic regression: common technique used in binary classification problems. It assigns coefficients to the predictor variables to obtain the probability of the dependent variable taking the value 1 (Minarno, 2020). To control the training process, we can adjust a series of parameters called hyperparameters in every machine learning model. The hyperparameters tunned in the logistic regression model were C (inverse of regularization strength), max_iter (maximum number of iterations taken for the solvers to converge), solver (algorithm to use in the optimization problem) and random_state (controls the randomness) (Pedregosa, 2011).

-Random forest: type of meta-estimator that trains multiple decision tree classifiers with different subsets of the dataset and employs averaging to enhance predictive accuracy and control the overfitting problem (Zhu, 2022). The parameters tunned were min_samples_leaf (minimum number of samples required to be at a leaf node), min_samples_split (minimum number of samples required to split an internal node), max_depth (maximum depth of the tree), criterion (function to measure the quality of a split), n_estimators (number of trees in the forest) and random_state (controls the randomness of the bootstrap and the sampling of the features) (Pedregosa, 2011).

-Gradient boosting: algorithm that builds an additive model from decision trees (Singh, 2023). The parameters tunned were min_samples_leaf, min_samples_split, max_depth (maximum depth of the individual regression estimators), n_estimators (number of boosting stages to perform), learning_rate (shrinks the contribution of each tree) and random_state (controls the random seed given to each tree estimator at each boosting iteration) (Pedregosa, 2011).

-K-nearest neighbours: algorithm that uses proximity between instances to make classifications (Jääsaari, 2019). The parameter tunned was n_neighbors (number of neighbours to use) (Pedregosa, 2011).

-Multilayer perceptron: artificial neural network made up of multiple layers, such that it has the ability to solve problems that are not linearly separable (Karaki, 2020). The parameters tunned were max_iter (maximum number of iterations), hidden_layer_sizes (number of neurons in the hidden layers), solver (solver used for weight optimization), activation (activation function for the hidden layer) and random_state (controls the random number generation for weights and bias initialization, the train-test split if early stopping is used, and batch sampling when the solver 'sgd' or 'adam') (Pedregosa, 2011).

In all of them it was used a parameter grid to try several parameters with the model_selection.GridSearchCV function from the scikit-learn library, which allows to perform cross validation. For each classifier it was created a model in combination the four types of resampling methods and one with the basal data, each one with their own best parameters extracted from the parameter grid (Pedregosa, 2011).

3.3.4 MODEL SELECTION AND EVALUATION

After creating the models, metrics were calculated with scikit-learn library. The accuracy, F1 and recall scores and the confusion matrices were used to compare the different models.

Confusion matrices are matrices that represent the number of instances that are classified in a known group (true label) in one dimension and the number of instances that are predicted to belong to a group (predicted label) in the other matrix dimension. To analyse the matrix, it is necessary to know the meaning of each cell of it. The first one is the number of instances that are labelled in the group one and are predicted by a model to be in the same group, which are called True Positives or TP; True Negatives or TN is the number of instances that belong to group zero and are predicted as members of this group; False Positives or FP is the number of instances predicted as group one when their original group is zero; and False Negatives or FN are the ones labelled as group one but predicted to be in group zero.

Accuracy represents the percentage of correctly classified instances (True Positives or TP and True Negatives or TN) over the total number of instances (TP, TN, False Positives or FP and False Negatives or FN), and its equation is: $accuracy = \frac{TN+TP}{TN+FP+FN+TP}$

Recall is the True Positive rate, a measure of correctly classified instances in the positive class. Its equation is: $recall = \frac{TP}{TP+FN}$

F1 is a metric that relates accuracy and recall. It is the harmonic mean of precision and recall, and it can be calculated with:

$$F1 = 2 * (accuracy * recall) / (accuracy + recall)$$

The best models were selected according to the highest F1 scores because it combines precision and recall measurements into a single value, and it is usually the best metric to work with unbalanced datasets (Mohammed, 2020).

In this process of selection, it was calculated the F1 score for a model that classifies all the individuals in the same class with the corresponding equation in order to obtain a basal value to compare the models' performance. The resulting score 0.15 is the baseline that has to be overcome by the models (Zach, 2021).

3.4 EXTRACTION OF SHAPLEY VALUES

Shapley values reflect the individual contribution of each value of the input variables in the difference between the actual prediction and the average prediction. These values provide an overall understanding of how the variables impact the model output for a particular instance (Smith, 2021).

In order to show the implication of each variable, it was used the shap library version 0.41.0 to extract the Shapley values and create some graphical representations: the bar plot, which presents the different variables in order of average impact on the model and the summary plot, which combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance, and the colour represents the value of the feature from low to high (Lundberg, 2017).

3.5 GENDER BASED ANALYSIS

After the selection of the three best models, the dataset was split by gender to perform a further variable analysis. The classifiers selected were applied to each gender dataset, also with the corresponding training and test sets and using cross validation. Finally, the Shapley values were extracted to compare between genders and with the general model.

4. RESULTS

In order to find more key factors in the development of depression, it was decided to use the data from the Barcelona Brain Health Initiative (BBHI) and select some of the self-reported health, socio-economic and lifestyles factors that were described in the literature having an implication in brain health. It was found that purpose in life and personal growth, two of the six factors composing the psychological wellbeing model proposed by Carol D. Ryff (1995) are related to the sense of coherence, which in turn is related to high levels of mental health (Cattaneo, 2022). These two variables are part of the vital plan test of the BBHI, whose questions are precisely from the Ryff Scales of Psychological Well-Being(Ryff, 1995).

The lack of social interaction can lead to psychological sequelae (Cabello-Toscano, 2023), so it was introduced as a variable to study from the Sleep and Social BBHI questionnaire which contains the Lubben Social Network test.

Also, cardiovascular diseases like heart attacks (Zhang, 2018), psychological diagnoses like anxiety (Choi, 2020) and neurodegenerative diseases like dementia (Bennett, 2014) have a usual comorbidity with depression, so several diagnoses variables were included in the data set: number of cardiovascular diseases, number of neurodegenerative disorders, number of psychiatric diagnoses, and total number of diagnoses.

Finally, it was found that early life stress is associated with an increased risk for the development of depressive disorder (LeMoult, 2020), so some questions related to childhood were also included among the rest of the factors to study, which can be found in the 3.3.1 Data selection section in materials and methods with the descriptions of all the variables used in this project.

To be able to assess the implications of the selected factors in the development of depression, it was proposed to create a classification model by using supervised machine learning techniques and calculate the Shapley values of these variables which gives the implication of each one in the depression classification model. Also, to compare this depression implication, it was included in the input dataset some of the already known variables that are related with depression diagnoses, such as physical activity (FPHS questionnaire)(Sutherland, 2019), sleep quality (Jenkins sleep scale)(Steiger, 2019), cognitive function (PROMIS cognitive abilities and cognitive concerns questions)(Talarowska, 2009), diabetes (Mahabadi, 2021), cancer (Chochinov, 2001), alcohol consumption (Boden, 2011), smoking habits (Paperwalla, 2004), nutrition (Bakırhan, 2022)(Mediterranean Diet Adherence Screener, MeDAS), age (Mirowsky, 1992), gender (Parker, 2010), medication for pain (Roughan, 2021) and general health (Middelboe, 1995) (PROMIS general health test). Consequently, a screening of models was conducted to select the best ones from which to calculate the significance of the variables. The models evaluated were logistic regression, random forest, gradient boosting, k-nearest neighbours, and multilayer perceptron, which are described in the 3.3.3 Modelling section from materials and methods.

The process of applying a machine learning algorithm involves several steps, including the input data preprocessing step. However, the imbalance between the number of participants who were diagnosed with depression and the number who were not can cause a problem in the prediction, making the model to classify the minority class (depression diagnosed people) as the majority class (non-depressed people). To solve this issue, it was decided to apply resampling methods to the input data, in this case, two undersampling methods, one non-heuristic like random undersampling and one heuristic as Tomek links, and two oversampling methods, random oversampling (non-heuristic) and SMOTE (heuristic), described in 3.3.2 Data preprocessing section of materials and methods (Mohammed, 2020). With that being said, all the models were tested with the data resampled by the four resampling techniques, and it was decided to maintain the non-heuristic methods for further analysis as they create the models with better metrics than the heuristic ones, as it is shown in the following table that includes the F1 score for all the combinations between the four resampling methods and the five classifiers (Table 2).

Table 2. F1 scores from different combinations of models and resampling

	basal	random undersampling	Tomek links	random oversampling	SMOTE
logistic regression	0.0	0.220	0.0	0.245	0.225
random forest	0.0	0.232	0.0	0.172	0.127
gradient boosting	0.1	0.234	0.021	0.140	0.135
knn	0.2	0.182	0.022	0.077	0.167
MLP	0.3	0.229	0.021	0.144	0.175

The following is a further explanation of the models created by training them with three different data sets, one with the unaltered data, one with the data resampled by random undersampling and another with random oversampling.

4.1 LOGISTIC REGRESSION

The best logistic regression models obtained by training with the unaltered data and the two types of resampling have the parameters shown in the Table 3.

Table 3. Logistic regression model parameters.

	Base data	Random undersampling	Random oversampling
C	0.001	0.01	0.01
max_iter	100	200	100
solver	sag	saga	saga

As it is shown in the table, the low C value denotes there is a high regularization strength in all the models to avoid overfitting and reduce model complexity. The solvers introduced sag and saga are both adequate for the large input dataset.

With respect to the metrics, we can see in Table 4 how the accuracy was higher in the baseline model, however, both F1 and recall gave values of zero. This is because the model classified all instances of the test set as the negative class (no depression) which is the majority class.

Table 4. Logistic regression model evaluation metrics.

	Base data	Random undersampling	Random oversampling
accuracy	0.921	0.684	0.709
recall	0	0.568	0.602
f1	0	0.220	0.245

On the other hand, we can see how the model trained with the random oversampling data provides higher values of accuracy, recall and F1 values, so it is the best logistic regression model (Table 4).

4.2 RANDOM FOREST

In the case of random forest classifier, the parameters selected by cross-validation for the three models, were the ones showed in the Table 5.

Table 5. Random forest model parameters.

	Base data	Random undersampling	Random oversampling
criterion	gini	gini	entropy
max_depth	10	10	15
min_samples_leaf	1	2	1
min_samples_split	10	5	2
n_estimators	50	25	100

The basal model and the random undersampling model use gini criterion uses the Gini impurity as the function to measure the quality of a split which measures the frequency at which any element of the dataset will be mislabelled when it is randomly labelled, meanwhile, the random oversampling model uses entropy criterion, that uses the Shannon information gain, which minimizes the log loss between the true labels and the probabilistic predictions of the tree model (Pedregosa, 2011). The depth of 10-15 indicates that the generated trees are large, but it makes sense since there are thirty variables in the input table.

This time, the metrics obtained do not point to a clearly better model. According to accuracy, the best model is the random oversampling one, but recall and F1 are higher with the random undersampling model. Since we are taking F1 as the most complete metric, the best random forest model is the one trained with the random undersampling resampling method (Table 6).

Table 6. Random forest model evaluation metrics.

	Random forest metrics		
	Base data	Random undersampling	Random oversampling
accuracy	0.918	0.675	0.914
recall	0	0.625	0.113
f1	0	0.232	0.172

4.3 GRADIENT BOOSTING

With the gradient boosting classifier, the parameters selected are the following (Table 7):

Table 7. Gradient boosting model parameters.

	Base data	Random undersampling	Random oversampling
min_samples_leaf	10	5	5
min_samples_split	2	2	2
max_depth	2	2	6
n_estimators	100	50	100
learning_rate	0,1	0,1	1

In this case, the max depth indicates that the trees that were used to build the gradient boosting models were smaller than the ones for the random forest models. Respecting the learning rate, a low value like 0.1 is considered better than one because it attributes a smaller weight to each tree contribution and makes the model to perform more iterations to converge.

Looking the metrics, it is the same case as random forest, the best model was obtained using the random undersampling technique because it is the one with the highest F1 score (Table 8). The F1 score of the random oversampling model is lower than the F1 value of a baseline model that classifies all the instances as the positive class (with depression), which is 0.15.

Table 8. Gradient boosting model evaluation metrics.

	Gradient boosting metrics		
	Base data	Random undersampling	Random oversampling
accuracy	0.918	0.684	0.901
recall	0	0.613	0.102
f1	0	0.234	0.140

4.4 K-NEAREST NEIGHBOURS

The parameters obtained for the three KNN models are shown in the Table 9.

Table 9. K-nearest neighbours model parameters.

	Base data	Random undersampling	Random oversampling
n_neighbors	20	7	2

The basal model needs twenty neighbours to classify. A higher number of neighbours to consider is better for classification as it leads to smoothening the decision boundaries, by contrast, the small number of neighbours of the random oversampling model leads to unstable decision boundaries.

The best model for this classifier is the random undersampling one. The F1 score of the random oversampling model has the same issue as the one of the random oversampling random forest model. It is lower than the F1 value of the baseline model (Table 10).

Table 10. K-nearest neighbours model evaluation metrics.

	K-nearest neighbours metrics		
	Base data	Random undersampling	Random oversampling
accuracy	0.921	0.624	0.851
recall	0	0.534	0.079
f1	0	0.182	0.077

4.5 MULTILAYER PERCEPTRON

The parameters of the MLP models are shown in the following Table 11.

