salud mental

ORIGINAL ARTICLE
Volume 45, Issue 5, September-October 2022
doi: 10.17711/SM.0185-3325.2022.028

Machine learning approaches to identify profiles and predictors of psychosocial discomfort among Italian college students

Chiara Buizza,[1] Jessica Dagani,[1] Clarissa Ferrari,[2] Herald Cela,[1] Alberto Ghilardi[1]

[1] Department of Clinical and Experimental Sciences, University of Brescia, Italy.

[2] Service of Statistics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.

Correspondence:
Chiara Buizza
Department of Clinical and Experimental Sciences, University of Brescia.
Viale Europa 11,
Brescia, Italy.
Phone: +39 030 371-7149
Email: chiara.buizza@unibs.it

Received: 6 August 2021
Accepted: 4 February 2022

Citation:
Buizza, C., Dagani, J., Ferrari, C., Cela, H., & Ghilardi, A. (2022). Machine learning approaches to identify profiles and predictors of psychosocial discomfort among Italian college students. *Salud Mental*, *45*(5), 213-226.

DOI: 10.17711/SM.0185-3325.2022.028

ABSTRACT

Introduction. College students live a crucial period of transition from late adolescence to adulthood when they have to deal with important stressful tasks. Thus, university often represents a stressful environment, pushing students to cope with a high academic pressure. As a result, this period constitutes a sensitive age for the onset of mental disorders. Typically, students are not aware of the early signs of their own compromised mental health until symptoms aggravate to an overt disorder. Therefore, it is important to timely detect subthreshold symptoms mostly related to generic mental distress. **Objective.** First, to assess psychophysical well-being and mental distress among college students in northern Italy, and to detect predictors, among socio-demographic and academic characteristics, and risky drug use of these two outcomes. **Method.** The study involved 13,886 students who received an email explaining the purpose of the e-research. The questionnaires used were the General Health Questionnaire (GHQ-12), the University Stress Scale (USS), and a modified version of World Health Organization-ASSIST v3.0. **Results.** 3,754 students completed the web-survey. Students showed poor well-being and mental distress. The strongest predictor of mental distress and compromised well-being was physical health, followed by sex, study field, risky drug use, and academic performance concerns. **Discussion and conclusion.** This study shows that it is very important to promote in college students healthy behaviors in order to increase their physical exercise and reduce substance use. Moreover, it would be desirable to improve academic counselling facilities as an important front-line service to intercept mental health issues among young adults.

Keywords: Well-being, mental distress, college students, predictors, web-survey.

RESUMEN

Introducción. Los estudiantes universitarios pasan por un periodo crucial en su transición de la adolescencia tardía a la edad adulta, periodo en que tienen que lidiar con tareas estresantes. La universidad representa un entorno estresante, que empuja a los estudiantes a hacer frente a una alta presión académica. Como resultado, este periodo constituye una edad sensible para la aparición de trastornos mentales. En general, los estudiantes no cobran consciencia de los primeros signos de que su propia salud mental está en riesgo sino hasta que los síntomas se agravan y se convierten en un trastorno manifiesto. Por tanto, es importante detectar oportunamente los síntomas subumbrales relacionados ante todo con la angustia mental genérica. **Objetivo.** Evaluar el bienestar psicofísico y la angustia mental entre estudiantes universitarios del norte de Italia, y en segundo lugar, detectar predictores entre las características sociodemográficas y académicas, y el uso de drogas de estos dos resultados. **Método.** En el estudio participaron 13,886 estudiantes que recibieron un correo electrónico que explicaba el propósito de la investigación. Los instrumentos utilizados fueron el Cuestionario de Salud General (GHQ-12), la Escala de Estrés Universitario (USS) y una versión modificada de la Organización Mundial de la Salud-ASSIST v3.0. **Resultados.** 3,754 estudiantes completaron la encuesta en línea. Los estudiantes mostraron bienestar y angustia mental. El predictor más fuerte de angustia mental y bienestar comprometido fue la salud física, seguido del sexo, el campo de estudio, el uso de drogas y el rendimiento académico. **Discusión y conclusión.** Este estudio muestra que es muy importante promover entre los estudiantes universitarios comportamientos saludables para promover el ejercicio físico y reducir el consumo de sustancias. Además, sería deseable mejorar la orientación académica que es un importante servicio de primera línea para interceptar los problemas de salud mental en los estudiantes.

Palabras clave: Bienestar, malestar mental, estudiantes universitarios, predictores, encuesta web.

INTRODUCTION

College students go through a crucial period of transition from late-adolescence to adulthood, when they have to deal with various stressful tasks. This period of development is considered a sensitive age for the onset of common psychiatric conditions (Kang, Rhodes, Rivers, Thornton, & Rodney, 2021; Kessler et al., 2005; Liu, Stevens, Wong, Yasui, & Chen, 2018; Patel, Flisher, Hetrick, & McGorry, 2007). The WHO World Mental Health International College Student project displayed that one in three college students screened positive for at least one of the common lifetime mental disorders, showing a high level of need for mental health services in university contexts (Auerbach et al., 2018). Furthermore, several evidences found that college students present poorer mental health as compared to their peers in the general population (Blanco et al., 2008; Bruffaerts et al., 2019; Kang et al., 2021; Lovell, Nash, Sharman, & Lane, 2015; Rith-Najarian, Boustani, & Chorpita, 2019). Also, alcohol/substance use disorders are highly prevalent among college students (Bruffaerts et al., 2019; Cordero-Oropeza, García-Méndez, Cordero-Oropeza, Corona-Maldonado, 2021; Skidmore, Kaufman, & Crowell, 2016), which eventually may lead to suicide or attempted suicide, an ever-growing issue of concern in academic settings (Gunnel, Caul, Appleby, John, & Hawton, 2020; Mortier et al., 2018; Oh, Marinovich, Jay, Zhou, & Kim, 2021; WHO, 2016).

Despite early mental disorders onset, effective treatment is typically not initiated until years later. Although often universities provide integrated support services, a major cause of concern is that students consistently show low levels of help-seeking behaviours (Eisenberg, Hunt, & Speer, 2012; McLafferty et al., 2017), due to a lack of knowledge about mental health problems, stigma, or denial of the severity of their problems. Typically, students are not aware of early signs of their own compromised mental health until symptoms aggravate to an overt disorder. Therefore, it is important to promptly detect those subthreshold symptoms mostly to generic mental distress and well-being.

This study aims to assess psychophysical well-being and mental distress among college students, and to identify possible predictors associated with them. To achieve this aim, we applied different machine learning approaches in order to detect predictors at different levels of analysis and give major reliability and scientific grounding to the findings. With these approaches, comprehensive profiles of students well-being and distress were built.

