10-3-2024

Toward an explainable public health intelligence to detect depression using mobile application usage

Ehsan Sabzizadeh
IE University, ehsan.sabzizadeh@student.ie.edu

Luz Rello
IE University, luz.rello@ie.edu

Alberto Urueña
Universidad Politécnica de Madrid, alberto.uruena@upm.es

TOWARD AN EXPLAINABLE PUBLIC HEALTH INTELLIGENCE TO DETECT DEPRESSION USING MOBILE APPLICATION USAGE

Research full-length paper

Sabzizadeh, Ehsan, IE University, Madrid, Spain, ehsan.sabzizadeh@student.ie.edu

Rello, Luz, IE University, Madrid, Spain, luz.rello@ie.edu

Urueña, Alberto, Universidad Politécnica de Madrid, Madrid, Spain, alberto.uruena@upm.es

Abstract

The purpose of this research is to design a public mental health intelligence based on explainable machine learning. Historical data on user application usage behavior for four subsequent semesters from the Cybersecurity and Confidence in the Spanish Households National Survey were used, and an IT artifact was developed to detect depression. Historical use of mobile applications can partially predict user depression symptoms, and when we add sociodemographic data (gender, educational level, and age), our model performance reaches acceptable results. Finally, we implement post-hoc explainable algorithms at local and global levels, providing us with a detailed analysis of the variables that derive depression.

Keywords: Public health intelligence, explainable machine learning, design science, depression, e-health.

1 Introduction and Motivation

According to the World Health Organization (WHO), a person dies due to suicide or suicidal behavior every 40 seconds, indicating that over 1.34% of all deaths are related to suicide. In particular, suicide ranks as the fourth leading cause of death for people aged 15 to 29 years. (WHO, 2021). As indicated by WHO (2021), although the suicide mortality rate has decreased during the past years across all regions (except the United States), in the context of the *Sustainable Development Goals* of the United Nations, it is necessary to accelerate the reduction of the suicide mortality rate to meet the global goal of reducing the death rate by one third till 2030. Moreover, if we look at the data from Institute of Health Metrics and Evaluation (2019), we notice that suicide-caused deaths compared to the total number of deaths have increased in countries with a low Socio-Demographic Index (SDI)[1]. Furthermore, from an economic point of view, Rockett et al. (2023) claimed that the total cost of self-injury mortality (SIM) in the United States is approximately 1.19 trillion dollars, considering costs related to the quality of life factor. This is more than twice the estimated cost for the period of 1999/2000, which was around 0.46 trillion dollars (Rockett *et al.*, 2023). One way to decrease suicide attempts and suicide mortality rates is to identify the main reason behind this catastrophic decision.

Depression disorder is the most widespread mental health problem, resulting in a mood of depression or a long period of loss of enjoyment and interest in a particular activites (WHO, 2023). Several research studies in the literature indicate that there is a strong relationship between depression and suicide attempts. For example, Cavanagh et al. (2003) performed a systematic review of the literature (including 154 articles) and realized that mental disorders (mainly caused by depression) were strongly associated with suicide. In another study in the literature, Obuobi-Donkor, Nkire and Agyapong (2021) found that in a large proportion of studies, major depression was identified as the most influential factor in suicide death among older people. In addition to the significant relationship between depression and suicide, depression costs the world economy one trillion dollars a year based on WHO's estimations from 1990 to 2013 (WHO, 2016). Furthermore, between 1990 and 2013, there was an increase of almost 50% in the number of people affected by depression and/or anxiety, from a total of 416 million to a total of 615 million (WHO, 2016). Approximately 10% of the world's population experiences mental disorders, which comprise a significant 30% of all non-fatal diseases (WHO, 2016).

Previous research has found that problematic smartphone use may be a sign of current mental disorders or maybe a potential risk of future mental disorders (Demirci, Akgönül and Akpinar, 2015; Sharma *et al.*, 2019; Panova *et al.*, 2020). Consequently, mobile application usage can contain useful information on individuals' mental health status. Moreover, the number of smartphone users is around 4.6 billion, which means that more than 50% of the world's population uses a smartphone (Statista, 2023). Furthermore, according to another report, people around the world spent five hours on average using their mobile applications (Ceci, 2023).

In this research, our objective is to develop a public mental health intelligence based on Explainable Machine Learning (XML) and mobile application usage, as our IT artifact. This public mental health intelligence not only predicts people prone to depression but also provides more granular information on factors that drive depression for each individual (local explainability). Therefore, our main research question goal is:

Create an IT artifact based on XML to detect depression using mobile application usage.

This piece is organized as follows. First, we will look at the theoretical background of XML. Then, we review related literature on emotion detection and developing design science theory. For the next step, we will articulate our research approach and explain how the design science approach in IS will help us design our IT artifact. Then, we delve into designing our IT artifact based on guidelines from design

[1] SDI is a number between 0 and 1, which combines three kinds of information together: income per capita, educational level, and fertility rates.

science research. To do so, we first develop a machine algorithm to detect depression based on the usage of mobile applications and demographic data. Finally, we present our global (dataset) and local (case) level explainability for our model.