Table 11. Multilayer perceptron model parameters.

	Base data	Random undersampling	Random oversampling
activation	identity	indentity	indentity
hidden_layers_sizes	50	50	100
max_iter	5000	5000	5000
solver	lbfgs	adam	adam

Both the basal model and the random undersampling one have an identity activation function, which means that there are no changes in the input values, meanwhile, logistic activation function converts the input data into a probability value between 0 and 1. The high number of hidden layers in all the models indicates that the complexity is also high. The solver adam works pretty well on relatively large datasets so it makes sense that it is the selected parameter for the models trained with resampled data (Pedregosa, 2011).

Once again, the best model is the one trained with data resampled with the random undersampling algorithm because is the only one that gives a F1 score higher than the 0.15 baseline (Table 12).

Table 12. Multilayer perceptron model evaluation metrics.

	Multilayer perceptron metrics		
	Base data	Random undersampling	Random oversampling
accuracy	0.921	0.788	0.871
recall	0	0.465	0.284
f1	0	0.257	0.257

4.6 MODEL SELECTION AND VARIABLE RELEVANCE

Since the main objective is focused on finding new potential depression predictors, we selected three models instead of the best one to compare the variables importance between them. The models were selected by their F1 score, being the multilayer perceptron model trained with the data resampled by random oversampling the highest with 0.2577, followed by the multilayer perceptron/undersampling with 0.2570 and the logistic regression/ oversampling with 0.245. Nevertheless, the confusion matrices from these models were extracted and it was decided to discard the two Multilayer perceptron models due to the fact that the percentage of False negatives was higher than the True positives one, which means that there are more depressed people classified as non-depressed than classified as depressed (Figure 1).

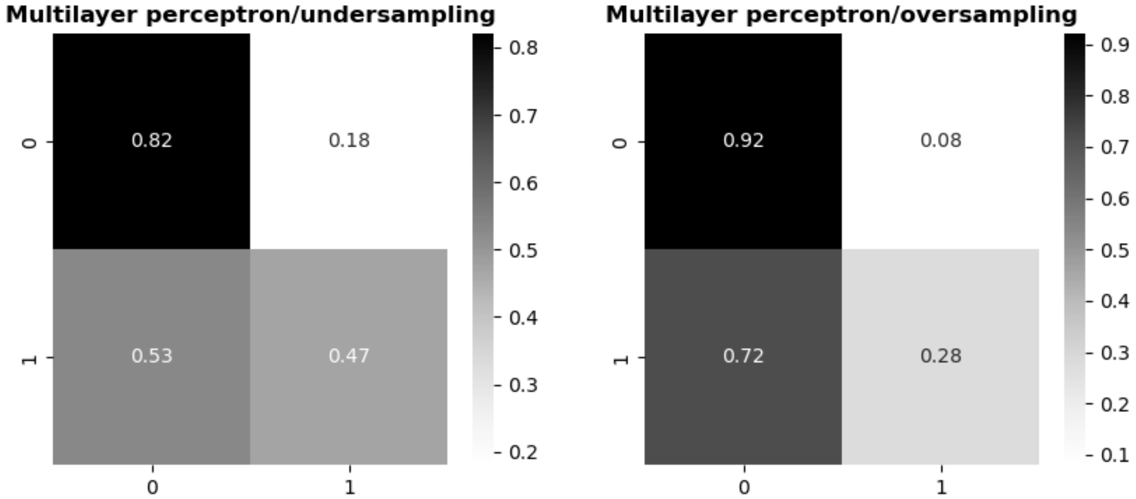

Figure 1. Confusion matrices of the MLP/undersampled data and MLP/oversampled data models.

In consequence, the following two models with higher F1 score were selected to continue the analysis, the random forest/undersampling and the gradient boosting/undersampling. The confusion matrices of these models were also represented, but this time, the True positives were greater in proportion than the False negatives (Figure 2).

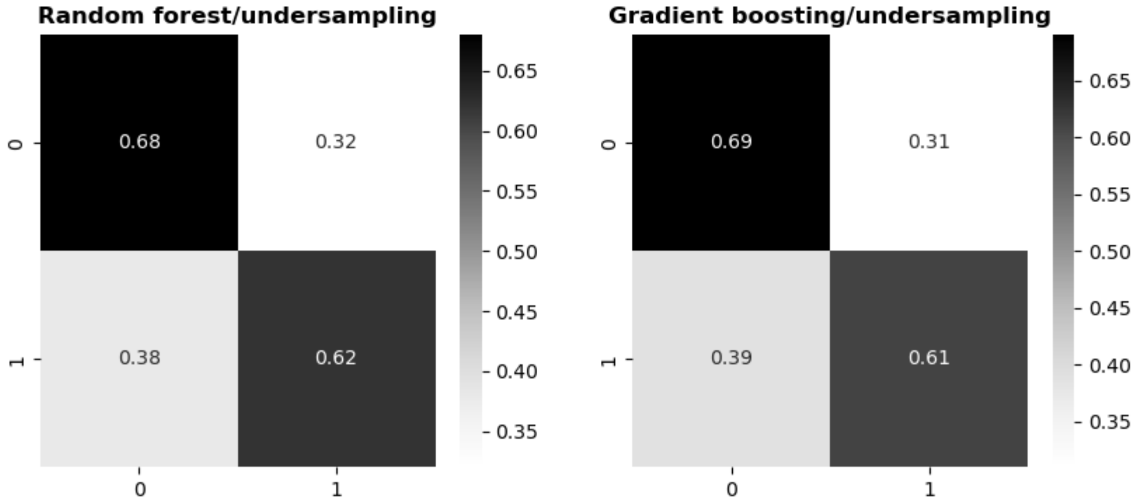

Figure 2. Confusion matrices of the Random Forest/undersampled data and Gradient boosting/oversampled data models.

Then, the Shapley values for these models were calculated. As we can see in the following bar plot, the importance of the variables in the logistic regression/random oversampling shows that the factors of psychiatric diagnoses, the sixth question related to childhood, that was “What was your father´s usual job / job?”, total diagnoses and purpose in life have high Shapley values, similar to gender and sleep quality, only surpassed by global health (Figure 3).

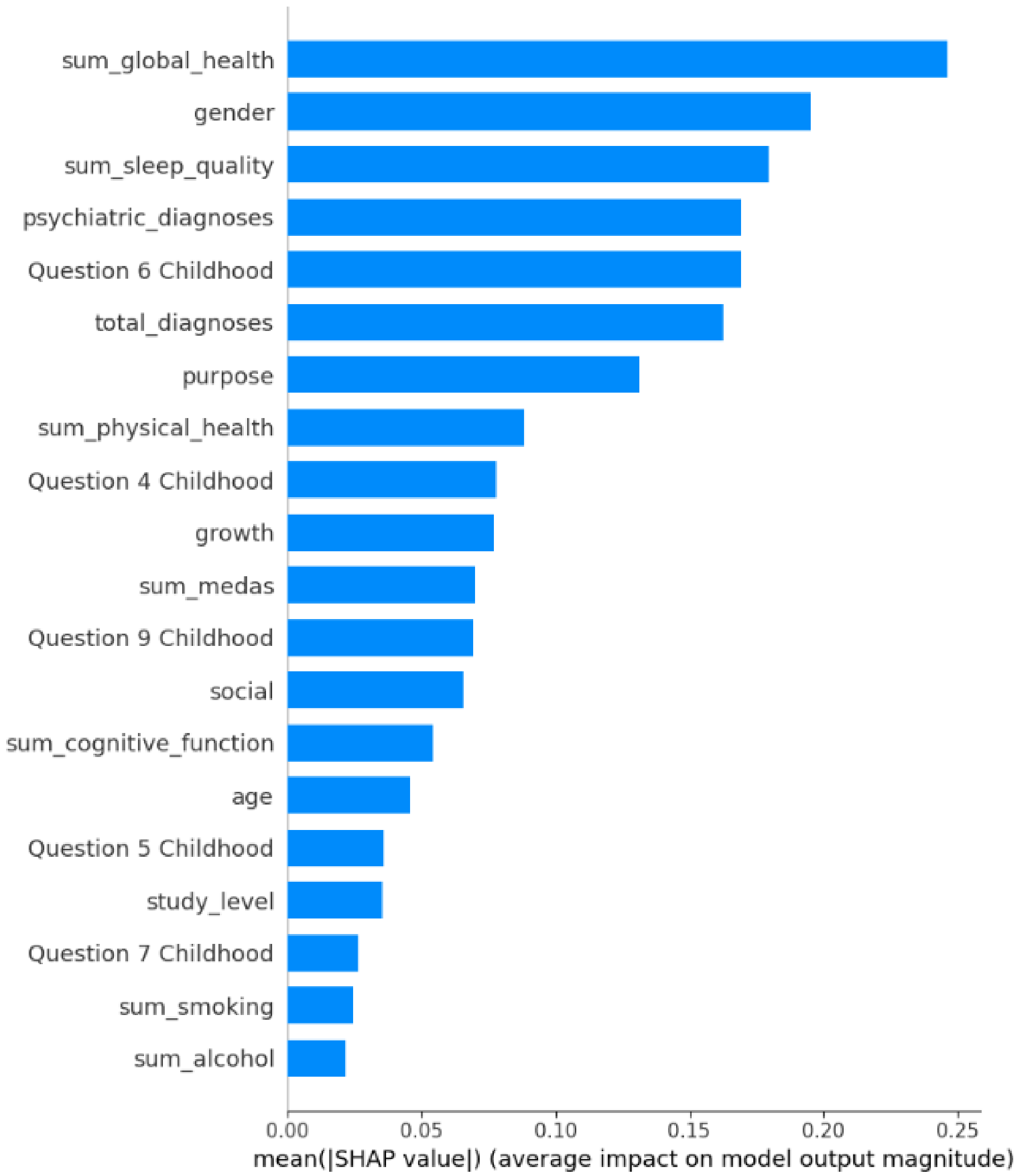

Figure 3. Shapley values in logistic regression model trained with Random oversampled data.

In the next graphic we can see not only the relevance order of the variables, but also the effects of the factor values (Figure 4). A high number of psychiatric diagnoses or total diagnoses increases the probability of depression. If the father was a householder or had an unpaid job during childhood significantly reduces the risk of depression, as opposed to having a job as a director or manager with university education. And a low score of purpose in life scale, which means that the person lacks a sense of meaning in life and sense of direction, has few goals or aims and does not see purpose of past life, increases the depression risk.

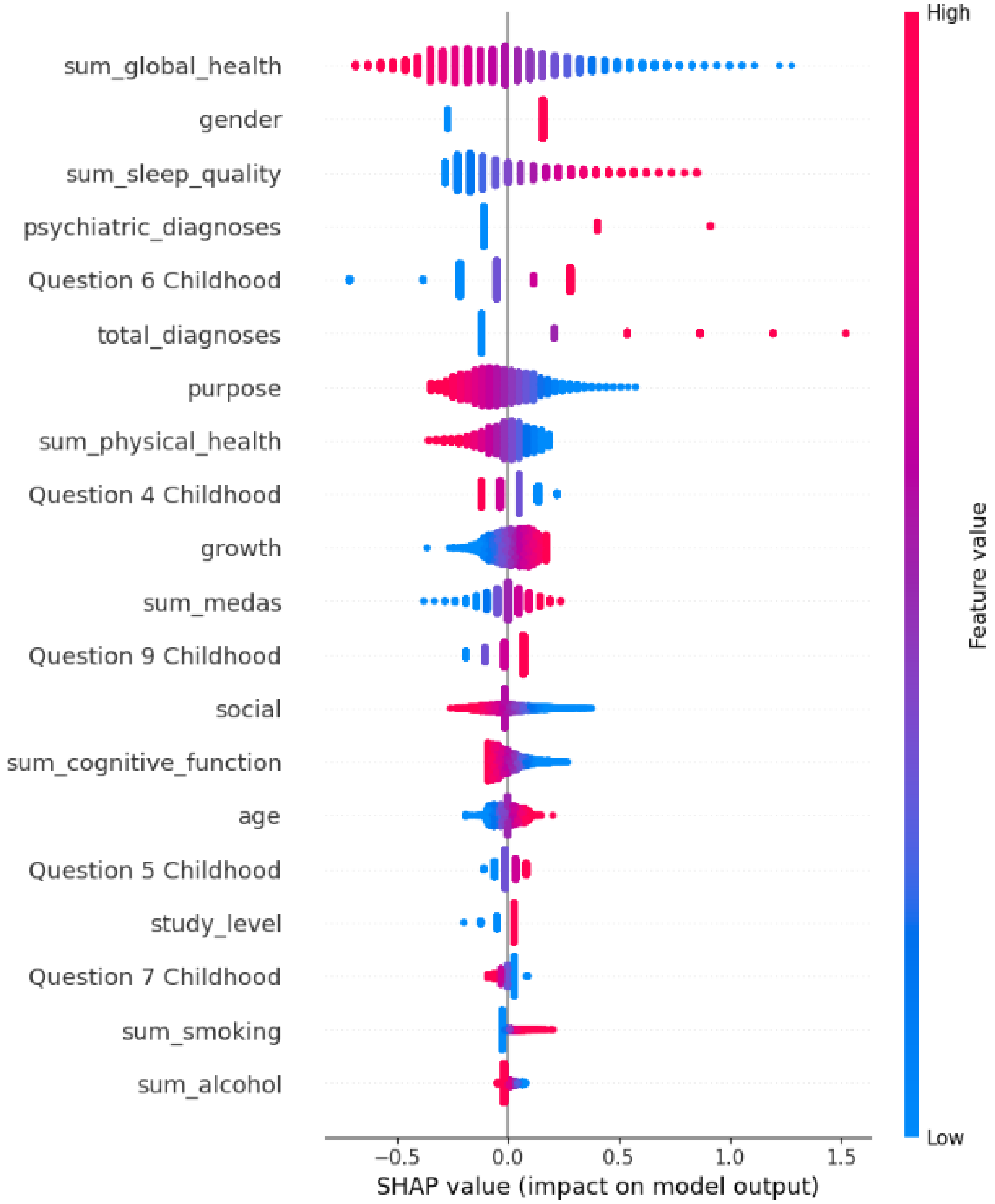

Figure 4. Feature importance vs feature effects in Logistic regression/Random oversampling model.