METHOD

Study design

The present cross-sectional study took place in Brescia University, a medium-sized public college in Northern Italy, from May to June 2019. The target sample was composed of 13,886 students. All of them received a first email asking them to participate in a web-survey on their psychological well-being, together with the link to access the survey and a detailed description of the study. They were informed that participation was voluntary and that the survey was completely anonymous. Through the web-link, students were asked to confirm their informed consent to participate. The web-survey was created with LimeSurvey (www.limesurvey.org), a proprietary survey tool that allows for completely anonymous data collection. Indeed, the software automatically sent via email a personal link to access the survey each participant. Once a participant completed the survey, LimeSurvey deleted any link between the participant and their answers to the survey. Only de-identified data were delivered to the investigators to preserve participants' anonymity. The survey was implemented following the guidelines proposed by Pealer and Weiler (2003). In order to maximize response rate, we used some of the strategies proposed by Edwards et al. (2009) such as using user-friendly questions, choosing close-ended options for answers, and sending reminders. Indeed, every week (for a total of six weeks) an email was sent by the software to the students who did not complete the survey to remind them to participate.

Measurements

For the purposes of this paper the following questionnaires were considered:

The General Health Questionnaire (GHQ-12). It is composed by 12 items to assess psychophysical well-being (Goldberg & Blackwell, 1970). Each item scores from 0 to 3. The standard method 0-0-1-1 of scoring was used in this study. In this method, a score of 0 was assigned to the first two low-stress alternatives and a score of 1 was given to the two high-stress alternatives. The maximum score was 12, with a cut-off point > 3, indicating psychological distress. The GHQ-12 proved to be a reliable instrument, as indicated by a Cronbachs' alpha of .81 (Politi, Piccinelli, & Wilkinson, 1994).

The University Stress Scale (USS). It is composed by 21 items that capture the cognitive appraisal of demands across the range of environmental stressors experienced by students (Stallman & Hurst, 2016). Students are asked to rate on a 4-point Likert scale, ranging from 0 (Not at all) to 3 (Constantly). The total score ranged from 0 to 63: the higher the score, the higher the perceived stress level. Extent score ≥ 13 is predictive of significant mental distress. The USS proved to be a reliable instrument as indicated by a Cronbach's alpha of .83, test-retest reliability r = .82 (Stallman, 2008).

A modified version of the World Health Organization-ASSIST v3.0: a questionnaire based on the self-report adaptation of Barreto, Oliveira, and Boerngen-Lacerda

(2014) to detect harmful and hazardous drug use in primary health care, general medical care and other settings (Poznyak, 2008). It contains eight questions about 10 substance categories: tobacco, alcohol, marijuana, crack/cocaine, methamphetamine/amphetamine type stimulants, inhalants, sedatives, hallucinogens, opioids, and other drugs. A score is determined for each substance and is categorized as low, moderate, or high risk. The ability of the ASSIST to classify patients based on the degree of drug use has been extensively validated (Humeniuk et al., 2008; 2012). Cronbach's alpha was considered moderate to good for alcohol, tobacco, and cannabis. Moreover, the self-report version had acceptable levels of sensitivity (66.7%-100%) and specificity (83.5%-97.1%) for tobacco, alcohol, cannabis and cocaine, using the same cut-off scores of the interview format (Barreto et al., 2014).

Furthermore, students were requested to fill out a socio-demographic and academic form to collect information such as sex, age, nationality, living status, field of study, and academic performance. Additionally, more personal information such as sexual orientation, religion, perceived physical health was requested.

Statistical analysis

Descriptive statistics for the socio-demographic and academic characteristics and the questionnaire scores are given in terms of mean and standard deviation for numerical variables and percentage distribution for categorical variables. Normality assumption for the GHQ-12 and USS scales was assessed with the Shapiro-Wilk test and graphical inspection by QQ-plots. ANOVA tests were applied for comparing GHQ-12 and USS scales across categories of socio-demographic and academic variables, and risky drug use. The regression trees (RT) technique was applied to detect the most important predictors of the target outcomes (see e.g., Saeys, Inza, & Larrañaga, 2007; Loh, He, & Man, 2015; Speiser, 2021) among the socio-demographic student features, of the two main outcomes: psychophysical well-being (GHQ-12) and mental distress (USS). RTs allow to highlight homogeneous groups (i.e., student profiles) with low or high scores at the two outcomes. In details, two separate RTs were carried out on GHQ-12 and USS as dependent variables and socio-demographic variables (in our case categorical regressors) resulted significantly associated with the two scales. The output of the RT is given by different pathways (defined by the estimated regressor cut-offs) and, for each of them, the estimated predicted mean of the dependent variable is provided. Prediction accuracy of the regression trees was evaluated in terms of risk estimate and corresponding standard error (James, Witten, Hastie, & Tibshirani, 2013). See methodological details for RT in Supplementary Materials.

Finally, in order to provide comprehensive student profiles in terms of both psychophysical well-being and mental distress, a Multiple Corresponding Analysis (MCA) technique was performed. GHQ-12 and USS scales were dichotomized based on their corresponding cut-off values (3 and 13, respectively) and the association between sociodemographic and substance variables with the dichotomized outcomes was investigated. The variables included in the MCA were selected among those significantly associated to both GHQ-12 and USS scales. The outcome of this method was represented by a two-dimensional space plot (Biplot) showing the relationship between individuals (points) and the categorical variables. Variable categories that were in the same quadrant or that were close enough were considered in mutual relationship and in association with closer individual subgroups. This graphical representation of MCA results allows to identify specific subject profiles (Rencher, 2003).

All tests were two-tailed and the probability of a type I error was set at $p< .05$. Descriptive analyses were performed using IBM SPSS Statistics for Windows, Version 26.0. The RT machine learning technique was performed through IBM SPSS by applying both Exhaustive CHAID (Chi-squared Automatic Interaction Detection) and CRT (Classification and Regression Tree) methods. The multivariate MCA technique was carried out with software R (R Core Team, 2020, version 3.6.3) with package *FactoMineR*.

Ethical considerations

Ethical approval was obtained from the Institutional Committee of the University of Brescia. The study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. Students were informed that their participation was confidential, anonymous, not compulsory, and that all their personal data would be respected.

RESULTS

Characteristics of the sample

The response rate was 27% (3.754/13.886). The mean age of participants was 23.0 years (SD = 4.6) and most students were female (58.0%). Almost all the sample was of Italian nationality (94.7%), and most of them lived with their families (79.7%). Most of the sample attended a degree course in the Medicine area (35.4%). Most of the sample declared to be heterosexual (91.4%) and Christian (51.9%).

Psychophysical well-being and mental distress

The GHQ-12 mean score was 6.4 (SD = 2.9), indicating psychological distress (cut-off > 3). Students who had high scores, revealing a worst well-being, were atheist/agnostic compared to students of Christianity religion. Medical students showed significantly higher scores than students of the Engineering and Economics areas. Also, students registered on second supplementary year had higher scores than those registered on a regular academic year. Even students with low mean grades had higher scores than students with medium/high mean grades (Table 1).

The USS mean score was 14.5 (SD = 7.7), indicating mental distress (cut-off ≥ 13). Students with the worst mental distress were: those who were married or cohabiting as compared to students who were in a relationship, but did not live with a partner, or were single; medical students compared to students of the Engineering and Economics areas; students with a low mean grade compared to students with a high mean grade (Table 1).