2 Theoretical Background

The term Explainable Artificial Intelligence or XAI emerged formally in research and discussions around the mid-2010s. The US Defense Advanced Research Projects Agency (DARPA) launched a special program on XAI in 2017 (Gunning and Aha, 2019). Gunning and Aha (2019) stated that XAI aims at an end user who depends on an AI system's decisions, suggestions, or actions and thus requires comprehension of the system's reasoning. XML, which is a part of XAI, aims to enhance the interpretability of Machine Learning (ML) models. XML seeks to ensure stakeholders can apprehend, trust, and efficiently employ ML models.

When we are talking about the interpretability of an ML model, we can look at two aspects discussed by Lipton (2016): Transparency (answering "how the model is working?") and post-hoc explanation (which indicates "What else can the model tell me?"). Transparency (opposite to black-box) has three dimensions. The first one is Simulatability, which suggests that the whole model can be understood by the user at a glance. The second aspect of transparency is decomposability, which refers to the property of a model where each component - input, parameter, and calculation - can be understood independently. The third aspect is algorithm transparency, which shows us exactly how the model learns.

As mentioned in Lipton (2016) post-hoc interpretation does not pay attention to how the model is working, instead, it provides users with useful information (for example, which variable has more effect on categorizing a company as fraudulent). Our focus in this research will be on the post-hoc explainability of an ML model. As a result, we will dive deeper into this kind of model interpretability.

Rai (2020) defined two aspects for categorizing post-hoc interpretable methods. The first aspect is model specificity. Explanation techniques can be classified as *model-specific* and *model-agnostic*. *Model-specific* methods integrate interpretability constraints directly into the essential architecture and operational processes of the models (Rai, 2020) and they can be applied to certain types of ML methods (Kim *et al.*, 2023). Conversely, *model-agnostic* methods take advantage of the inputs and outputs of opaque models to produce explanations. Consequently, they can be used for different types of ML models. The second aspect is called the scope of the explanation. This aspect can be divided into two classes: *global for the model* and *local for the prediction*. *Global* explanations seek to clarify patterns across the entire dataset by looking at both the model and the dataset comprehensively (Kim *et al.*, 2023), for example, the coefficients of independent variables in a linear regression analysis or the rules derived from a decision tree. On the contrary, *local* explanations focus on individual instances, examining the complex and frequently non-linear interactions and relationships between various variables linked to a specific data point (Kim *et al.*, 2023).

We can classify the interpretable ML method into four general categories based on the categories mentioned above. As we are interested in taking advantage of heterogeneous kinds of ML methods, our focus will be on model-agnostic types of XMLs:

- **Model-agnostic global explanation:** This XML technique aims to unravel the decision-making processes of convoluted models, including deep learning systems, by employing interpretable models such as decision trees. In addition, diagnostic tools such as partial dependence plots and individual conditional expectations are used to investigate the impact of specific features on the model predictions, which results in comprehensive and in-depth insights into the model output on the data set level. This approach not only clarifies how decisions are made but also helps verify the model's reliability through expert evaluation.
- **Model-agnostic local explanation**: This type of XML technique aims to produce model-agnostic explanations tailored to an individual instance or its immediate surroundings. Local

Interpretable Model-Agnostic Explanation (LIME) (Ribeiro, Singh and Guestrin, 2016), Local Surrogate (LS) (Laugel *et al.*, 2018), RObust Local EXplanations (ROLEX) (Kim *et al.*, 2023) are some of the well-known of this kind of interpretable techniques.

In this article, our primary focus will be on post hoc explainability on both local and global levels. The global explainability shows which parameters are the main predictors of depression for the entire sample. On the other hand, local-level explainability provides more analysis on the case level.

3 Literature Review.

In this section, we only cover studies that used ML methods to predict behavioral characteristics and emotional states (especially depressive mode) based on mobile application usage.

First, Hung et al. (2016) developed a mobile application through which they collected mobile usage data from 28 participants (all graduate students from the same department) over two weeks (as a comparison data set) and five days (to evaluate the model). The collected data contains raw data for application research and user responses about their negative emotions through a visual scale on the app. Then, they used well-known classification methods, including the Super Vector Machine, Naive Bayes with Decision Trees, and Naive Bayes together, which achieved the highest performance with an average accuracy of 86.17%.

Razavi, Gharipour and Gharipour (2020) studied 412 participants who provided data on their use of a cell phone. The study found that participants with depression sent more texts, had fewer contacts saved on their mobile phones, and spent more time on their phones making, receiving, and making shorter phone calls. The best model (random forest classification) resulted in a balanced accuracy of 76.8%, AUC of 75.8%, and sensitivity (recall) of 74.7% (when the participant's gender and age were added to the model, the model's performance increased to a balanced accuracy of 81.1%, AUC of 79.8% and sensitivity of

78.7%).

Alibasa, Calvo and Yacef (2023) gathered data from 72 users over 15 days (around 236,008 data points). The data contain self-reported data (a question about users' feelings (positive and negative modes) that was asked multiple times a day) and mobile application usage data. Then, they used a clustering algorithm to group the mobile usage data into eight categories. In addition, they took advantage of sequential pattern mining tools to extract mobile usage patterns (which they called digital context patterns or DSP) and added them to the model to see if it could increase its prediction power. Finally, they used a random forest classifier that resulted in 77.8% precision.