In the case of the random forest model trained by random undersampled data, the highest Shapley values correspond to psychiatric diagnoses. Other factors of those under study that received high Shapley values are social interactions, purpose in life and total diagnoses (Figure 5).

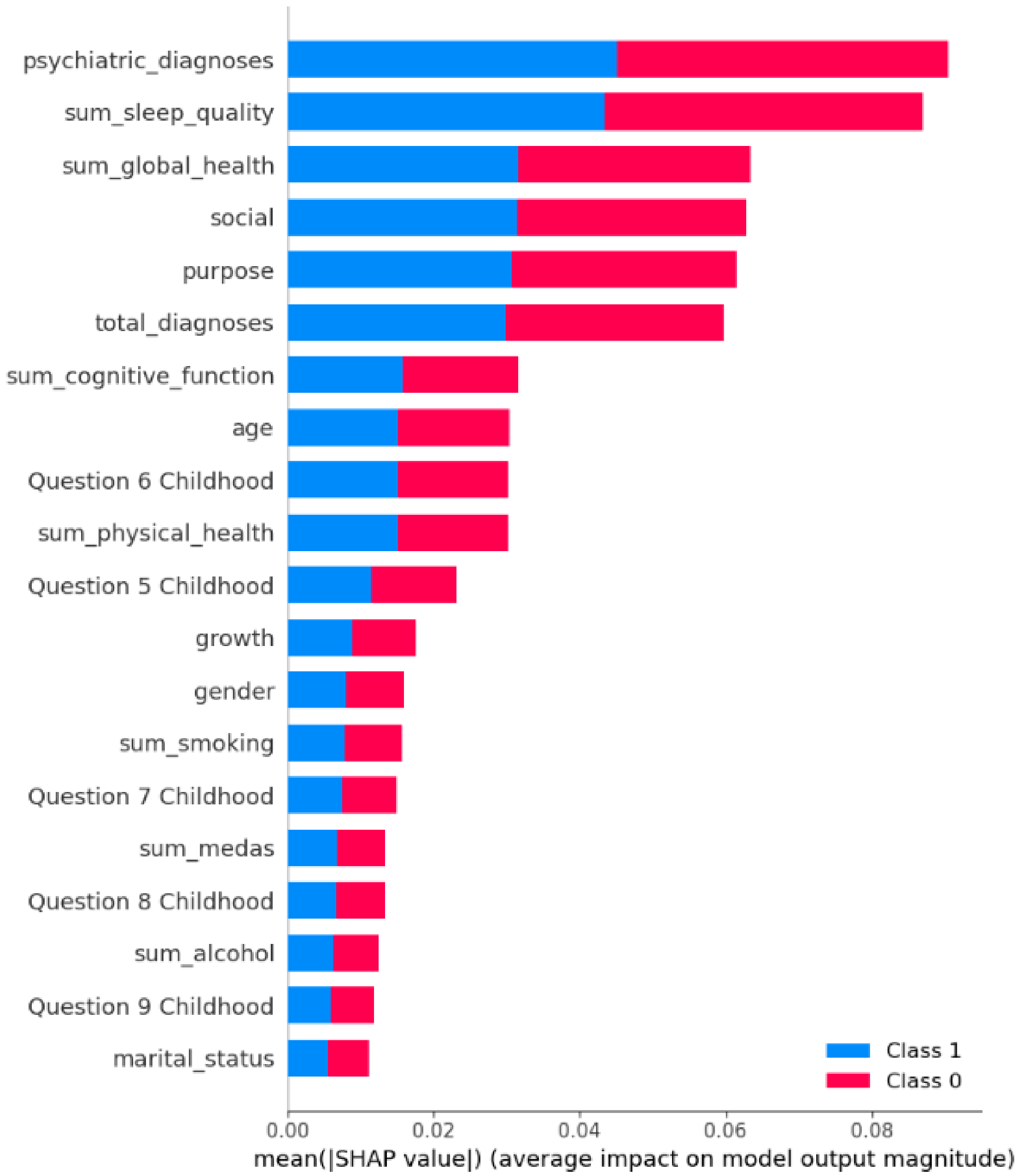

Figure 5. Shapley values in random forest model trained with Random undersampled data.

As the previous summary plot from logistic regression model, the increase in the number of psychiatric and total diagnoses makes the development of depression more likely to occur and the lack of purpose in life also increases the depression risk. This time it is clearer that higher scores in life purpose, which means that the person has goals in life and a sense of directedness and feels there is meaning to present and past life, decreases the probability of having depression. In this random forest model, social interaction has a high importance, and it is shown that an intermediate level of social engagement decreases the depression risk, a high level of social engagement does not affect the probability of depression, and a low level of social interaction increases its risk (Figure 6).

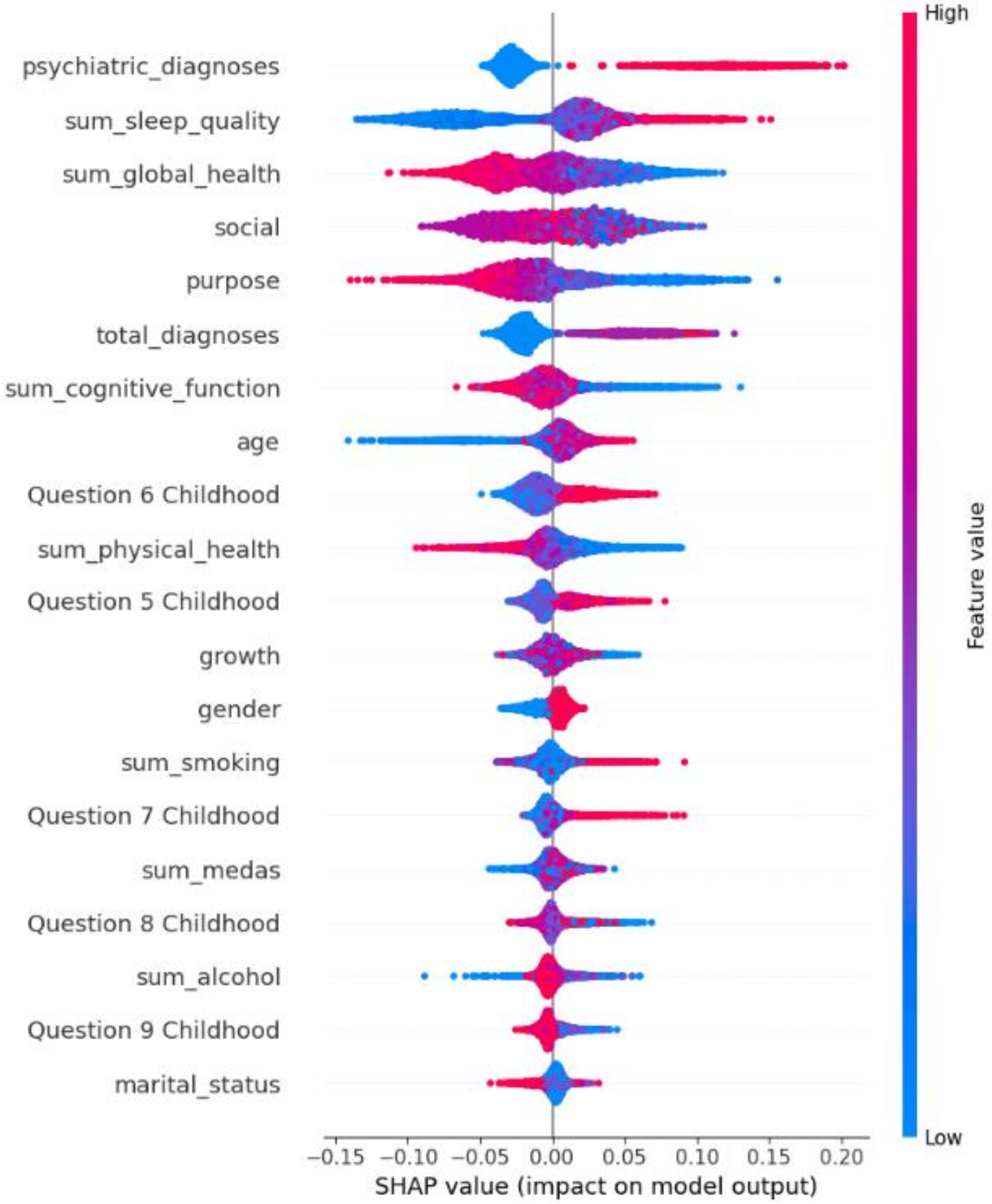

Figure 6. Feature importance vs feature effects in Random forest/Random undersampling model.

Finally, looking at the bar plot of the gradient boosting model trained with Random undersampled data, we can see that psychiatric diagnoses is, once again the most important variable according to the Shapley values. Social interaction repeats as a high value and total diagnoses and the childhood question related to the father's job (Figure 7).

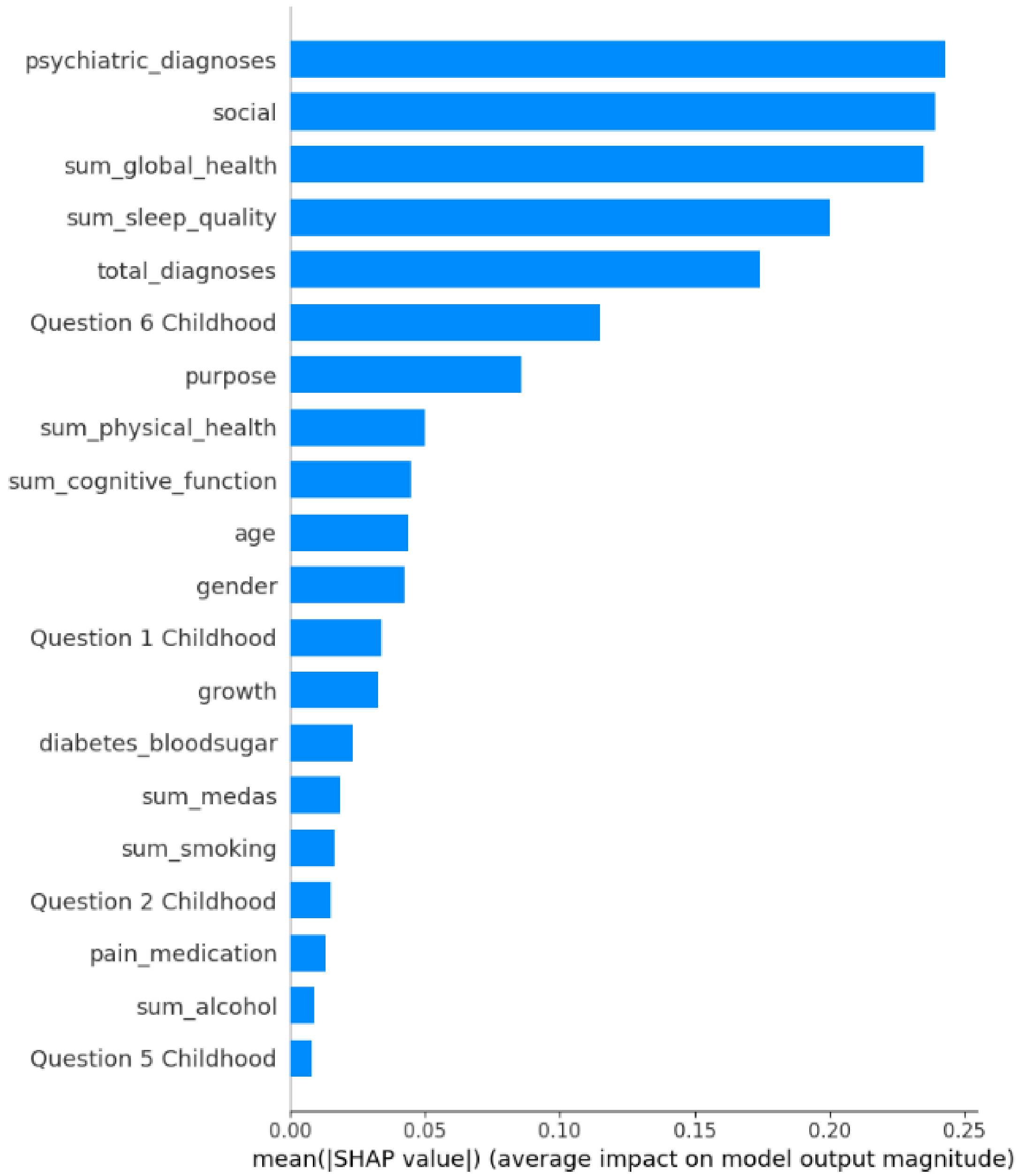

Figure 7. Shapley values in gradient boosting model trained with Random undersampled data.