Physical health, psychological problems in the past, and familiar for mental disorder tendency

Half of the sample (52.9%) rated their physical health between excellent and good, and students who reported to have better physical health had lower GHQ-12 and USS scores (Table 1). Twenty-three percent of the sample stated they had requested professional help for psychological problems in the past, and 7.5% stated they had used pharmacological therapy for life. Furthermore, 17.4% of the students reported having a first or second-degree relative suffering from a mental disorder.

Risky drug use

Students reported having used at least once in the past: alcohol (55.3%), tobacco (34.4%), marijuana (20.4%), and sedatives (3.2%). All other psychoactive drugs were below 1%. The use of each drug category was assessed as at low risk for at least 65% of the students. Students who had a moderate/high risky drug use had also significantly higher scores in both GHQ-12 and USS than students with a low drug risk, showing a worse psychophysical well-being and a higher mental distress (Table 1).

Psychophysical well-being and mental distress student pathways

The RT applied to the dependent variable GHQ-12 with eleven significantly associated variables (Table 1) as predictors is depicted in Figure 1. The best RT was the one obtained by the CHAID method (predictive accuracy estimated by cross-validation risk (within node variance) = 7.5, standard error = .16). The variable that most discriminated between high (above cut-off) and low (below cut-off) GHQ-12 scores was "physical health." Other predictors were risky use of at least one psychoactive drug, sex, field of study, mean grade, and alcohol consumption. Interestingly, four main pathways appeared.

Students with excellent physical health had the lowest GHQ-12 estimated mean score (5.2, SD = 2.7; Figure 1: node 1). Among the students who had a good physical health, those who had a low risky use for at least one substance had lower GHQ-12 scores (estimated mean = 5.8, SD = 2.8) than students with a moderate/high risk (estimated mean = 6.6, SD = 2.7). Among students with a low risk, those who had a medium/high mean grade had lower GHQ-12 scores (estimated mean = 5.7, SD = 2.8 vs mean 6.5, SD = 2.7, *p* < .001; Figure 1: nodes 2, 5, 6, 11, 12).

Among the students who had a fair physical health, females had higher GHQ-12 scores than males. Also, among these females, those who had a moderate/high risky use of alcohol had a higher GHQ-12 score than females with a low risky use of alcohol (estimated mean = 8.4, SD = 2.6 vs mean = 7.4, SD = 2.7, *p* < .001; Figure 1: nodes 3, 7, 14). Finally, among students with a poor physical health, medical and economics students had the highest GHQ-12 score (estimated mean = 9.4, SD = 2.2; Figure 1: nodes 4, 10).

The sociodemographic and drug risky pathways related to USS scale are depicted in the RT shown in Figure 2. The RT applied included all variables described in Table 1 as predictors. Also, for the USS outcome, the best RT was the one obtained by the CHAID method (predictive accuracy estimated by cross-validation risk [within node variance] = 53.9, standard error = 2.0). Even in this case, "physical health" was the variable that more discriminated between high and low USS scores. Further predictors were sexual orientation, university registration, living status, alcohol risk, religion, and university status.

Among students who had an excellent physical health, in-town students had lower scores than students living away from home (estimated USS means score 10.8, SD = 5.9 *vs* 12.0, SD = 7.8). Likewise, among in-town students, Christians had lower USS scores than atheists/agnostics and Muslims (estimated mean = 9.5, SD = 5.2; Figure 2: nodes 1, 5, 11).

Among students with good physical health, heterosexuals had lower scores than homosexuals and bisexuals (estimated means 13.1, SD = 7.1 vs 17.9, SD = 7.6; *p* < .001). Similarly, among heterosexuals, students with a moderate/high risky use of alcohol had higher USS scores than those with a low risky use (estimated means 15.0 SD = 7.1 *vs* 12.6 SD = 7.0; *p* < .001; Figure 2: nodes 2, 7, 8).

Among students who had a fair physical health, those who were registered on a regular academic year or on the first supplementary year had lower scores than students registered on the second (or subsequent) supplementary year (estimated means 16.2 SD = 7.4 *vs* 20.0 SD = 8.1; *p* < .001). Among students registered on a regular academic year or on the first supplementary year, heterosexuals had lower mental distress compared to homosexuals and bisexuals (estimated means 15.8 SD = 7.2 *vs* 19.4 SD = 8.1; *p* < .001; Figure 2: nodes 3, 10, 18). Finally, the highest level of mental distress was esti-

Table 1
Comparison among GHQ-12 and USS mean scores and socio-demographic and academic variables, and drug risk (N = 3.754)

	N	*GHQ-12* *M (SD)*	*P value* *(post-hoc)*	*USS* *M (SD)*	*P value* *(Bonferroni post-hoc)*
All sample		6.4 (2.9)	--	14.5 (7.7)	--
Sex					
Male (1)	1.569	6.1 (2.9)	< .001	13.5 (7.6)	< .001
Female (2)	2.178	6.6 (2.9)	(1 *vs* 2)	15.1 (7.7)	(1 vs 2)
Citizenship					
Italian (1)	3.556	6.4 (2.9)	.911	14.3 (7.5)	< .001
Other EU country (2)	99	6.5 (2.9)		17.2 (9.6)	(1 vs 2,3)
Other non-EU country (3)	95	6.3 (3.1)		17.9 (5.8)	
Living with					
Alone (1)	141	6.2 (3.0)	.089	16.5 (8.9)	< .001
Family (2)	2.991	6.4 (2.9)		14.0 (7.4)	
Partner (3)	163	6.3 (2.8)		16.9 (9.4)	(2 *vs* 1,3,4,5)
Friends/other students (4)	208	6.9 (3.1)		15.8 (7.7)	
Other (5)	152	6.3 (2.9)		16.9 (8.6)	
Marital status					
Single (1)	1.659	6.1 (2.7)	.670	14.5 (7.8)	.008
In a relationship (2)	1.825	6.4 (2.9)		14.2 (7.3)	
Married /cohabiting (3)	222	6.4 (2.9)		16.3 (9.1)	(3 *vs* 1,2)
Other (4)	48	6.3 (3.1)		14.8 (8.4)	
Sexual orientation					
Heterosexual (1)	3.429	6.4 (2.9)	.077	14.1 (7.5)	< .001
Bisexual (2)	116	7.2 (3.1)		19.0 (7.7)	
Homosexual (3)	75	6.3 (2.7)		19.6 (8.2)	(1 *vs* 2,3)
Other	20	6.6 (3.1)		17.2 (8.9)	
Do not want to answer	112	6.6 (3.1)		16.1 (8.7)	
Religion					
Atheist/agnostic (1)	1.266	6.6 (3.0)	.009	15.0 (7.6)	< .001
Christianity (2)	1.948	6.2 (2.8)		13.8 (7.6)	
Judaism (3)	2	--	(1 *vs* 2)	--	(2 *vs* 1,7)
Islam (4)	105	7.0 (2.8)		16.1 (7.0)	
Hinduism (5)	5	8.0 (2.9)		15.5 (12.1)	
Buddhism (6)	17	5.8 (2.5)		17.5 (11.2)	
Other (7)	80	6.8 (3.2)		17.9 (8.8)	
Do not want to answer (8)	328	6.5 (3.0)		14.7 (7.5)	
University status					
Full time student	2.730	6.3 (2.9)	.025	14.0 (7.6)	< .001
Student and worker	727	6.6 (3.0)		16.1 (7.9)	
Living status					
In-town students	2.171	6.3 (2.9)	.005	13.9 (7.5)	< .001
Students living away from home	1.287	6.6 (3.0)		15.5 (7.9)	
Field of study					
Medicine (1)	1.330	6.6 (2.9)	< .001	15.1 (7.5)	.009
Engineering (2)	1.069	6.4 (2.9)	(1 *vs* 3,4)	14.1 (7.7)	
Economics (3)	754	6.2 (2.9)	(2 *vs* 4)	14.0 (7.8)	(1 *vs* 2,3)
Law (4)	306	5.9 (2.7)		14.1 (8.0)	
Registration					
Registered on a regular academic year (1)	2.774	6.3 (2.9)	< .001	13.9 (7.4)	< .001
Registered on first supplementary year (2)	315	6.7 (2.9)		15.2 (7.4)	
Registered on second supplementary year (3)	142	7.4 (3.0)	(3 *vs* 1)	18.0 (8.1)	(1,2 *vs* others)
Registered on other supplementary year (4)	227	6.7 (3.0)		17.6 (9.7)	
Grades (Mean)					
Low (18-22) (1)	509	6.8 (2.9)	.002	15.5 (8.2)	.002
Medium (22-26) (2)	1.705	6.2 (2.9)	(1 *vs* 2,3)	14.5 (7.6)	(1 *vs* 3)
High (26-30) (3)	1.243	6.4 (2.9)		14.0 (7.4)	
How would you rate your physical health?					
Excellent	437	5.2 (2.7)	< .001	11.0 (6.6)	< .001
Good	1.551	5.9 (2.8)		13.4 (7.2)	
Fair	1.050	7.3 (2.8)	(each	16.7 (7.6)	(each *vs* others)
Poor	115	8.4 (2.6)	*vs* others)	20.8 (9.0)	
Risky drug use (ASSIST) Tobacco					
Low	1.905	6.3 (2.9)	.034	13.9 (7.6)	< .001
Moderate/high	910	6.6 (2.9)		15.5 (7.7)	
Alcohol					
Low	2.198	6.3 (2.9)	.006	13.9 (7.4)	< .001
Moderate/high	617	6.7 (3.0)		16.5 (8.3)	
Risk for at least one of the other psychoactive drugs					
Low	2.382	6.3 (2.9)	< .001	14.1 (7.4)	< .001
Moderate/high	433	6.9 (2.9)		16.4 (8.6)	