As we can see, most previous studies on the usage of mobile applications gathered data over a limited range of time for limited users (from 28 to 412). These limitations, especially when applying the ML method, may lead to a higher risk of overfitting and harder generalization.

Moreover, all the mentioned research used black-box ML algorithms, and their main evaluation criterion was accuracy-related measures (accuracy, precision, recall, AUC, etc.). As a result, there is a need for an explainable model which provides users with a more detailed analysis of both the case and the dataset levels.

4 Research Approach

Design science in IS aims to push the boundaries of human and organizational abilities by inventing new IT artifacts (Hevner *et al.*, 2004). Artifacts as an outcome of design science are constructs (the basic language of concepts), models (combined in higher-order constructions), methods (techniques for executing goal-oriented tasks like algorithms and procedures), and implementation (or instantiation of prototypes) (March and Smith, 1995; Hevner *et al.*, 2004). Hevner and Chatterjee, p. (2010, p. 5) defined design science research as "a research paradigm in which a designer answers questions relevant to

human problems through the creation of innovative artifacts, thus contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem."

Design science generally has two steps: building and evaluation (March and Smith, 1995). Building indicates the process of creating artifacts, and evaluation indicates how the designed artifacts are performing toward reaching design goals (March and Smith, 1995). Therefore, as March and Storey (2008) stated, the contribution of design science research requires (1) an accurate description and identification of an IT problem in the organization, (2) demonstrating that there are no adequate solutions in the current IT knowledge base, (3) the creation and demonstration of cutting-edge IT artifacts that resolve the issue, (4) a rigorous evaluation of the IT artifacts concerning their utility, (5) elaboration on the enhancement of the IT knowledge base and practice, and (6) explaining the impact on IT management and application.

Our approach to creating the IT artifacts for this project will be guided by the design science paradigm from two distinct angles, mainly based on the work of March and Storey (2008) and Peffers et al. (2007)Initially, the focus will be developing various IT artifacts, which will be explainable (or interpretable) ML algorithms tailored to our specific design scenario. Our IT artifact has experienced a critical evaluation and refinement phase after construction. This step was crucial in aligning the artifacts with the practical user requirements and expectations before implementation in actual design scenarios.

5 Data

To train our ML model, we used a secondary source of data. These data have been collected through the Cybersecurity and Confidence in Spanish Households National Survey[2] (CCSHNS), conducted by the National Observatory of Telecommunications and Information Society[3]. The CCSSNS is an annual survey of Spanish Internet users on cybersecurity. Data from four time panels were used in this study for the second half of 2020 to the first half of 2022. Each study panel was carried out on representative samples from the 18 to 75 year old population of Internet users.

Since the CCSSNS survey is designed by a Spanish national institution, it aims to collect a representative sample of the country's population. This survey is regularly obtained from two sources: the participants' responses and the information obtained from their phones remotely after consenting to the information. Self-reporting information includes a sociodemographic description and other psychosocial and psychological variables, including personality, mental distress, social well-being, and social pressure of digital media (gathered at the end of each semester). The data gathered from remote devices, e.g., smartphones, provides an account of the security vulnerability, the infection of malware, and, most importantly, application use data during scanning. Our data set contains data from four semesters: the second semester of 2020 and the first semester of 2022, resulting in 15,166 data points. Since we intended to study smartphone usage behavior, we removed all PC users and users with unknown devices, resulting in 7,944 data points. After preprocessing the data and removing outliers for the usage of mobile applications, we had 5,801 cases (user-semester) to use as input to our model.

5.1 Ground Truth

Now, the question is how to determine the state of depression. To assess the level of depression, CCSSNS took advantage of a brief and validated version of 7 elements of the Center for Epidemiological Studies on Depression (CESD), called CESD-7 (Herrero and Gracia, 2007). The CESD-7 scale is an auto-reporting scale that measures depression symptoms in general and is a simplified version of CESD-20.

[2] Estudio sobre la Ciberseguridad y Confianza en los hogares españoles

[3] El Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información (ONTSI)

Radloff (1977) the original version of CES-D was introduced with 20 questions that passed the reliability and validity tests. Furthermore, Lewinsohn et al. (1997) decided to examine the efficacy of CES-D-20 among 1,005 adults (50 years and older) as there is a risk of misinterpreting physical illnesses and their associated functional disabilities as depression symptoms, resulting in an over-diagnosis of depression. The results of their study showed that age, sex, cognitive dysfunction, functional impairment, physical illness, and desire to be socially accepted had no significant detrimental influence on the psychometric characteristics or screening effectiveness of the CES-D-20 survey (they also used a diagnostic interview to determine the level of depression).

Herrero and Gracia (2007) introduced a modified and shorter version of CES-D-20 with seven questions. To verify validity, they implemented a panel study in two groups of Spanish adults to study the relationship between depression and two other variables: health perceptions and social integration. They found that using CES-D-7 instead of CES-D-20 (as a depression measure) would not change the result of regression. Furthermore, Juarros-Basterretxea et al. (2021) studied 203 men in a Spanish prison (19-66 years), of whom 33 (16. 5%) had an official record of depression disorder and showed that CES-D-7 (with a cutoff score of 16 of 21) recognizes almost all cases with major depression and syndrome.