Confirming the results from the last two models, a higher number of total and psychiatric diagnoses, low levels of social interaction, the lack of life purpose and the highly qualified father's job during childhood increase the risk of depression diagnosis. Also, it is important to point out the effects of the already studied variables in the development of depression. A poor global health increases depression incidence, while a good global health decreases it; a low score in Jenkins sleep scale indicates that the sleeping quality is high and the graphic shows that it decreases depression risk; regular physical activity maintains low the depression risk; and problems in cognitive function increases the probability of developing depression (Figure 8).

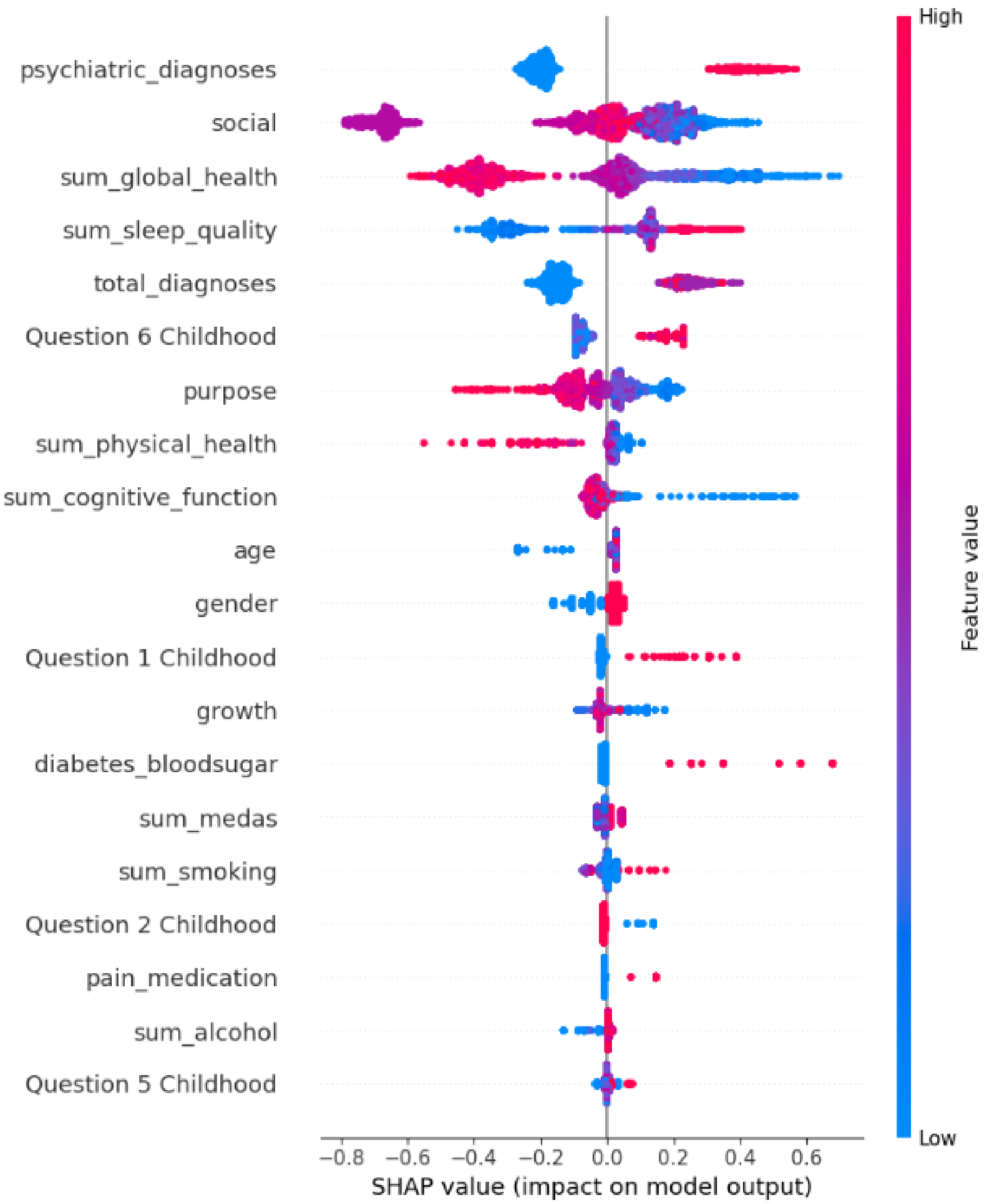

Figure 8. Feature importance vs feature effects in Gradient boosting/Random undersampling model.

Among the models, there are some variables in common from those that are under study, stand out psychiatric diagnoses, social interactions, purpose in life, total diagnoses and the father's job during childhood with them being near the top of relevance at least in two of the three selected models. In contrast, factors that are known to be related to depression diagnoses like nutrition, smoking and alcohol consumption, are at the bottom of the graphics, which implies that they have a low contribution to the construction of the model. Some of the input variables don't appear in the bar plots, this is because the function only shows the twenty more relevant variables, in any case, these factors have Shapley values near 0, which means that their implication in the model is insignificant.

4.7 STUDY OF VARIABLE RELEVANCE BY GENDER

To perform an analysis of the models by gender, the input data was divided into men and women datasets and the previously selected models were trained with these datasets.

The men's dataset consisted in 1805 instances divided in 82 men labelled with depression and 1723 instances of men without depression. Meanwhile, the women's dataset has a number of 3793 instances, 379 depression labelled and 3414 instances without depression.

The first model is Logistic regression with random oversampling. The F1 score obtained with the men dataset is worse than the original dataset score but overcomes the baseline model value for the men dataset (0.08), that classifies all the individuals as the same class. However, the women dataset obtained better metrics in general (Table 13).

Table 13. Logistic regression model metrics by gender.

	Logistic regression metrics	
	Men	Women
accuracy	0.684	0.719
recall	0.5	0.609
f1	0.109	0.319

The order of variable relevance in the logistic regression model trained with the men dataset is different from the one of the original dataset. Number of psychiatric diagnoses decreases its contribution and appear two childhood questions as highly important, "What was your father´s level of education?" (Question 4 Childhood) and "Did you live most of the time in a town or a city as a child?" (Question 9 Childhood). It also maintains the social interaction and life purpose relevance (Figure 9).

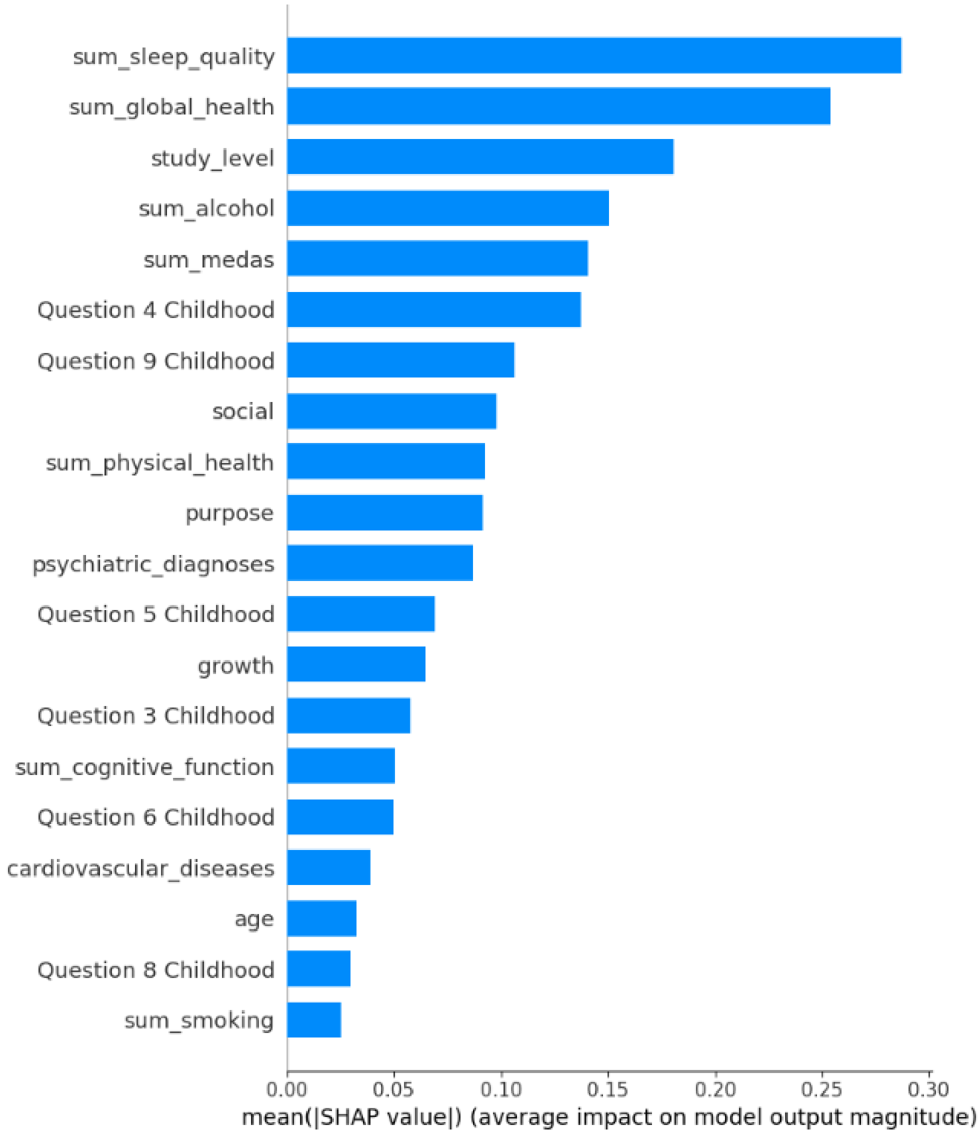

Figure 9. Shapley values in logistic regression model trained with men dataset.

In the following plot, it is shown the implication of the values of each variable in the model. There are some know variables with high relevance different from the original one. A low study level or no studies reduce the probability of developing depression. Even though the alcohol consumption is relevant, their values do not significantly change the model output, even so lower values in this variable means higher level of consumption and slightly increases depression risk. The nutrition variable shows that a low level of adherence to Mediterranean diet decreases the depression probability. In this case, high values of social interaction reduce the depression risk, contrary to low values (Figure 10).

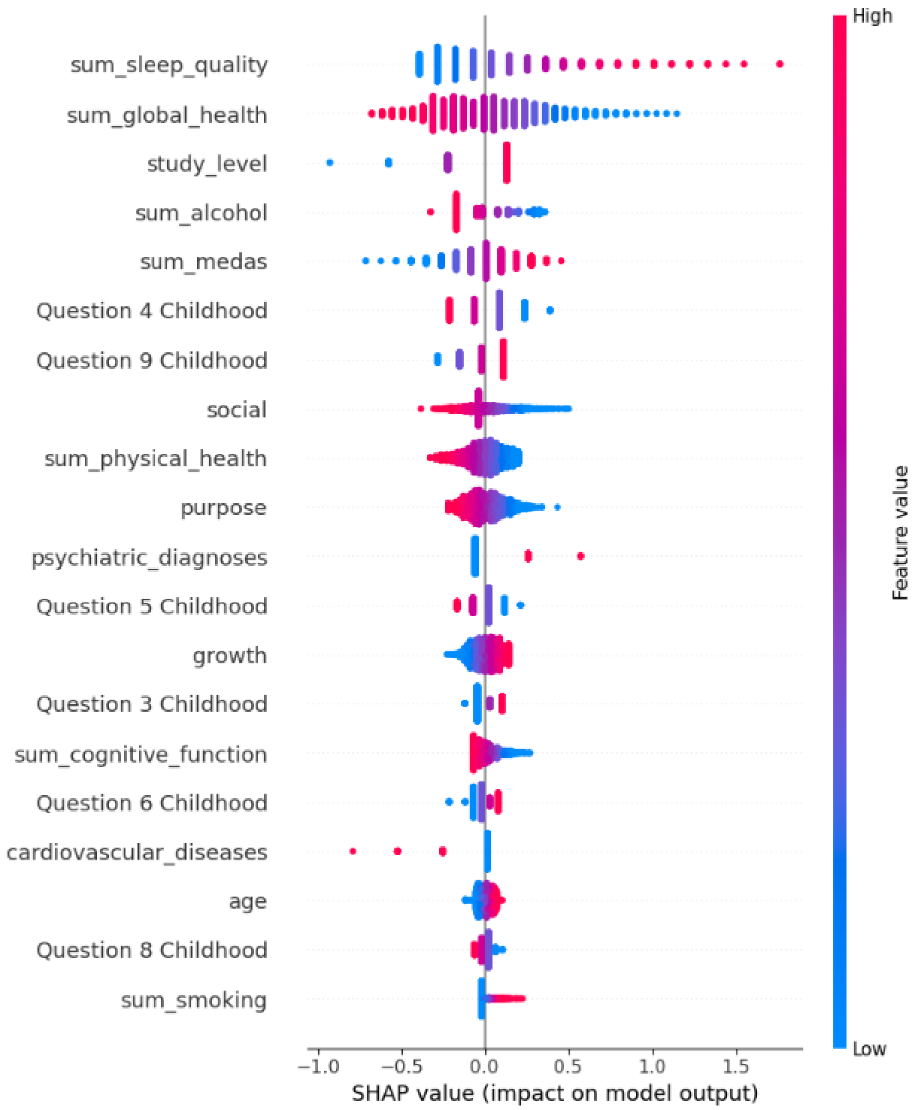

Figure 10. Feature importance vs feature effects in Logistic regression model trained with men dataset.