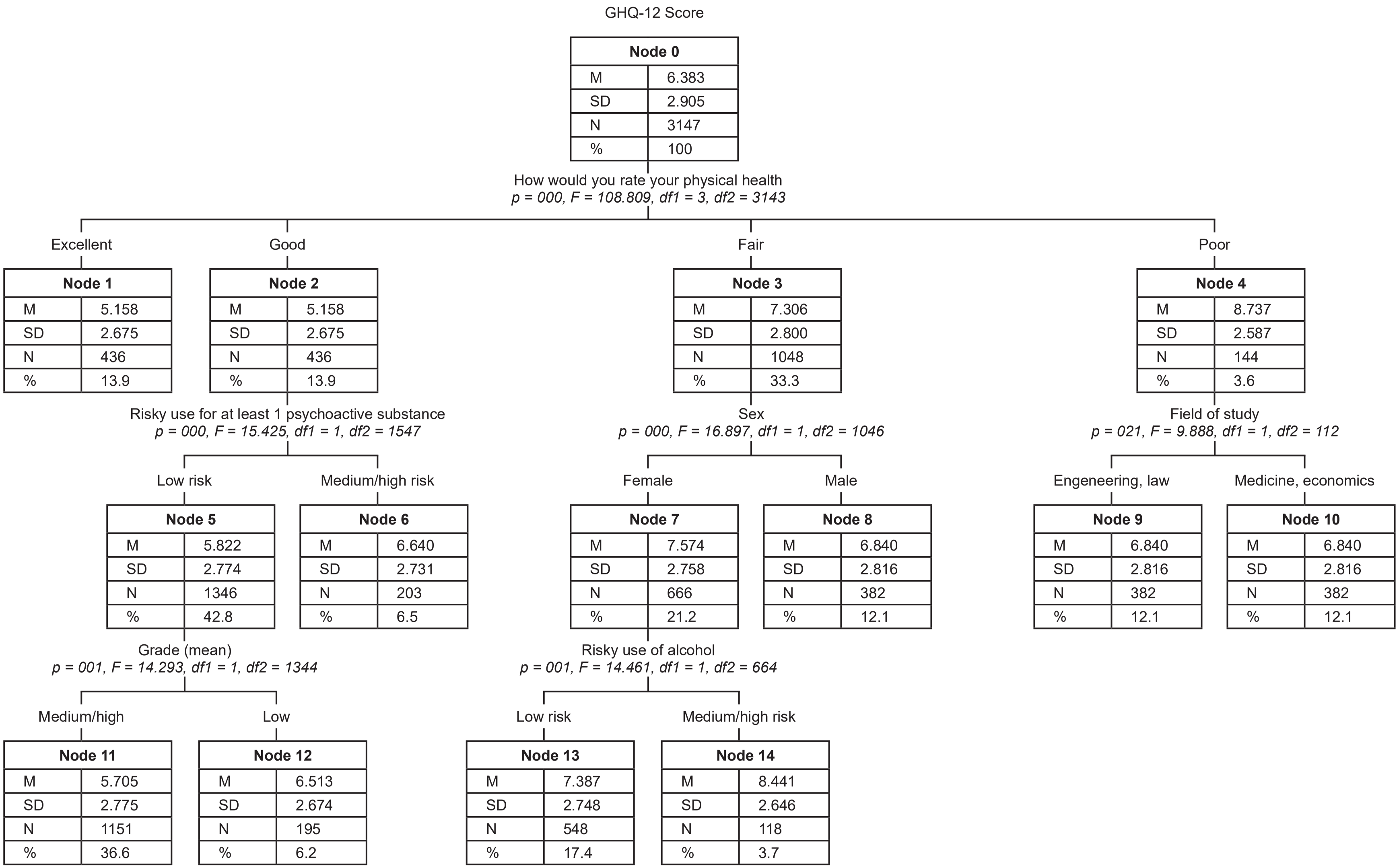

Figure 1. Regression tree on GHQ-12 as dependent variable.

Note. The most prominent (from top to bottom) predictors of the GHQ-12 score are reported. At each level of the RT, the name of the predictor and its corresponding test result about its association with the above (parent) variable is reported. In each node are reported: the number of observations (N) belonging to the node and the corresponding percentage (%) on the total study sample; the mean (M) and the standard deviation (SD) of the GHQ-12 score of the subgroup of N observations (subjects) of the node.