Total number of cases	Mean of depression score	SD of depression score	Depression cases
5,801	6.02	4.60	3.9%

Table 1 Depression statistics

Application	Mean	Median	SD	75% percentile	Max
Total usage	201.68	139.27	215.33	275.39	1435.11
Social media application usage	28.95	7.61	53.89	35.27	727.40
Entertainment application usage	14.09	2.86	36.91	11.74	840.74
Communication application usage	10.49	2.22	25.59	10.98	679.69
Shopping application usage	4.18	0.69	13.08	3.36	419.81
Messaging application usage	25.25	12.22	40.20	32.19	567.65

Table 2 Statistical summary of application usage (average minutes per day)

As a result, CES-D-7 can be considered an appropriate measure to detect depression and consists of seven items that measure how often each depressive symptom occurred over the past week. The response of the categories ranges from 0 (never or rarely, less than one day) to 3 (most of the time or all the time, 5–7 days). So we calculate the sum of the answers (resulting in a number between 0-21), then take the average, and finally turn it into a variable 0-1 by considering 16 as the optimal cut-off point based on Juarros-Basterretxea et al. (2021). Consequently, according to the summary of the statistics presented in

Table 1, 3.9% of the semester users in our sample are prone to depression, which is in line with the WHO report WHO (2023).

5.2 Features

The main variables that served as features in our ML method address users' usage of mobile applications (total usage plus the usage of 16 other apps). The data was collected for each user on different days, so we divided each application usage by the number of days and calculated the average usage per day for each month. Then, we categorize the app into six main groups:

- **Total application usage**
- **Social Media**: Twitter, Facebook and Instagram.
- **Entertainment:** YouTube, Netflix, and Spotify.
- **Communication**: Skype, Zoom, and Call.
- **Shopping**: Amazon and Wallapop.
- **Messaging**: Telegram, WhatsApp and Mail.

In Table 2, we show the statistical summary of the application usage. Another source of data that we can use in our model is the sociodemographic data collected through the survey during each semester. We have a total number of 5,801 cases in our data set, of which 2,999 were men, and 2,802 were women, 1,575 were unemployed, 4,226 were employed, 1,676 were not married (or divorced or widow), and 4,125 were married. Regarding the educational level, we had 2,953 cases with higher education, 2,785 with secondary school education, and 63 with primary education (and others).

5.3 Data pre-processing

Before analyzing the data, we should pre-process it. The app usage data was reported monthly in the data set, but the depression score was measured at the end of each semester. As a result, we take the mean of application usage for a whole semester and calculate the average application usage per day for each user for each semester. As a result, we have data points for each user over four time periods (semesters).

To remove outliers, we filtered the data for each application usage and considered only cases with total application usage in the range of 0 to 24 hours.

Furthermore, since we have two heterogeneous data sources and a wide range of application usage, we normalize the data using the max-min normalization method:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Moreover, our data was highly imbalanced (3.9% depressed case). To deal with this problem, we used an oversampling method developed by Chawla et al. (2002) called Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples for the minority class by interpolating existing examples from the minority class. For each sample in the minority class, SMOTE identifies its k-nearest neighbors and generates new synthetic samples by randomly selecting one of these neighbors and combining the feature values of the chosen neighbor with the original sample. This approach aids in balancing the class distribution, enhances the model's training, and mitigates the risk of overfitting that might arise from merely reproducing minority class instances.

6 Designing the IT Artifact

In this section, we want to design our preliminary IT artifact, an ML algorithm, together with a post-hoc explanation. This algorithm will be the backbone of our mental health application.

6.1 Model selection and optimization

First, based on the literature and experience, we selected eight ML methods that we found suitable for our research goal: Super Vector Classifier, Logistic Regression, Complement Naive Bayes, XGBoost Classifier, Balanced Random Forest Classifier, Gradient Boosting, Adaptive Boosting Classifier, Balanced Bagging Classifier, and Stacking. Each of these methods requires proper tuning of its hyperparameters.

To identify the optimal classification method for our dataset, we need to choose an evaluation metric to compare our ML method's performance. Here, we selected the F1 score as our main evaluation benchmark.

We followed Algorithm 1 to create the set of optimized hyperparameters for each of the ML models. After this step, we evaluated each model based on our test data set and selected the high-performance ML model based on the F1 score.

For the one-level stacking, we used the four best models (considering the f1 score) as the base model and the best model as the final classifier. For the two-level stacking classifier, we do the same for the base learners, select two of the best-performing models as the first level of stacking, and use the best-performing model as the final layer of classification.