In the model trained with women's dataset, the variable relevance is more similar to the original dataset. Nevertheless, age takes relevance and some of the childhood questions. Mostly the "The house where you habitually lived during your childhood, how many rooms did you have?" item (Question 8 Childhood) (Figure 11).

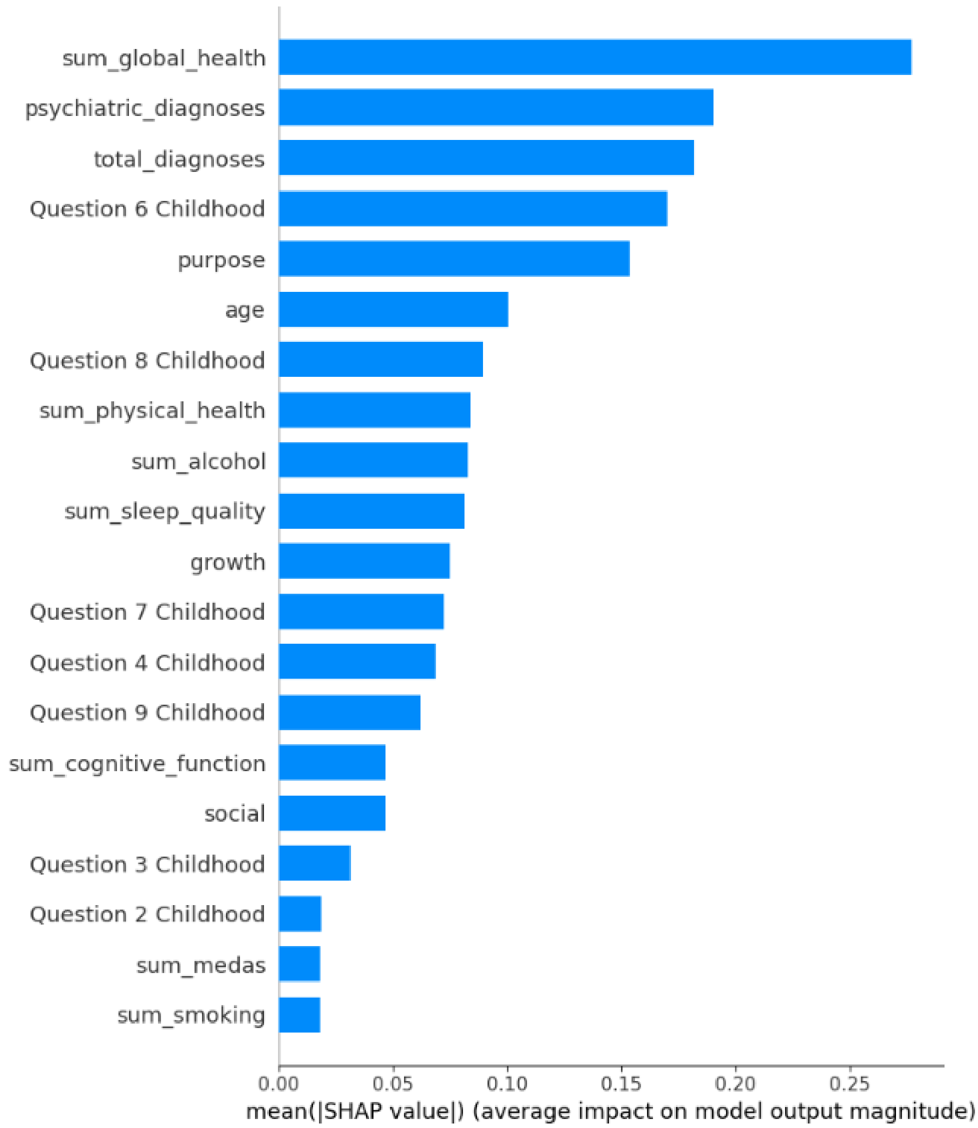

Figure 11. Shapley values in logistic regression model trained with women dataset.

The following plot reveals that older women have more probabilities of developing depression and a smaller number of rooms in the childhood residency decreases depression risk in women. And the alcohol consumption is inverted as the previous model, higher consumption decreases depression risk (Figure 12).

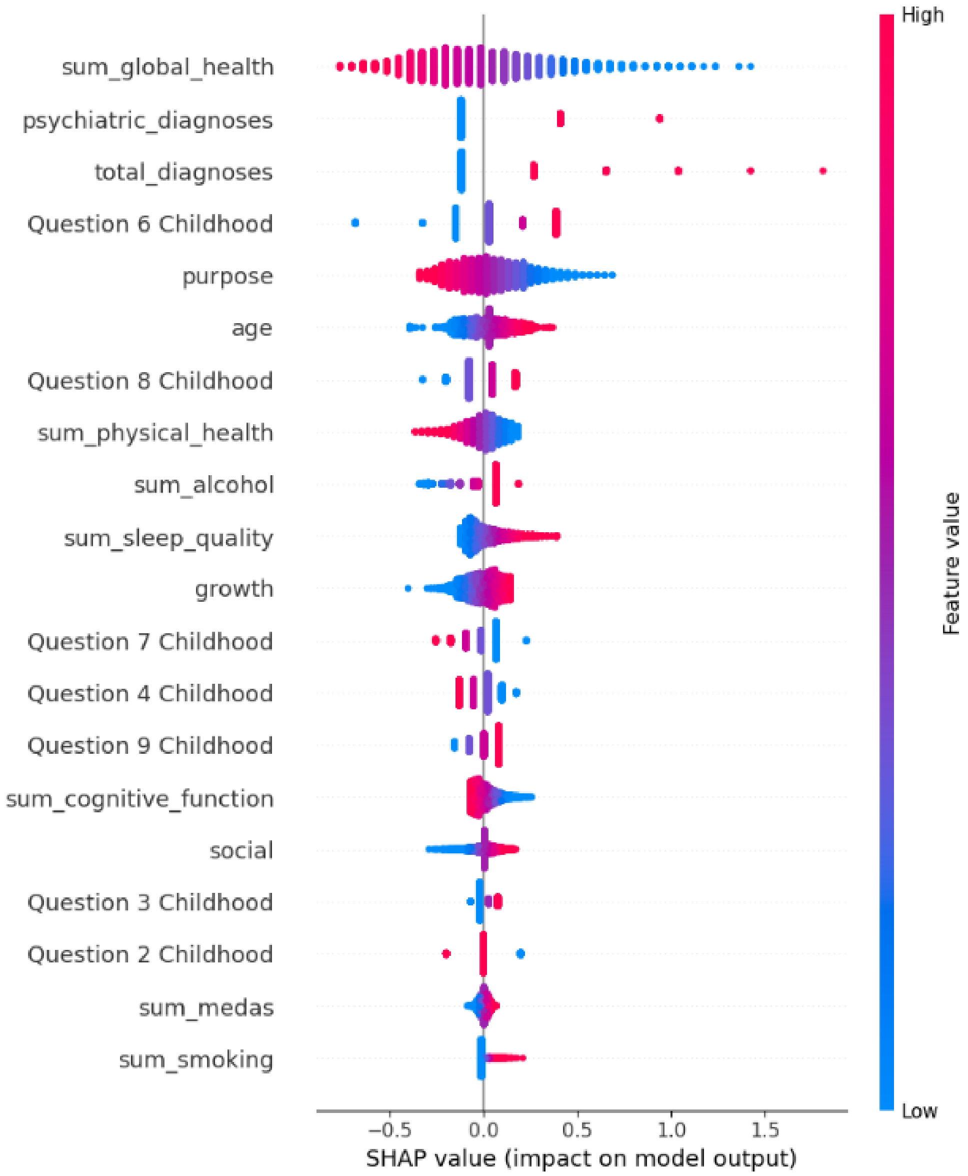

Figure 12. Feature importance vs feature effects in Logistic regression model trained with women dataset.

The second classifier is Random Forest with random undersampling. The F1 score obtained with the men dataset is again worse than the baseline model. However, the women dataset obtained better F1 score and recall (Table 14).

Table 14. Random forest model metrics by gender.

	Random forest metrics	
	Men	Women
accuracy	0.587	0.678
recall	0.714	0.646
f1	0.118	0.302

In the Random Forest model of the men's dataset, the studied variables purpose in life and personal growth take relevance, as well as the question 6 about childhood that asks the father's job. As the previous model, variables that were important with the complete dataset now have less relevance. It is the case of psychiatric and total diagnoses (Figure 13).

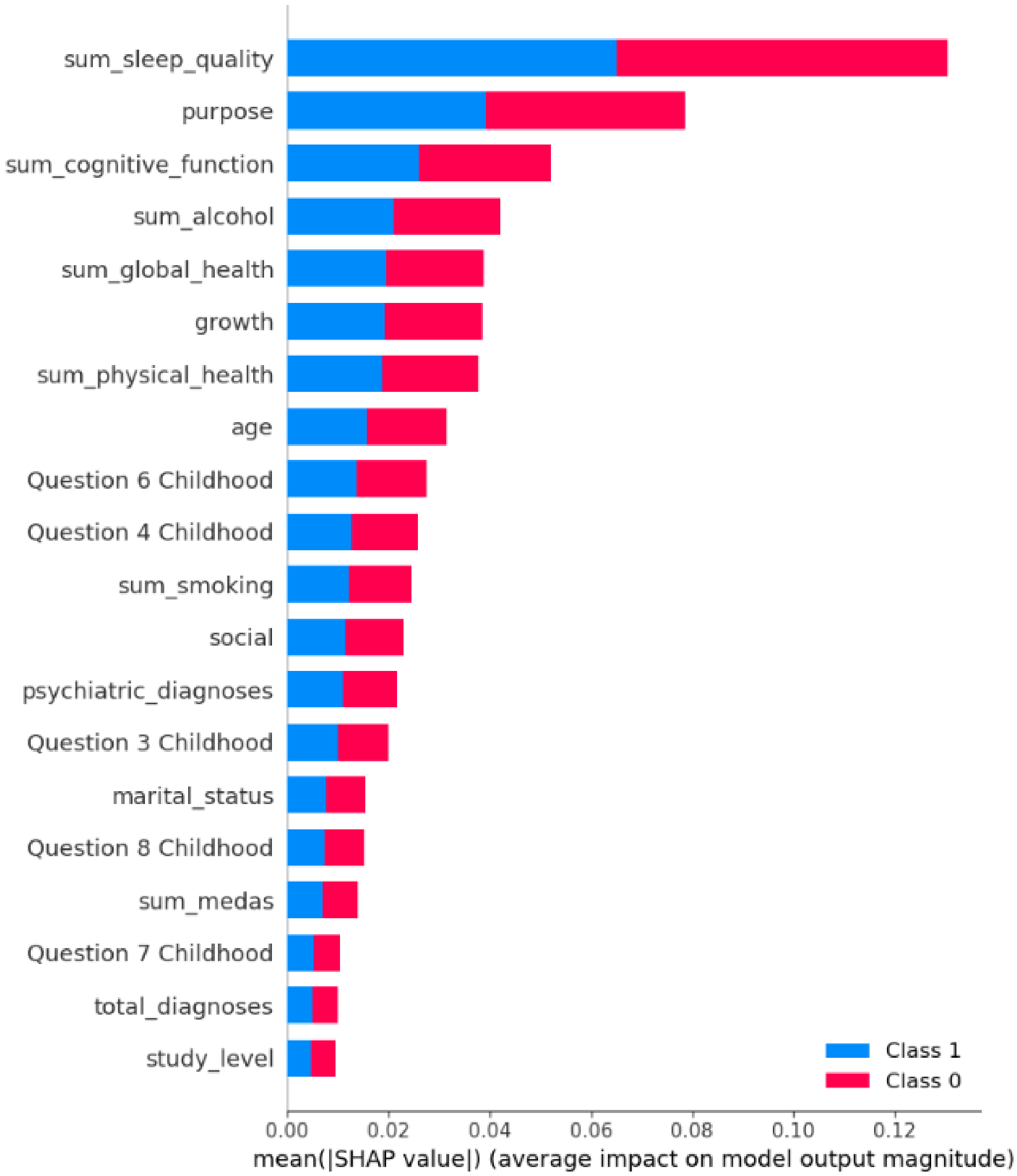

Figure 13. Shapley values in Random Forest model trained with men dataset.

As shown in the following graph, the greater the sleep quality (low Jenkins score) the lesser depression development probability for men. The lack of purpose in life increases the depression risk. In this model, cognitive function takes relevance and low levels of cognitive function increase the depression development. And higher alcohol intake also increases the depression probability. The lower values of personal growth tend to increase the depression risk, but these blue dots are also in the decrease section, so the effect of this variable is not clear (Figure 14).