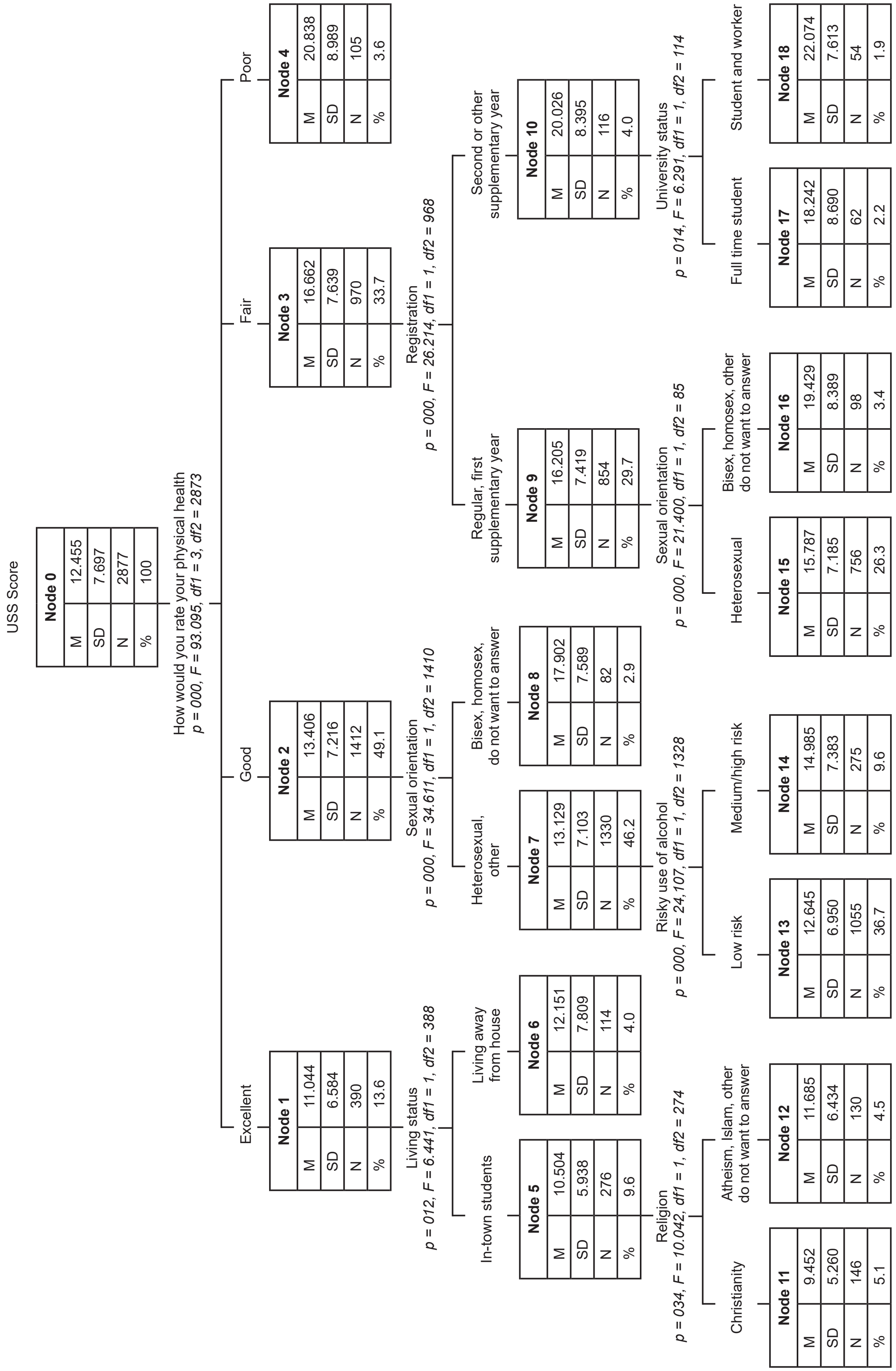

Figure 2. Regression tree on USS as dependent variable.

Note. The most prominent (from top to bottom) predictors of the USS score are reported. At each level of the RT, the name of the predictor and its corresponding test result about its association with the above (parent) variable is reported. In each node are reported: the number of observations (N) belonging to the node and the corresponding percentage (%) on the total study sample; the mean (M) and the standard deviation (SD) of the USS score of the subgroup of N observations (subjects) of the node.

mated for students registered on the second (or subsequent) supplementary year who were also working students (estimated USS mean score = 22.1, SD = 7.0; Figure 2: nodes 4, 18).

Psychophysical well-being and mental distress student profiles

The MCA was performed in order to assess the simultaneous association of the sociodemographic, academic, and drug use categorical variables with the high and low GHQ-12 and USS categories. We firstly included in the MCA all the variables associated to both USS and GHQ-12 (Table 1). Then we selected those that contributed the most to explain the total variability of the data set, resulting in the following list: physical health, field of study, risk of alcohol use, risk of psychoactive drug use, sex, and grades. This selection was based on a specific function of the MCA (fviz_contrib() of factoextra R package) which allowed to evaluate the contribution of each variable categories (in %) to the definition of the two-dimensional space plot (Biplot). The first horizontal dimension is mainly characterized by the risk of unhealthy behaviours and grades: from left towards right the grades decrease and the risks for unhealthy behaviours increase. The second dimension (vertical axe) is characterized by physical health with improved health condition moving from the bottom to the top.

The MCA applied in such subsample of variables confirms the findings obtained by RTs. Through the MCA Biplot of Figure 3A, two distinct subjects' profiles, based on the GHQ-12 score, were identified: azure and red ellipses represent subjects having low and medium-high GHQ-12 scores respectively. The low GHQ-12 score cluster is mainly characterized by excellent physical health. Conversely, the high GHQ-12 score profile is mainly characterized by fair and poor physical health together with high USS score, a high risky use of alcohol and psychoactive drug and field of study (Medicine). Interestingly, females turn out to be in the high GHQ-12 cluster.

Figure 3B shows that the low USS score cluster is mainly characterized by excellent physical health and low GHQ-12 score. Similarly to the high GHQ-12 profile, the high USS score profile is characterized by fair and poor physical health together with a high GHQ-12 score, a high risky use of alcohol, and psychoactive drugs, for a medical student. As for GHQ-12, females turn out to be in the high USS score cluster while males were not characteristic of any profile.

DISCUSSION AND CONCLUSION

To our knowledge this is the first study carried out in Italy on a large sample of college students. Still, the response rate (27%) was not very high, even if the available data on web-surveys among college students show a wide variability: from 10% to 80% approximately (Kim, Sinn, & Syn, 2018; Kenney, DiGuiseppi, Meisel, Balestrieri, & Barnett, 2018).