Algorithm 1 Model optimization

Considering train features ($features\_tr$), train target ($target\_tr$), train features ($features\_te$), train target ($target\_te$), and list with length of L (L=8) containing model names ($model\_list = [m_1, ..., m_L]$) and hyperparameters list $param\_list = [p_1, ..., p_L]$ (we assume each p_l have length of P):

for l in $1, .., L$ **do**
  $best\_model_l \leftarrow 0$
  $best\_f1\_score_l \leftarrow 0$
  for p in $1, ..., P$ **do**
    Holding all the other parameters constant
    Split train features ($features\_tr$) and train target ($target\_tr$) into K separate data set
    for k in $1, .., K$ **do**
      Set $features\_f_k$ *and* $target\_t_k$ *as test instances*
      Train $method_l$ *remaining* $K - 1$ *splits*
      Calculate $target\_pr_k$ *using trained model* $method_l$ *and* $features\_f_k$
      Calculate $f1\_score_k$ *using* $target\_pr_k$ *and* $features\_tr_k$
    end for
    $f1\_score_p = \frac{1}{K} \sum_{k=1}^{K} f1\_score$
    if $f1\_score_p > best\_f1\_score_l$ **then**
      $best\_model_l \leftarrow f1\_score_p$
      $best\_f1\_score_l \leftarrow method_l$
    end if
  end for
  Add $best\_model_l$ and $best\_f1\_score_l$ to $best\_models\_list$ and $best\_f1\_score\_list$
end for
Return $best\_models\_list$ and $best\_f1\_score\_list$

Figure 1 Optimization Algorithm for ML Methods

6.2 XML Algorithm

After developing the first version of our ML algorithm based on CCSHNS, we tried to develop our XML algorithm, which will be based on post hoc analysis.

6.2.1 Global model agnostic explainability

This section aims to offer global-level explainability, enhancing our understanding at the dataset level. Here, we use the SHAP values (SHapley additive explanation) introduced by Lundberg and Lee (2017). SHAP values distribute the contribution of each feature to the final prediction according to principles from cooperative game theory, particularly the Shapley value concept. In a model, each feature is considered a "player" in a cooperative game, and SHAP values are calculated to fairly distribute the "payout" (i.e., the prediction) among all features by evaluating their marginal contributions across all possible

subsets of features. This technique ensures a reliable and theoretically sound distribution of feature significance, making it suitable for various model types and enhancing interpretability. The global explainable part of our IT artifact will be based on the SHAP values approach.

6.2.2 Local model agnostic explainability

The global level of explainability will provides us with some insights into the relationship between depression and application usage at the dataset level. However, in most cases, we need to look at each case with more details (local-level explainability) to decide if a person is prone to depression or not. LIME (Local Interpretable Model-agnostic Explanations) is a technique designed to explain individual predictions of any ML model by approximating it locally with an interpretable model (Ribeiro, Singh and Guestrin, 2016). LIME works by perturbing the input data around the prediction instance and observing the resulting changes in predictions. The perturbations generate a new dataset on which a more straightforward and more interpretable model (like a linear model) is trained to replicate the behavior of the complex model near the specific instance. This method enables users to discern which features will likely affect the model's predictions for that particular instance. The local explainable part of our IT artifact will be based on the LIME algorithm.

7 Results

7.1 Our ML Algorithm Results

After turning the ML algorithms, we run the models, and you can see the results in Table 4, which show that SVC outperforms other methods. As a result, we will select this model as the input for our XML algorithms.

Method	AUC	F1 Score	Depressed		Not Depressed	
			Precision	Recall	Precision	Recall
SVC	**0.64**	**0.59**	0.28	0.55	**0.90**	0.73
Complement Naive Bayes	0.61	0.56	0.24	0.54	0.89	0.69
Logistic regression	0.63	0.55	0.25	**0.60**	**0.90**	0.66
XGBoost	0.52	0.53	0.23	0.13	0.85	0.92
Gradient Boosting	0.49	0.48	0.13	0.07	0.84	0.91
Adaptive Boosting	0.53	0.54	**0.31**	0.12	0.85	**0.95**
Balanced Random Forest	0.54	0.48	0.18	0.49	0.86	0.58
Balanced Bagging	0.56	0.56	0.27	0.27	0.86	0.86
One-level Stacking	0.63	0.57	0.26	0.58	0.90	0.69
Two-level Stacking	0.61	0.54	0.23	0.58	0.89	0.64

Table 3 Results from optimized classifiers with mobile usage and socio-demographic data