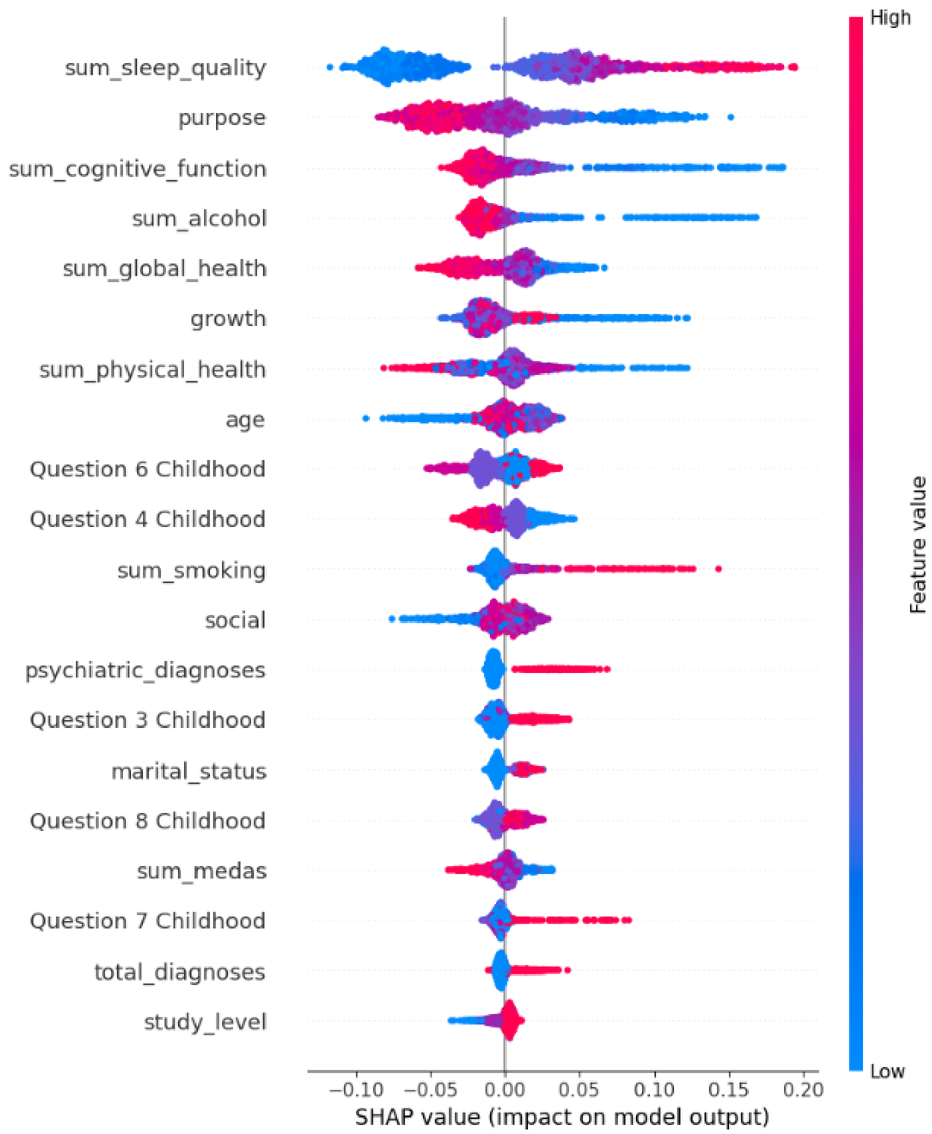

Figure 14. Feature importance vs feature effects in Random Forest model trained with men dataset.

Results obtained for the Random Forest model trained with the women's dataset, are very similar to those from the men's dataset. The only differences are the position of psychiatric diagnoses, more relevant in women's model, and the alcohol consumption, more relevant in the men's model (Figure 15).

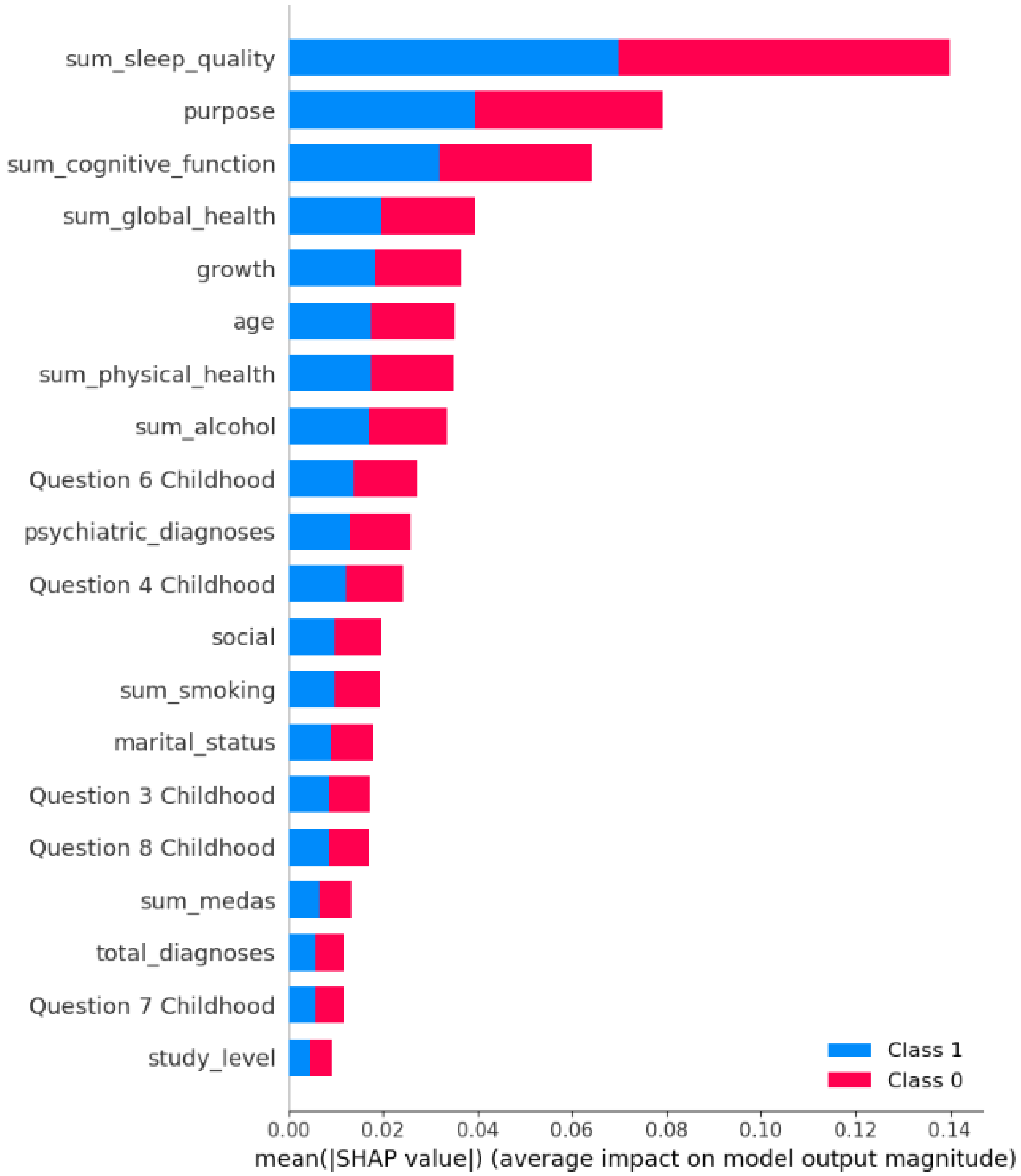

Figure 15. Shapley values in Random Forest model trained with women dataset.

As it is shown in the Figure 16, younger women tend to be less depressed. It is also remarkable that this time, the intermediate levels of the question about the father's job during childhood decrease the depression diagnosis. These levels correspond to Qualified manual labour and Professional (with university studies).

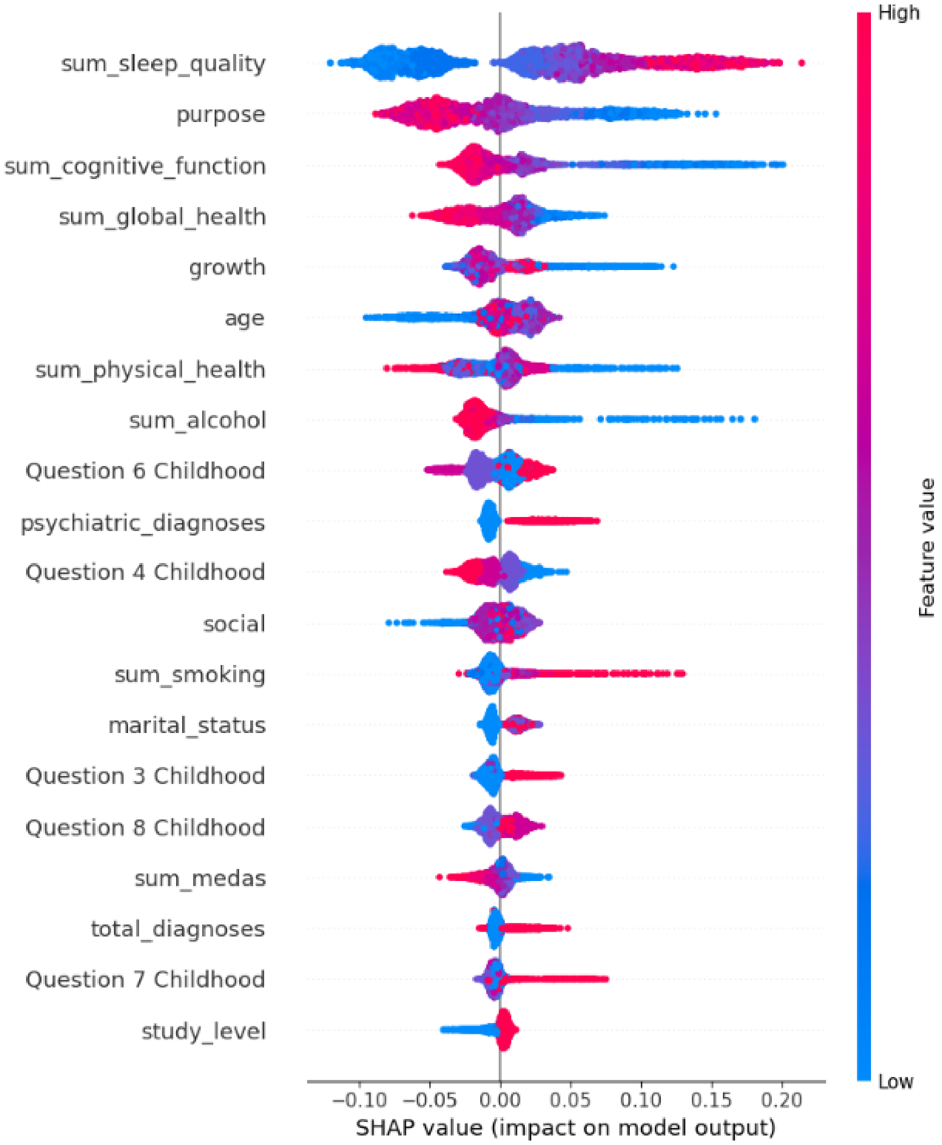

Figure 16. Feature importance vs feature effects in Random Forest model trained with women dataset.

The last classifier is Gradient boosting with random undersampling. The F1 score obtained with the men dataset is worse than the baseline model that classifies all the individuals as the same class. In the other hand, the women dataset obtained better metrics in general (Table 15).

Table 15. Gradient boosting model metrics by gender.

	Gradien boosting metrics	
	Men	Women
accuracy	0.601	0.691
recall	0.714	0.658
f1	0.121	0.315

Finally, the last men's model gradient boosting reveals the relevance of some of the studied variables, life purpose, social interactions, personal growth and the sixth childhood item (father's job). Also, among the other variables, takes importance smoking, alcohol, cognitive function and age (Figure 17).