The coherence of the student profiles and pathways found by the two multivariate/multiple techniques guarantee (MCA and RT) the robustness and reliability of the findings. Moreover, RT was a particularly suitable technique to allow for the identification of a rank of predictors of psychophysical well-being and mental distress, and to provide cut-offs for each of the prominent predictors which is very helpful in a clinical perspective to highlight subjects' profiles. This study showed that physical health, risk of use of alcohol and psychoactive drugs, sex, and field of study were the characteristics more associated and discriminant between low and high level of well-being and distress among the student population. In particular, fairly or poorly perceived physical health was the strongest predictor of mental distress and compromised well-being. This is in line with previous literature, that demonstrated how physical activity was strongly correlated to well-being (Budzynski-Seymour et al., 2020; Klainin Yobas et al., 2014). Being a female student was another factor towards higher perceived distress in college settings. Also, this result confirms data from the literature (Bayram & Bilgel, 2008; Beiter et al., 2015; Stowell, Lewis, & Brooks, 2019). A possible explanation of this result could be related to the different coping strategies characterizing sex differences. Typically, females have been associated to emotion-focused coping styles, sometimes doomed to be maladaptive, whereas males have been associated to approach-focused coping styles, often described as adaptive (Brougham, Zail, Mendoza, & Miller, 2009; Mahmoud, Staten, Hall, & Lennie, 2021). Moreover, the relationship between grades and psychological well-being could be seen both ways. Students with good grades have consequently more self-efficacy, more adaptive coping skills, and generally a better quality of life thanks to their academic achievements. On the other hand, students who are more exposed to distress, or whose well-being is compromised, for a reason will present most probably temporary cognitive issues, which would ultimately affect their academic performance. Either type of relation between academic performance and psychological distress has already been confirmed by the literature (Click, Huang, & Kline, 2017; Lin et al., 2020; Sajid, Ahmad, & Khalid, 2015), furthermore, once more confirming existing literature (Bhochhibhoya, Collado, Branscum, & Sharma, 2015; Blank, Connor, Gray, & Tustin, 2016; Sebena, El Ansari, Stock, Orosova, & Mikolajczyk, 2012). In this particular study there is a strong correlation between risky alcohol use and perceived mental distress. Tembo, Burns, & Kalembo (2017) suggest it is the academic performance that predicts high alcohol consumption, confirming that both variables contribute to mental distress in college settings. Whereas

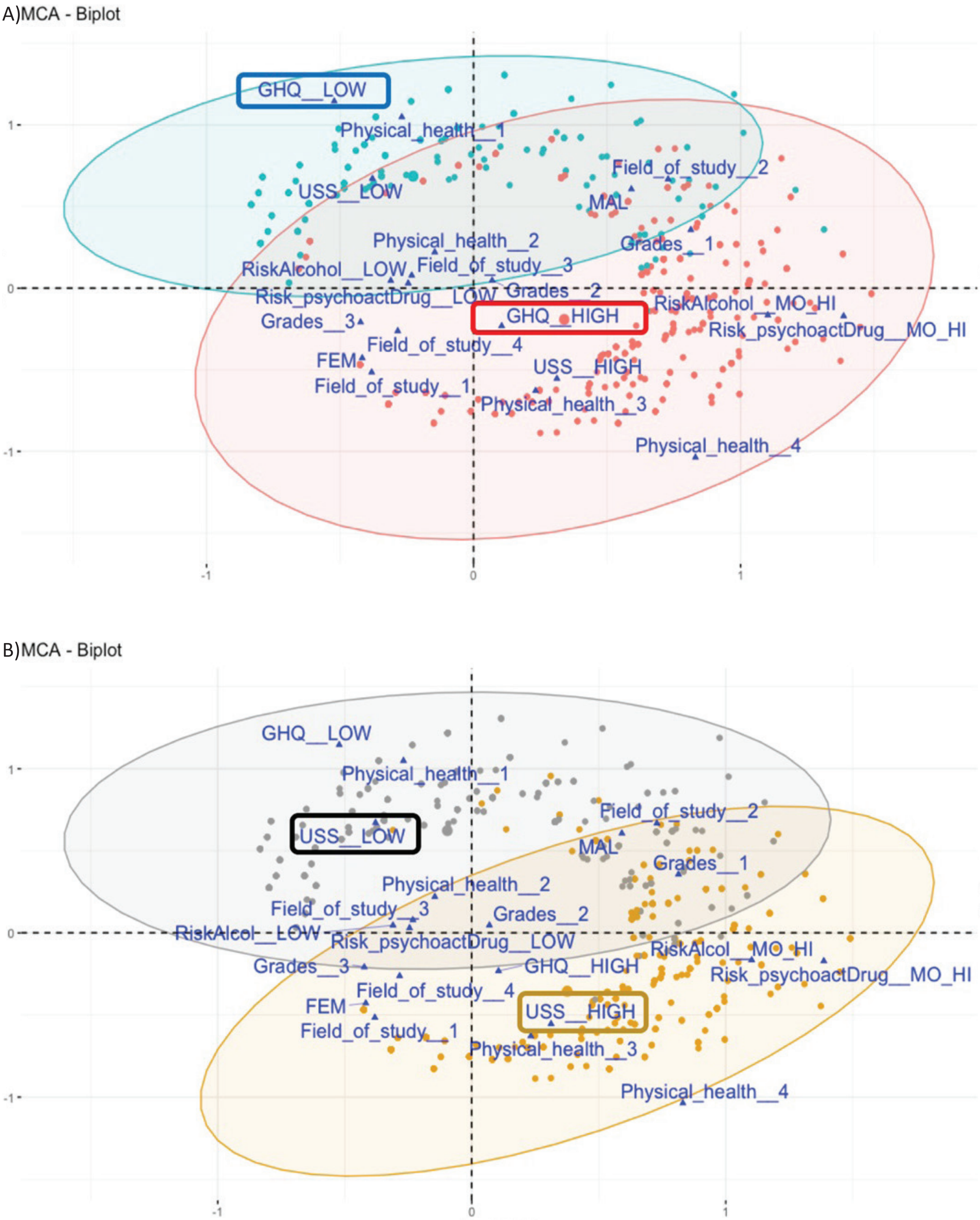

Figure 3. A) MCA Biplot (ellipses for GHQ-12 categories Low vs High). **B)** MCA Biplot (ellipses for USS categories Low vs High).

Notes: Variable categories legend: Physical_health_1: excellent; Physical_health_2: good; Physical_health_3: fair; Physical_health_4: poor; FEM: sex female; MAL: sex male; Field_of_study_1: Medicine; Field_of_study_2: Engineering; Field_of_study_3: Economics; Field_of_study_4: Law; Risk_psychoactDrug_LOW: low risk for at least one of the other psychoactive drugs; Risk_psychoactDrug_MO_HI: moderate/high risk for at least one of the other psychoactive drugs; RiskAlcohol_LOW: low risk of use alcohol; RiskAlcohol_MO_HI: moderate/high risk of use alcohol; Grades_1: low grades; Grades_2: medium grades; Grades_3: high grades.

The first horizontal dimension is mainly characterized by the risk of unhealthy behaviours and grades: from left toward right the grades decrease and the risks for unhealthy behaviours increase. The second dimension (vertical axis) is characterized by physical health with improved health condition moving from the bottom to the top.

in another study there is an interaction among sex female and risky alcohol use in increasing mental distress (Pedrelli, Borsari, Lipson, Heinze, & Eisenberg, 2016).