7.2 Global XML Algorithm Results

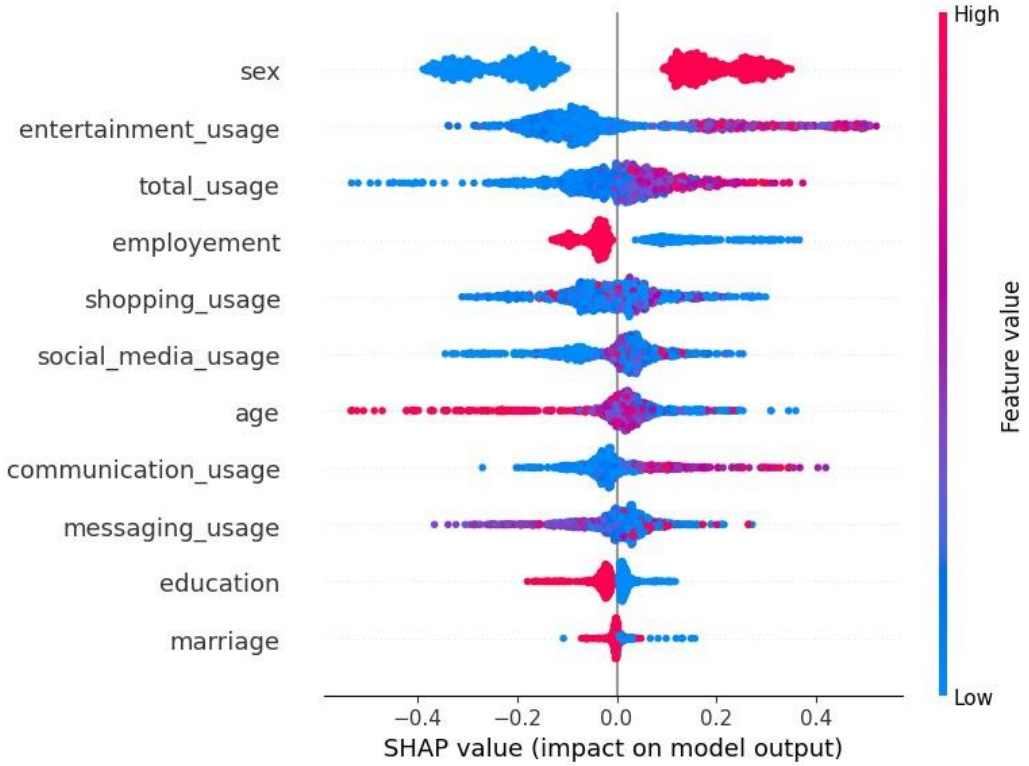

Figure 2 SHAP values on the dataset level

Figure 2 shows us the SHAP summary plot. SHAP values help understand the impact of each feature on the model's predictions. Here is how to interpret the plot:

- **X-Axis (SHAP value):** The SHAP value indicates the impact of each feature on the model's prediction. A positive SHAP value means that the feature contributes to an increase in the predicted value (depressed case). In contrast, a negative SHAP value implies that the feature causes a decrease in the predicted value (non-depressed case).
- **Y-Axis (Features):** Each row represents a different feature.
- **Color Bar:** The color indicates the value of the feature (blue for low values and pink for high values).
- **Sex**: High values (pink: female) have a positive impact on the model output, while low values (blue: male) have a negative impact. This indicates a higher probability of depression for women compared to men (in line with WHO (2023)).
- **Education**: Higher education level (pink) has a negative impact, while lower education level (blue) has a neutral to slightly negative impact.
- **Employment**: High values (pink: employed) have a negative impact, and low values (blue: unemployed) have a positive impact, indicating employment status is a significant feature in the model.
- **Age**: Younger ages (pink) have a negative impact, while older ages (blue) have a positive impact on the model output, indicating that age is an influential factor.
- **Entertainment** Usage: Higher values of entertainment usage generally have a positive impact on the model output, while lower values have a neutral to negative impact.

- **Social Media Usage**: High values show a mix of positive and negative impacts, indicating that social media use has a complex relationship with the target variable.
- **Total Usage**: Similarly to social media usage, total usage values show both positive and negative impacts, suggesting nuanced effects on prediction.
- **Communication Usage**: High values (pink) generally have a positive impact, while low values (blue) generally have a negative impact.
- **Marriage**: This feature has a less distinct pattern, but shows some variation, where high values (pink: married) slightly decrease the model output, while low values (blue: not married, divorced, or widow) have a positive impact.
- **Shopping Usage**: Similarly to social media usage, shopping application usage values show both positive and negative impacts, suggesting nuanced effects on prediction.
- **Messaging Usage**: Similarly to social media usage and total application usage, messaging usage values show both positive and negative impacts, suggesting nuanced effects on prediction.

7.3 Local XML Algorithm Results

Figure 3 (right) illustrates the model's predicted probabilities for a case of depression. The LIME visualization for this instance shows a 70% probability of depression and a 30% probability of not being depressed. Key elements that contribute toward the depressed prediction are as follows: higher usage of entertainment apps (over 11.25) contributing 0.09, female gender contributing 0.09, being unemployed contributing 0.05, extensive total usage (over 274.64) contributing 0.03, younger age (36 years or below) contributing 0.02, increased social media usage (above 7.47) contributing 0.01, lower educational levels contributing 0.01, and several other minor factors.

Alternatively, factors that result in a non-depressed prediction include increased shopping application usage (shopping usage > 0.64), contributing 0.02, and higher communication application usage, contributing 0.01. The model predicts with high confidence that the individual is depressed, primarily due to significant use of entertainment and social media applications, lack of employment, and being younger.

Figure 3 (left) illustrates the model's prediction probabilities for an individual who is not depressed. The model estimates a 54% chance of being non-depressed and a 46% chance of being depressed. For the non-depressed prediction, essential factors include gender (male) contributing 0.09, employment status (employed) contributing 0.04, increased messaging usage contributing 0.02, higher overall usage contributing 0.01, and marital status (married) contributing 0.01.

Elements that lead to a depressed prediction encompass elevated entertainment usage (over 11.25) contributing 0.09, heightened communication usage contributing 0.04, increased social media usage contributing 0.02, lower educational level contributing 0.01, greater shopping usage contributing 0.01, and being within the age bracket of 36 to 44 contributing 0.01. The model's balanced forecasts reveal a nuanced interaction among these elements, resulting in a slightly increased chance of not experiencing depression.

Figure 3 LIME analysis for a depressed case (right) and for a non-depressed case (left)

8 Discussion

These research findings resulted in interesting insights into building an XML public health intelligence that can detect depression based on analyzing mobile application usage by following a design science research approach. Our research contributes to the field of IS by demonstrating how mobile application usage data and sociodemographic information can forecast depression and how each feature will contribute to our model output on both case-level and dataset-level.