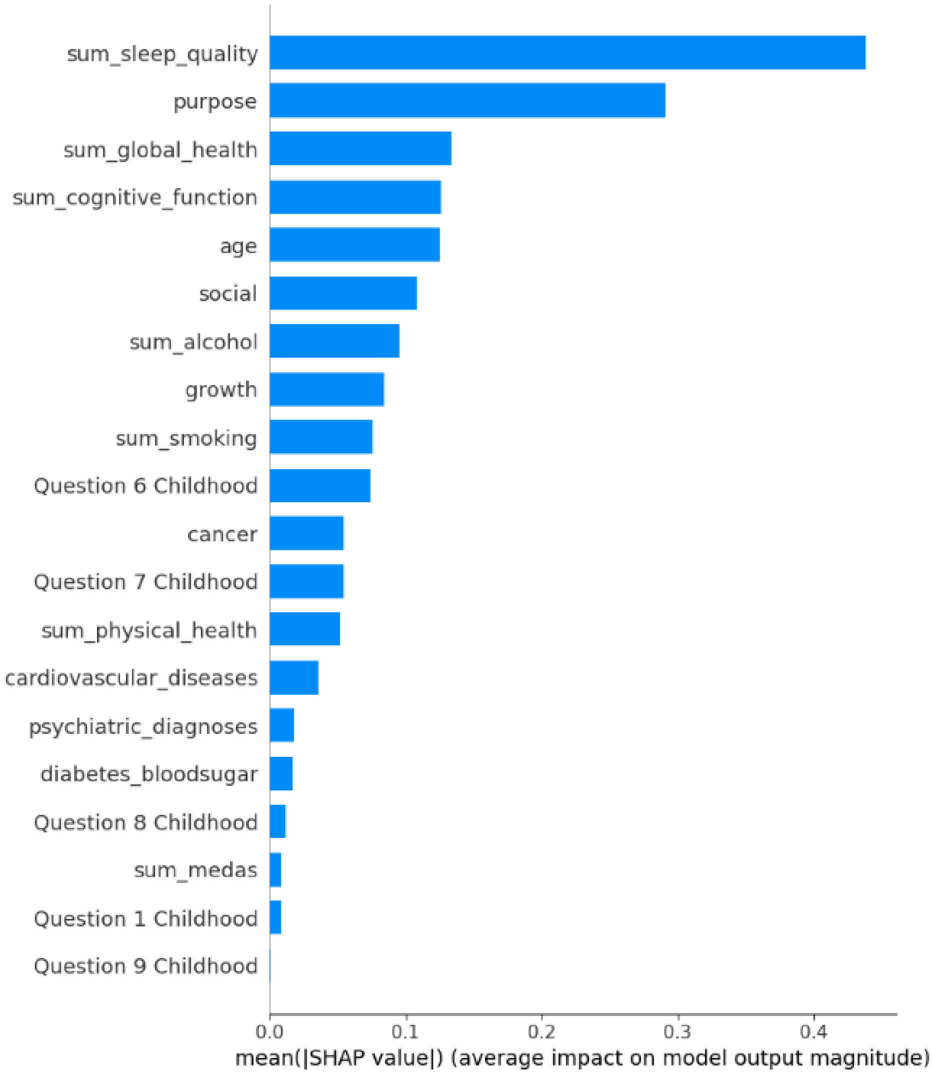

Figure 17. Shapley values in Gradient boosting model trained with men dataset.

As stated in the previous models, a lack of life purpose and a poor sleep quality increase the depression probability. It is remarkable that high values of the smoking variable, which translates to continuous smoking habits increase the depression risk; and a higher number of cancer suffer increases the depression diagnoses (Figure 18).

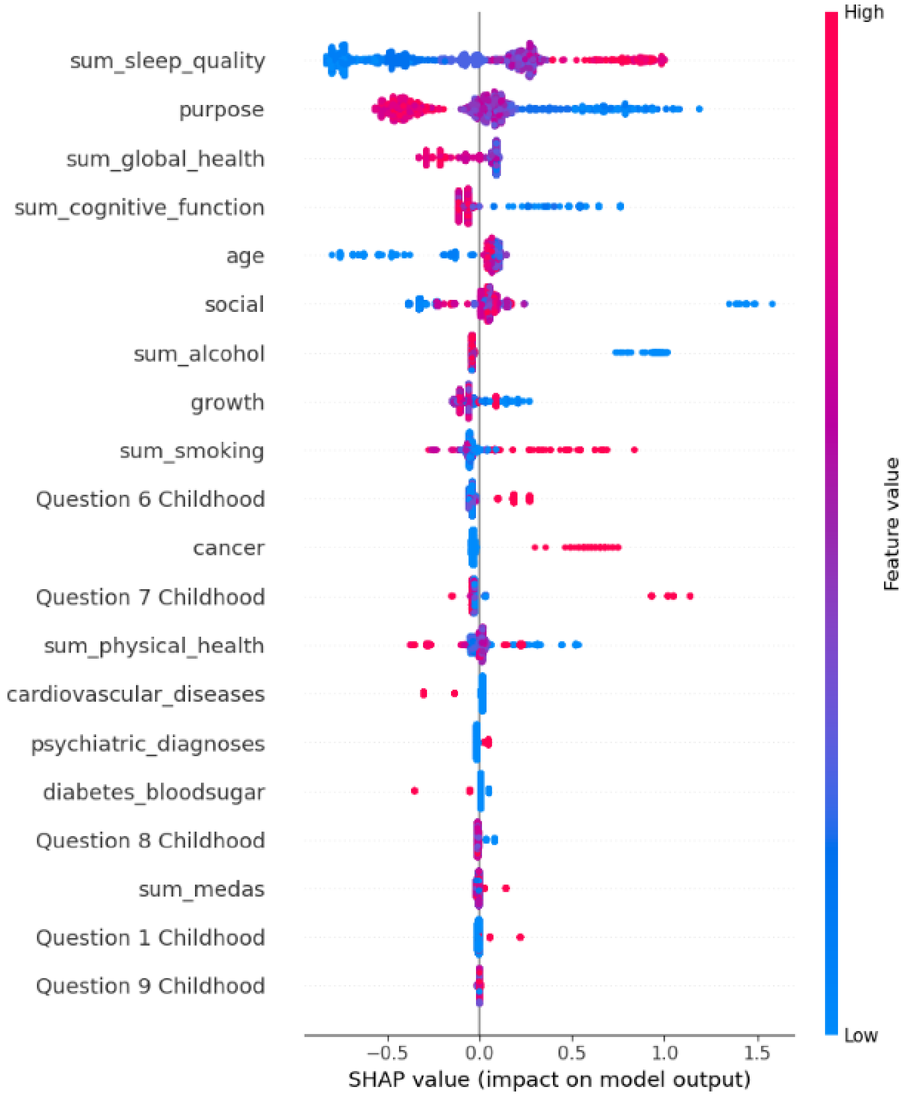

Figure 18. Feature importance vs feature effects in Gradient boosting model trained with men dataset.

On the other hand, the gradient boosting women's model is heavily influenced by the social interaction variable, the number of psychiatric and total diagnoses, life purpose and the ninth childhood question, which is "Did you live most of the time in a town or a city as a child?" (Figure 19).

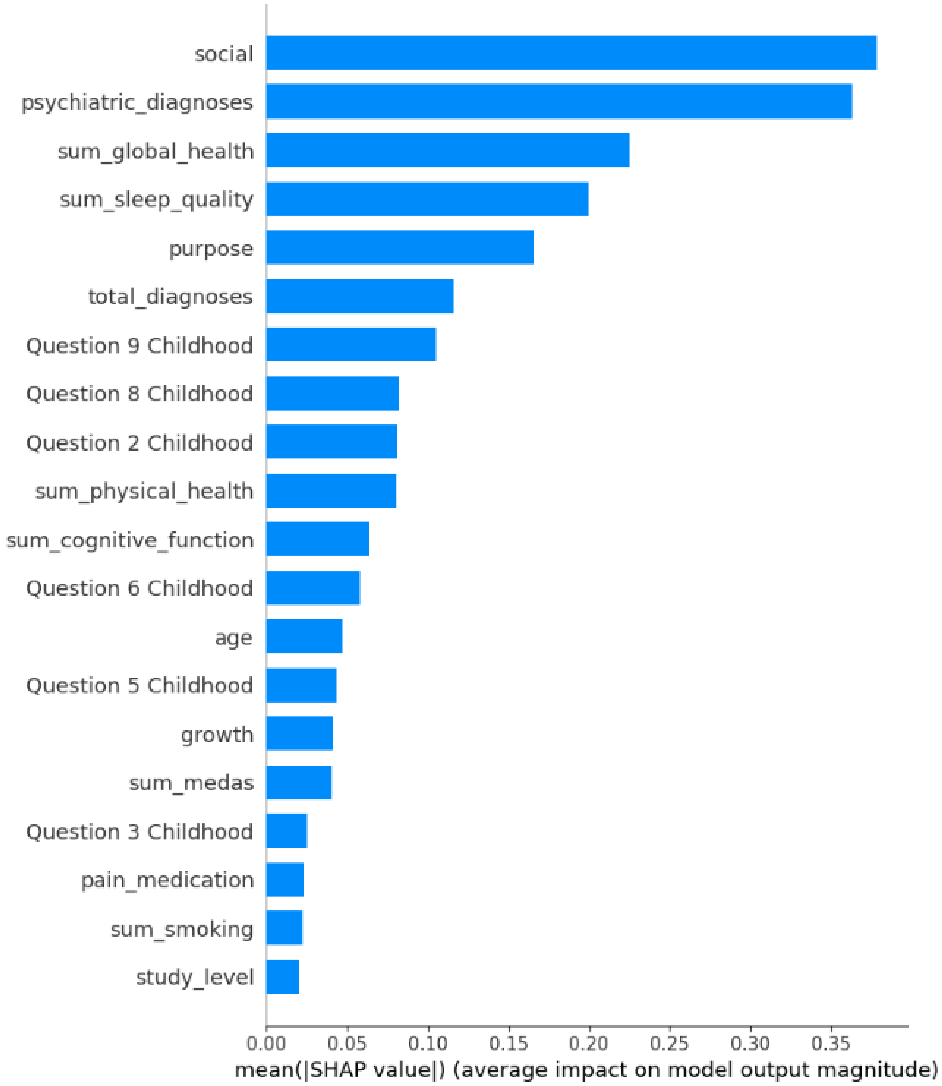

Figure 19. Shapley values in Gradient boosting model trained with women dataset.

The social variable changes a bit its behaviour. This time, an intermediate social engagement decreases depression risk, but a high level and a low one increases it. Living in a big city during the childhood increases the depression diagnoses. Higher values of the question eight about childhood, “The floor / house where you habitually lived during your childhood, how many rooms did you have?”, augment the depression development risk; and the question number two, “Had complications at birth (needed respirator, incubator or remained hospitalized)?” reveals that having complications at birth increase the depression risk (Figure 20).

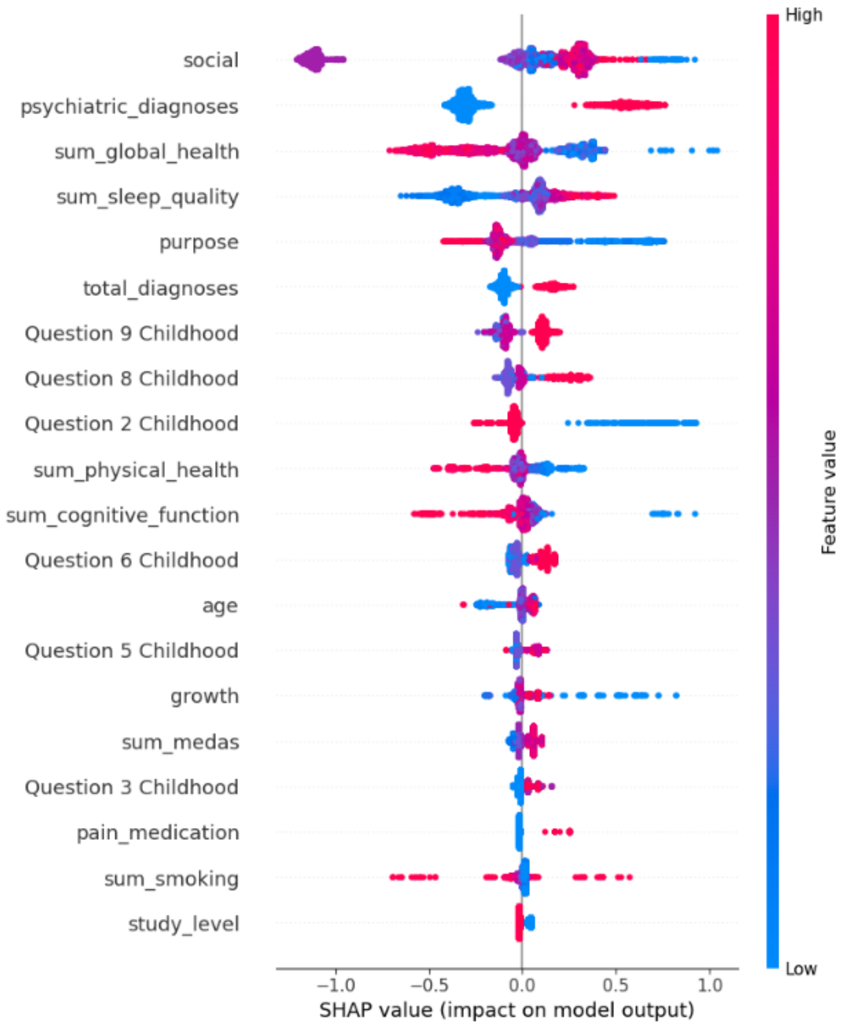

Figure 20. Feature importance vs feature effects in Gradient boosting model trained with women dataset.

5. DISCUSION AND CONCLUSIONS

In this study, whose main objective was to discover new factors affecting the development of depression, several variables have been discovered thanks to the creation of supervised machine learning models.

As has been observed in other articles about mental health, a person's vital plan affects the development of mental illness. This can be seen in the models constructed, with life purpose being between the fifth and seventh most influential variable in the development of depression. Having a good view of the past and the present, having goals and having a sense of directness can reduce the likelihood of developing depression. In the case of personal growth, the second variable included in the life plan, a rather low influence could be observed in the models, being below the tenth position in those trained with the full dataset. Future work remains to be done on the analysis of the rest of the Ryff test items as predictors of depression, for which another database would be needed to collect this information (Ryff, 1995).