Other results of this study suggest that being a working-student contributed to increase academic distress. This result is quite plausible since being responsible for two major commitments such as work and study, requires high organizational skills and several efforts on the part of working-students, compared to their non-working peers. Typically, working-students are also older and have more financial obligations deriving from the greater weight of the costs they have to deal with (Mukherjee, McKinney, Hagedorn, Purnamasari, & Martinez, 2017). Religion is another impacting factor in stress appraisal in academic settings. As it has been evidenced also in a recent meta-analysis, being atheistic or agnostic brings more distress than practicing or being oriented to some religion (Forouhari et al., 2019). Religion is a coping strategy that has often been connected to well-being because of its assimilation to the transcendental need to rely in a bigger strength (Mahmoud, Staten, Hall, & Lennie, 2021). Lastly, in this study it has been appreciated that students living alone and/or far from family suffer more from mental distress and present lower well-being. This correlation has been investigated before and it has been reported that off-campus students presented higher stress levels (Beiter et al., 2015). These correlations are interpretable in the light of the social network support: students with a small or no social network (i.e., living alone) have less opportunities to bond with peers or to ask for instrumental support. Conversely, students living far from home feel a higher perception of distress since they are not supported by key or attachment figures in this period of transition. Social support is a key coping strategy in college environments (Chao, 2012; Hefner & Eisenberg, 2009; Stallman, 2010), and compromised ability in this skill amounts to lower chances in getting help and coping with stress optimally in academic settings.

The results of this study shows that it would be very important to promote in college students healthy behaviors with the aim of increasing physical exercise and reducing substance use, most of all alcohol. Moreover, it would be desirable to improve academic counselling services in order to detect more vulnerable students; especially those students who belonging to the identified vulnerability framework could be involved in interventions, such as psychoeducational sessions, in order to learn more on healthy behaviors and how these affect their academic performance and quality of life. College counselling can represent a key front-line service in early detecting sub-threshold symptoms mostly related to generic mental distress in young adults. This is particularly important since most mental diseases have their onset between the ages of 18 and 24 (Kessler et al., 2005; 2007). Thus, counselling services turn out to be a necessity and a first step to support and avoid degeneration, sorrow, and burnout.

This study has several limitations due to the generalizability of the results, since data was gathered sampling only students of a single university in Northern Italy. In this way, it is not far-fetched that evidences obtained were affected by a cultural bias in perceiving stress or university life. Moreover, sampling only a single university carries on further issues on generalizability, regarding groups scarcely represented. This is the case of foreign students as well as those from religions underrepresented in our sample. Finally, it should be noted that predictors found in this study were conditioned by the analyzed sample and the statistical strategies used as well as by the variables collected. For instance, we had no information on the personality characteristics of these students or on their coping strategies for managing stress. Had we gathered more data we might have been able to find other predictors. Further studies are needed to better understand which are the best predictors of psychophysical well-being and mental distress in college students.

Funding

None.

Conflicts of interest

The authors declare they have no conflicts of interest.

REFERENCES

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., … WHO WMH-ICS Collaborators. (2018). The WHO World Mental Health Surveys International College Student Project: Prevalence and Distribution of Mental Disorders. *Journal of Abnormal Psychology*, *127*(7), 623-638. doi: 1037/abn0000362

Barreto, H. A., de Oliveira, A., & Boerngen-Lacerda, R. (2014). Development of a self-report format of ASSIST with university students. *Addictive Behaviors*, *39*(7), 1152-1158. doi: 10.1016/j.addbeh.2014.03.014

Bayram, N., & Bilgel, N. (2008). The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Social Psychiatry and Psychiatric Epidemiology*, *43*(8), 667-672. doi: 10.1007/s00127-008-0345-x

Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M., & Sammut, S. (2015). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders*, *173*, 90-96. doi: 10.1016/j.jad.2014.10.054

Bhochhibhoya, A., Collado, M., Branscum, P., & Sharma, M. (2015). The role of global mental health and type-D personality in predicting alcohol use among a sample of college students. *Alcoholism Treatment Quarterly*, *33*(3), 283-295. doi: 10.1080/07347324.2015.1050932

Blanco, C., Okuda, M., Wright, C., Hasin, D. S., Grant, B. F., Liu, S.-M., & Olfson, M. (2008). Mental health of college students and their non-college-attending peers: Results from the National Epidemiologic Study on Alcohol and Related Conditions. *Archives of General Psychiatry*, *65*(12), 1429-1437. doi: 10.1001/archpsyc.65.12.1429

Blank, M.-L., Connor, J., Gray, A., & Tustin, K. (2016). Alcohol use, mental well-being, self-esteem and general self-efficacy among final-year university students. *Social Psychiatry and Psychiatric Epidemiology*, *51*(3), 431-441. doi: 10.1007/s00127-016-1183-x

Brougham, R. R., Zail, C. M., Mendoza, C. M., & Miller, J. R. (2009). Stress, sex differences, and coping strategies among college students. *Current Psychology*, *28*(2), 85-97. doi: 10.1007/s12144-009-9047-0

Bruffaerts, R., Mortier, P., Auerbach, R. P., Alonso, J., Hermosillo De la Torre, A. E., Cuijpers, P., WHOWMH-ICS Collaborators. (2019). Lifetime and 12-month treatment for mental disorders and suicidal thoughts and behaviors among first year college students. *International Journal of Methods in Psychiatric Research*, *28*(2), e1764. doi: 10.1002/mpr.1764

Budzynski-Seymour, E., Conway, R., Wade, M., Lucas, A., Jones, M., Mann, S., & Steele, J. (2020). Physical activity, mental and personal well-being, social isolation, and perceptions of academic attainment and employability in university students: the Scottish and British Active Students Surveys. *Journal of Physical Activity and Health*, *17*(6), 610-620. doi: 10.1123/jpah.2019-0431

Chao, R. C.-L. (2012). Managing perceived stress among college students: The roles of social support and dysfunctional coping. *Journal of College Counseling*, *15*(1), 5-21. doi: 10.1002/j.2161-1882.2012.00002.x

Click, K. A., Huang, L. V., & Kline, L. (2017). Harnessing inner strengths of at-risk university students: relationships between well-being, academic achievement and academic attainment. *Perspectives: Policy and Practice in Higher Education*, *21*(2-3), 88-100. doi: 10.1080/13603108.2016.1273260

Cordero-Oropeza, R., García-Méndez, M., Cordero-Oropeza, M., & Corona-Maldonado, J. J. (2021). Characterization of alcohol consumption and related problems in university students from Mexico City. *Salud Mental*, *44*(3), 107-115. doi: 10.17711/SM.0185-3325.2021.015

Edwards, P. J., Roberts, I., Clarke, M. J., DiGuiseppi, C., Wentz, R., Kwan, I., … Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, *8*(3), MR000008. doi: 10.1002/14651858.MR000008.pub4

Eisenberg, D., Hunt, J., & Speer, N. (2012). Help-seeking for mental health on college campuses: review of evidence and next steps for research and practice. *Harvard Review of Psychiatry*, *20*(4), 222-232. doi: 10.3109/10673229.2012.712839

Forouhari, S., Teshnizi, S. H., Ehrampoush, M. H., Mahmoodabad, S. S. M., Fallahzadeh, H., Tabei, S. Z., ... Dehkordi, J. G. (2019). Relationship between religious orientation, anxiety, and depression among college students: A systematic review and meta-analysis. *Iranian Journal of Public Health*, *48*(1), 43-52.