8.1 Model Performance and Explainability

The performance of our ML models (Table 4), specifically the SVC, indicates acceptable predictive power. The utilization of SHAP values for overall comprehensibility and LIME for specific comprehensibility yielded distinct insights into the aspects that significantly influence depression predictions. By looking at results from SHAP values, we can see that features like sex, age, employment, and communication usage exhibit distinct patterns, whereas features such as total application usage, social media usage, and shopping application usage present more intricate effects and cannot be interpreted easily on dataset level. As a result, we need a more granular analysis on the case level, which leads to implementing the LIME algorithm. For the depressed case (Figure 3, right), our analysis indicates that increased engagement with entertainment and social media apps, lack of employment, extensive overall usage, being younger, and having lower levels of education are associated with a higher likelihood of depression. On the other side, it has been shown that an increase in shopping and communication application usage is linked to a reduced probability of depression. For the non-depressed case (Figure 3, left), our analysis suggests that being male, employed, frequently using messaging apps, having high overall usage, and being married are associated with a lower likelihood of depression. In contrast, high involvement in entertainment, communication, and social media, coupled with lower educational attainment, increased shopping behavior, and belonging to a particular age range, are linked to a higher probability of depression.

8.2 Implications for Public Health Intelligence

The results of our research emphasize the possibility of using mobile application usage data to monitor and forecast the probability of depression. This method can result in timely identification and intervention of depression, which are essential for enhancing mental health at the individual and collective levels. By integrating XML techniques, the output of our ML models becomes more understandable, resulting in a comprehensive understanding of factors that lead to depression. Moreover, explainability is essential for gaining the trust of both users and mental health professionals.

8.3 Challenges and Limitations

Although the results of our study are encouraging, there are some limitations. A major challenge we face is the imbalanced distribution of data sets, with only 3.9 percent of patients classified as depressed. We have used SMOTE methods to address this problem, but this method is not perfect and future research should explore alternative ways to address imbalanced data sets to improve model performance further.

In addition, here, we used self-reported data as our ground truth, which may introduce bias to our results. One way to overcome this issue is to use other sources of data, such as physiological and behavioral data or a mental health expert's opinion, which can help us validate users' depression states.

9 Conclusion

Our study shows the advantages of using XML to detect depression using actual data from mobile application usage. This approach improves the predictive powers of mental health monitoring systems and guarantees interpretability and reliability, which are crucial for the acceptance of ML methods in public health settings.

10 Acknowledgements

The work presented here has been developed in the context of the Digymatex project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 870578. The results presented here reflects only the authors’ view, and the European Union is not responsible for any use that may be made of the information it contains

References

Alibasa, M.J., Calvo, R.A. and Yacef, K. (2023) 'Predicting Mood from Digital Footprints Using Frequent Sequential Context Patterns Features', *International Journal of Human–Computer Interaction*, 39(10), pp. 2061–2075. Available at: https://doi.org/10.1080/10447318.2022.2073321.

Cavanagh, J.T.O. *et al.* (2003) 'Psychological autopsy studies of suicide: a systematic review', *Psychological Medicine*. 2003/04/10, 33(3), pp. 395–405. Available at: https://doi.org/DOI: 10.1017/S0033291702006943.

Ceci, L. (2023) *Number of hours spent per day using apps worldwide from 2019 to 2022, by country*. Available at: https://www.statista.com/statistics/1269704/time-spent-mobile-apps-worldwide/ (Accessed: 29 August 2023).

Chawla, N. V. *et al.* (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal Of Artificial Intelligence Research*, 16, pp. 321–357. Available at: https://doi.org/10.1613/jair.953.

Demirci, K., Akgönül, M. and Akpinar, A. (2015) 'Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students', *Journal of Behavioral Addictions*, 4(2), pp. 85–92. Available at: https://doi.org/https://doi.org/10.1556/2006.4.2015.010.

Gunning, D. and Aha, D.W. (2019) 'DARPA's Explainable Artificial Intelligence (XAI) Program', *AI Magazine*, 40(2), pp. 44–58. Available at: https://doi.org/10.1609/AIMAG.V40I2.2850.

Herrero, J. and Gracia, E. (2007) 'Una medida breve de la sintomatología depresiva (CESD-7)', *Salud mental*, 30(5), pp. 40–46.

Hevner, A. *et al.* (2004) 'Design Science in Information Systems Research', *MIS Quarterly*, 28(1), pp. 75–105. Available at: https://doi.org/10.2307/25148625.

Hevner, A. and Chatterjee, S. (2010) *Design research in information systems. Theory and practice*. Springer.

Hung, G.C.-L. *et al.* (2016) 'Predicting Negative Emotions Based on Mobile Phone Usage Patterns: An Exploratory Study', *JMIR Res Protoc 2016;5(3):e160 https://www.researchprotocols.org/2016/3/e160*, 5(3), p. e5551. Available at: https://doi.org/10.2196/RESPROT.5551.

Institute of Health Metrics and Evaluation (2019) *Global Health Data Exchange (GHDx)*. . Available at: https://vizhub.healthdata.org/gbd-results/ (Accessed: 4 March 2023).