Social interaction came as relevant as well. It is known that loneliness can contribute to the development of mental disorders, and it has been shown that according to the models developed in this study, one of these disorders is depression. However, it is not a variable that produces a linear response. The less social interaction, the higher the risk of depression, but it is an intermediate level that produces the opposite effect, so it could be interpreted that maintaining a balance between moments of solitude and moments of social interaction contributes to maintaining mental health.

Having psychiatric disorders is one of the most important risk factors for depression. This makes sense, since one of the illnesses recorded in this variable is anxiety, and it is demonstrated in many articles that there is a comorbidity between anxiety and depression (Lenze, 2001). In fact, comparing the values of anxiety with those of psychiatric diagnoses, we obtain that out of 1238 people with at least one diagnosis of this type, 1191 correspond to diagnoses of anxiety. This would also explain the relevance of the number of total diagnoses in the prediction models, as there are 1581 persons with at least one diagnosis of some illness, which would reveal that most of them are anxiety disorders. None of the diagnoses of cardiovascular and neurodegenerative diseases have obtained a high relevance, so we can conclude that, although they may be determining factors in other mental problems, they do not affect the development of depression.

The last of the variables to be analysed is childhood. When added by individual items, we can see which aspect of childhood is influential. The result of calculating Shapley values has pointed to the question of what job the subject's father had during childhood as important. What one would expect before plotting the graphs is that the better the job, the better the quality of life and therefore the lower the probability of developing depression. However, the summary graphs have shown that if the father had an unpaid job or was a householder during the participant's childhood, the likelihood of depression decreases. One assumption could be that this kind of domestic work or unpaid work involves more time in the family home and increases the level of parent-child interaction, which would be related to the social interaction variable, an important factor in the development of depression. The father's level of education also appeared in several models with some significance. It is directly related to the job he has, so it makes sense that it would have the same behaviour as the job variable. However, these are only assumptions. Additional research would be needed to explain the behaviour of these variables.

The model with the best metrics corresponds to the logistic regression model trained with the resampled dataset by random oversampling with an accuracy of 0.70, a recall of 0.60 and an F1 of 0.24. The random oversampling algorithm does not remove any instances from training and test datasets, so there is a larger number of data available for training, however, the repetition of instances of the minority class, in this case, people with depression, can lead to amplification of outliers, observations that present extreme values for some variables, very different from the mean.

The other generated models obtained an accuracy close to the seventy percent described in other articles about depression prediction models (Kwang-Sig, 2022). However, F1 scores between 0.20 and 0.25 and recall slightly above 0.60 are far from the ideal value of 1 for both metrics. Adding more hyperparameters to the automatic selection of the best ones to generate a model is likely to improve the metrics, but one factor to consider is the run time of the models, which increases exponentially with each option added for testing. However, the three models selected for the analysis of new factors are within the acceptable metric range and meet the objective of predicting depression using test-collected variables.

On the other hand, the gender separation of the dataset shows that, although the accuracy percentage is maintained, the recall and F1 metrics indicate a better fit for the female model. Related to the relevant variables in these models, there are certain behaviours that would require further study. In the logistic regression model, which has the best metrics for the female dataset, high alcohol consumption decreases the likelihood of depression, the opposite of what appears in the other models. Also, another childhood variable related to the number of rooms in the family residence increases in importance, the higher the number of rooms, the higher the risk of developing depression.

The dataset separation by gender reveals that the variables predicting depression have mostly the same risk or prevention behaviour, but they do not have the same predictive significance for each gender. In addition, in each model generated for the same gender, different significance has been shown for all variables. This could be because although the selected models have performed well with the full data set, they may not necessarily be the best fitting models for gender-specific data. On the other hand, the number of participants on the models trained with the men's dataset is quite small, with eighty-two men who have suffered from depression and 1723 men who have not. Moreover, in two of the three models created, the random undersampling algorithm was used, which eliminates instances of the majority class, so the result is two models trained with one hundred and sixty-two data, a very small dataset that may affect the predictions resulting in an unrealistic list of variables relevance. All this together leads to the suggestion of a more comprehensive study with a larger dataset of male participants and a more in-depth search for gender-specific models.

In conclusion, three models have been obtained that predict depression including additional factors different from those of the scientific literature. Among these new predictors, life purpose, social interaction, number of psychiatric diagnoses and father's job during the subject's childhood have been shown to be important in the development of depression.

7.BIBLIOGRAPHY

1. Aleem S. et al. Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions. Electronics. 2022; 11(7):1111. doi: https://doi.org/10.3390/electronics11071111
2. Bakırhan H. et al. Diet, Sleep and Depression: Does Adherence to the Mediterranean Diet Matter? Journal of Turkish Sleep Medicine 2022;9:172-179. doi: 10.4274/jtsm.galenos.2022.15046
3. Bennett S, Thomas AJ. Depression and dementia: cause, consequence or coincidence? Maturitas. 2014 Oct;79(2):184-90. doi: 10.1016/j.maturitas.2014.05.009.
4. Boden, J.M. and Fergusson, D.M. (2011), Alcohol and depression. Addiction, 106: 906-914. https://doi.org/10.1111/j.1360-0443.2010.03351.x
5. Brent D.A. Antidepressants and Suicidality. Psychiatric Clinics of North America, Volume 39, Issue 3, 2016, p. 503-512. doi: https://doi.org/10.1016/j.psc.2016.04.002
6. Cabello-Toscano M. et al. Functional Brain Connectivity Prior to the COVID-19 Outbreak Moderates the Effects of Coping and Perceived Stress on Mental Health Changes: A First Year of COVID-19 Pandemic Follow-up Study. Biol Psychiatry Cogn Neurosci Neuroimaging. 2023 Feb;8(2):200-209. doi: 10.1016/j.bpsc.2022.08.005.
7. Cattaneo G. et al. The Barcelona Brain Health Initiative: A Cohort Study to Define and Promote Determinants of Brain Health. Front Aging Neurosci. 2018 Oct 11;10:321. doi: 10.3389/fnagi.2018.00321
8. Cattaneo G. et al. The Barcelona Brain Health Initiative: Cohort description and first follow-up. PLoS One. 2020 Feb 11;15(2):e0228754. doi: 10.1371/journal.pone.0228754
9. Cattaneo G. et al. Sense of Coherence Mediates the Relationship Between Cognitive Reserve and Cognition in Middle-Aged Adults. Front Psychol. 2022 Mar 28;13:835415. doi: 10.3389/fpsyg.2022.835415
10. Chochinov H.M. Depression in cancer patients. The Lancet Oncology, Volume 2, Issue 8, 2001, p.499-505. doi: https://doi.org/10.1016/S1470-2045(01)00456-9
11. Choi KW, Kim YK, Jeon HJ. Comorbid Anxiety and Depression: Clinical and Conceptual Consideration and Transdiagnostic Treatment. Adv Exp Med Biol. 2020;1191:219-235. doi: 10.1007/978-981-32-9705-0_14
12. Delgado-Gallén S. et al. Brain system segregation and pain catastrophizing in chronic pain progression. Front Neurosci. 2023 Mar 16;17:1148176. doi: 10.3389/fnins.2023.1148176
13. Deo R.C. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920-30. doi: 10.1161/CIRCULATIONAHA.115.001593
14. Furukawa T.A. et al. Optimal dose of selective serotonin reuptake inhibitors, venlafaxine, and mirtazapine in major depression: a systematic review and dose-response meta-analysis. The Lancet Psychiatry, Volume 6, Issue 7, 2019, p.601-609. doi: https://doi.org/10.1016/S2215-0366(19)30217-2

15. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol. 2005 Jun;44(Pt 2):227-39. doi: 10.1348/014466505X29657
16. Jääsaari E.et al. (2019). Efficient Autotuning of Hyperparameters in Approximate Nearest Neighbor Search. Advances in Knowledge Discovery and Data Mining. PAKDD 2019. Lecture Notes in Computer Science, Vol 11440. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-16145-3_46
17. Karaki Y, Ivanov N. Hyperparameters of Multilayer Perceptron with Normal Distributed Weights. Pattern Recognit. Image Anal. 30, 170–173 (2020). doi: https://doi.org/10.1134/S1054661820020054
18. Kwang-Sig L, Byung-Joo H. Machine Learning on Early Diagnosis of Depression. Psychiatry Investigation 2022;19(8):597-605. doi: https://doi.org/10.30773/pi.2022.0075
19. LeMoult J. et al. Meta-analysis: Exposure to Early Life Stress and Risk for Depression in Childhood and Adolescence. J Am Acad Child Adolesc Psychiatry. 2020 Jul;59(7):842-855. doi: 10.1016/j.jaac.2019.10.011
20. Lenze E.J.et al. (2001) Comorbidity of depression and anxiety disorders in later life. Depress. Anxiety, 14: 86-93. doi: https://doi.org/10.1002/da.1050
21. Lundberg S.M. et al. A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems 30 (NIPS 2017), 2017, Vol.30, ISSN: 1049-5258
22. Mahabadi B. et al. Diabetes and Depression. Diabetologe 2021, 17:727–732. doi: https://doi.org/10.1007/s11428-021-00791-3
23. Middelboe T. et al. The relationship between self-reported general health and observed depression and anxiety in cancer-patients during chemotherapy. Nordic Journal of Psychiatry, Volume 49, 1995 - Issue 1. doi: https://doi.org/10.3109/08039489509011880
24. Minarno A.E. et al. Performance Comparisson Activity Recognition using Logistic Regression and Support Vector Machine. 2020 3rd International Conference on Intelligent Autonomous Systems (ICoIAS), Singapore, 2020, pp. 19-24, doi: 10.1109/ICoIAS49312.2020.9081858.
25. Mirowsky, J., & Ross, C. E. (1992). Age and Depression. Journal of Health and Social Behavior, 33(3), 187–205. doi: https://doi.org/10.2307/2137349
26. Mohammed R, Rawashdeh J, Abdullah M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020, 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556.
27. Muñoz-Navarro R. Modelo jerárquico de diagnóstico y derivación de los trastornos mentales comunes en centros de atención primaria. Una propuesta a partir del ensayo clínico PsicAP. Ansiedad y Estrés 23 (2017) 124–129. doi: 10.1016/j.anyes.2017.10.002

28. Paperwalla K.N. et al. Smoking and depression. Medical Clinics of North America, Volume 88, Issue 6, 2004, p.1483-1494. doi: https://doi.org/10.1016/j.mcna.2004.06.007
29. Park LT, Zarate CA Jr. Depression in the Primary Care Setting. N Engl J Med. 2019 Feb 7;380(6):559-568. doi: 10.1056/NEJMcp1712493.
30. Parker G. and Brotchie H. (2010) Gender differences in depression. International Review of Psychiatry, 22:5, 429-436. doi: 10.3109/09540261.2010.492391
31. Pedregosa F. et al. (2011). Scikit-learn: Machine learning in Python. JMLR, 12, 2825-2830.
32. Reichardt M, M. Gundall and H. D. Schotten. Benchmarking the Operation Times of NoSQL and MySQL Databases for Python Clients. IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society, Toronto, ON, Canada, 2021, pp. 1-8. doi: 10.1109/IECON48115.2021.9589382
33. Roughan W.H. et al. (2021) Comorbid Chronic Pain and Depression: Shared Risk Factors and Differential Antidepressant Effectiveness. Front. Psychiatry 12:643609. doi: 10.3389/fpsyt.2021.643609
34. Ryff C. D. (1995). Psychological Well-Being in Adult Life. Current Directions in Psychological Science, 4(4), 99–104. https://doi.org/10.1111/1467-8721.ep10772395
35. Singh P. et al. Multi-objective hyperparameter optimization on gradient-boosting for breast cancer detection. Int J Syst Assur Eng Manag (2023). doi: https://doi.org/10.1007/s13198-023-01955-8
36. Smith M, Alvarez F. A machine learning research template for binary classification problems and shapley values integration. Software Impacts, Volume 8, 2021. doi: https://doi.org/10.1016/j.simpa.2021.100074
37. Steiger A, Pawlowski M. Depression and Sleep. International Journal of Molecular Sciences. 2019; 20(3):607. https://doi.org/10.3390/ijms20030607
38. Sutherland, M.E. Physical activity and depression. Nat Hum Behav 3, 320 (2019). doi: https://doi.org/10.1038/s41562-019-0559-9
39. Talarowska M. et al. (2009). Cognitive functions and depression. Psychiatria polska, 43(1), 31-40. PMID: 19694398
40. World Health Organization. 31 March, 2023. https://www.who.int/es/news-room/fact-sheets/detail/depression
41. Yuan Z. et al. Application of antidepressants in depression: A systematic review and meta-analysis. Journal of Clinical Neuroscience, Volume 80, 2020, p.169-181. doi: https://doi.org/10.1016/j.jocn.2020.08.013
42. Zach, September 8, 2021. What is Considered a “Good” F1 Score? https://www.statology.org/what-is-a-good-f1-score/
43. Zhang Y, Chen Y, Ma L. Depression and cardiovascular disease in elderly: Current understanding. J Clin Neurosci. 2018 Jan;47:1-5. doi: 10.1016/j.jocn.2017.09.022.
44. Zhu N. et al. Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search

Algorithm. Applied Sciences. 2022; 12(20):10456. doi: https://doi.org/10.3390/app122010456