Juarros-Basterretxea, J. *et al.* (2021) 'Using the CES-D-7 as a Screening Instrument to Detect Major Depression among the Inmate Population', *International Journal of Environmental Research and Public Health*, 18(3), pp. 1–10. Available at: https://doi.org/10.3390/IJERPH18031361.

Kim, B. *et al.* (2023) 'ROLEX: A Novel Method for Interpretable Machine Learning Using Robust Local Explanations', *MIS Quarterly*, 47(3), pp. 1303–1332. Available at: https://doi.org/10.25300/MISQ/2022/17141.

Laugel, T. *et al.* (2018) 'Defining Locality for Surrogates in Post-hoc Interpretablity'. Available at: https://arxiv.org/abs/1806.07498v1 (Accessed: 24 April 2024).

Lewinsohn, P.M. *et al.* (1997) 'Center for epidemiologic studies depression scale (CES-D) as a screening instrument for depression among community-residing older adults', *Psychology and Aging*, 12(2), pp. 277–287. Available at: https://doi.org/10.1037/0882-7974.12.2.277.

Lipton, Z.C. (2016) 'The Mythos of Model Interpretability', *Communications of the ACM*, 61(10), pp. 35–43. Available at: https://doi.org/10.1145/3233231.

Lundberg, S.M. and Lee, S.I. (2017) 'A Unified Approach to Interpreting Model Predictions', *Advances in Neural Information Processing Systems*, 2017-December, pp. 4766–4775. Available at: https://arxiv.org/abs/1705.07874v2 (Accessed: 2 December 2023).

March, S.T. and Smith, G.F. (1995) 'Design and natural science research on information technology', *Decision Support Systems*, 15(4), pp. 251–266. Available at: https://doi.org/10.1016/0167-9236(94)00041-2.

March, S.T. and Storey, V.C. (2008) 'Design science in the information systems discipline: An introduction to the special issue on design science research', *MIS Quarterly: Management Information Systems*, 32(4), pp. 725–730. Available at: https://doi.org/10.2307/25148869.

Obuobi-Donkor, G., Nkire, N. and Agyapong, V.I.O. (2021) 'Prevalence of Major Depressive Disorder and Correlates of Thoughts of Death, Suicidal Behaviour, and Death by Suicide in the Geriatric Population—A General Review of Literature', *Behavioral Sciences*, 11(11). Available at: https://doi.org/10.3390/bs11110142.

Panova, T. *et al.* (2020) 'Specific smartphone uses and how they relate to anxiety and depression in university students: a cross-cultural perspective', *Behaviour & Information Technology*, 39(9), pp. 944–956. Available at: https://doi.org/10.1080/0144929X.2019.1633405.

Peffers, K. *et al.* (2007) 'A Design Science Research Methodology for Information Systems Research', *Journal of Management Information Systems*, 24(3), pp. 45–77. Available at: https://doi.org/10.2753/MIS0742-1222240302.

Radloff, L.S. (1977) 'The CES-D scale: A self-report depression scale for research in the general population', *Applied psychological measurement*, 1(3), pp. 385–401.

Rai, A. (2020) 'Explainable AI: from black box to glass box', *Journal of the Academy of Marketing Science*, 48(1), pp. 137–141. Available at: https://doi.org/10.1007/S11747-019-00710-5/TABLES/1.

Razavi, R., Gharipour, A. and Gharipour, M. (2020) 'Depression screening using mobile phone usage metadata: a machine learning approach', *Journal of the American Medical Informatics Association*, 27(4), pp. 522–530. Available at: https://doi.org/10.1093/jamia/ocz221.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 'Model-Agnostic Interpretability of Machine Learning'. Available at: https://arxiv.org/abs/1606.05386v1 (Accessed: 27 April 2024).

Rockett, I.R.H. *et al.* (2023) 'Escalating costs of self-injury mortality in the 21st century United States: an interstate observational study', *BMC Public Health*, 23(1), p. 285. Available at: https://doi.org/10.1186/s12889-023-15188-8.

Sharma, M. *et al.* (2019) 'Nomophobia and its relationship with depression, anxiety, and quality of life in adolescents', *Industrial Psychiatry Journal*, 28(2). Available at: https://journals.lww.com/inpj/Fulltext/2019/28020/Nomophobia_and_its_relationship_with_depression,.11.aspx.

Statista (2023) *Number of smartphone users worldwide from 2013 to 2028*. Available at: https://www.statista.com/forecasts/1143723/smartphone-users-in-the-world (Accessed: 14 August 2023).

WHO (2016) 'Investing in treatment for depression and anxiety leads to fourfold return'. Available at: https://www.who.int/en/news-room/detail/13-04-2016-investing-in-treatment-for-depression-and-anxiety-leads-to-fourfold-return (Accessed: 13 May 2016).

WHO (2021) *Suicide worldwide in 2019: global health estimates*. Available at: https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data#:~:text=More%20than%20700%20000%20people,and%20occurs%20throughout%20the%20lifespan. (Accessed: 16 June 2021).

WHO (2023) *Depressive disorder (depression), WHO*. Available at: https://www.who.int/news-room/fact-sheets/detail/depression (Accessed: 29 June 2023).