Goldberg, D. P., & Blackwell, B. (1970). Psychiatric illness in general practice: A detailed study using a new method of case identification. *British Medical Journal*, *2*, 439-443. doi: 10.1136/bmj.2.5707.439

Gunnel, D., Caul, S., Appleby, L., John, A., & Hawton, K. (2020). The incidence of suicide in University students in England and Wales 2000/2001-2016/2017: Recors linkage study. *Journal of Affective Disorders*, *261*, 113-120. doi: 10.1016/j.jad.2019.09.079

Hefner, J., & Eisenberg, D. (2009). Social support and mental health among college students. *American Journal of Orthopsychiatry*, *79*(4), 491-499. doi: 10.1037/a0016918

Humeniuk, R., Ali, R., Babor, T. F., Farrell, M., Formigoni, M. L., Jittiwutikarn, J., … Simon, S. (2008). Validation of the alcohol, smoking and substance involvement screening test (ASSIST). *Addiction*, *103*(6), 1039-1047. doi: 10.1111/j.1360-0443.2007.02114.x

Humeniuk, R., Ali, R., Babor, T., Souza-Formigoni, M. L. O., Boerngen de Lacerda, R., Ling, W., … Vendetti, J. (2012). A randomized controlled trial of a brief intervention for illicit drugs linked to the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) in clients recruited from primary health-care settings in four countries. *Addiction*, *107*(5), 957-966. doi: 10.1111/j.1360-0443.2011.03740.x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications to R*. New York: Springer.

Kang, H. K., Rhodes, C., Rivers, E., Thornton, C. P., & Rodney, T. (2021). Prevalence of Mental Health Disorders Among Undergraduate University Students in the United States: A Review. *Journal of Psychosocial Nursing and Mental Health Services*, *59*(2), 17-24. doi: 10.3928/02793695-20201104-03

Kenney, S. R., DiGuiseppi, G. T., Meisel, M. K., Balestrieri, S. G., & Barnett, N. P. (2018). Poor mental health, peer drinking norms, and alcohol risk in a social network of first-year college students. *Addictive Behaviors*, *84*, 151-159. doi: 10.1016/j.addbeh.2018.04.012

Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Ustün, T. B. (2007). Age of onset of mental disorders: A review of recent literature. *Current Opinion in Psychiatry*, *20*(4), 359-364. doi: 10.1097/YCO.0b013e32816ebc8c

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 593-602. doi: 10.1001/archpsyc.62.6.593

Kim, S., Sinn, D., & Syn, S. Y. (2018). Analysis of College Students' Personal Health Information Activities: Online Survey. *Journal of Medical Internet Research*, *20*(4), e132. doi: 10.2196/jmir.9391

Klainin Yobas, P., Keawkerd, O., Pumpuang, W., Thunyadee, C., Thanoi, W., & He, H.-G. (2014). The mediating effects of coping on the stress and health relationships among nursing students: A structural equation modelling approach. *Journal of Advanced Nursing*, *70*(6), 1287-1298. doi: 10.1111/jan.12283

Lin, X.-J., Zhang, C.-Y., Yang, S., Hsu, M.-L., Cheng, H., Chen, J., & Yu, H. (2020). Stress and its association with academic performance among dental undergraduate students in Fujian, China: a cross-sectional online questionnaire survey. *BMC Medical Education*, *20*(1), 181. doi: 10.1186/s12909-020-02095-4

Liu, C. H., Stevens, C., Wong, S. H. M, Yasui, M., & Chen, J. A. (2018). The prevalence and predictors of mental health diagnoses and suicide among U.S. college students: Implications for addressing disparities in service use. *Depression and Anxiety*, *36*(1), 8-17. doi: 10.1002/da.22830

Loh, W.-Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatmenteffects. *Statistics in Medicine*, *34*(11), 1818-1833. doi: 10.1002/sim.6454

Lovell, G. P., Nash, K., Sharman, R., & Lane, B. R. (2015). A cross-sectional investigation of depressive, anxiety, and stress symptoms and health-behavior participation in Australian university students. *Nursing & Health Sciences*, *17*(1), 134-142. doi: 10.1111/nhs.12147

Mahmoud, J. S. R., Staten, R. T., Hall, L. A., & Lennie, T. A. (2021). The relationship among young adult college students' depression, anxiety, stress, demographics, life satisfaction, and coping styles. *Issues in Mental Health Nursing*, *33*(3), 149-156. doi: 10.3109/01612840.2011.632708

McLafferty, M., Lapsley, C. R., Ennis, E., Armour, C., Murphy, S., Bunting, B. P., … O'Neill, S. M. (2017). Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland. *PLoS One*, *13*, 12(12), e0188785. doi: 10.1371/journal.pone.0188785

Mortier, P., Cuijpers, P., Kiekens, G., Auerbach, R. P., Demyttenaere, K., Green, J. G., … Bruffaerts, R. (2018). The prevalence of suicidal thoughts and behaviours among college students: a meta-analysis. *Psychological Medicine*, *48*(4), 554-565. doi: 10.1017/S0033291717002215

Mukherjee, M., McKinney, L., Hagedorn, L. S., Purnamasari, A., & Martinez, F. S. (2017). Stretching every dollar: The impact of personal financial stress on the enrollment behaviors of working and nonworking community college students. *Community College Journal of Research and Practice*, *41*(9), 551-565. doi: 10.1080/10668926.2016.1179602

Oh, H. Y., Marinovich, C., Jay, S., Zhou, S., & Kim, J. H. J. (2021). Abuse and suicide risk among college students in the United States: Findings from the 2019 Healthy Minds Study. *Journal of Affective Disorders*, *282*, 554-560. doi: 10.1016/j.jad.2020.12.140

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: a global public-health challenge. *Lancet*, *369*(9569), 1302-1313. doi: 10.1016/S0140-6736(07)60368-7

Pealer, L., & Weiler, R. M. (2003). Guidelines for designing a Web-delivered college health risk behavior survey: lessons learned from the University of Florida Health Behavior Survey. Health Promotion Practice, *4*(2), 171-179. doi: 10.1177/1524839902250772

Pedrelli, P., Borsari, B., Lipson, S. K., Heinze, J. E., & Eisenberg, D. (2016). Gender differences in the relationships among major depressive disorder, heavy alcohol use, and mental health treatment engagement among college students. *Journal of Studies on Alcohol and Drugs*, *77*(4), 620-628. doi: 10.15288/jsad.2016.77.620

Politi, P. L., Piccinelli, M., & Wilkinson, G. (1994). Reliability, validity and factor structure of the 12-item General Health Questionnaire among young males in Italy. *Acta Psychiatrica Scandinavica*, *90*(6), 432-437. doi: 10.1111/j.1600-0447.1994.tb01620.x

Poznyak, V. (2008). *Management of substance abuse: The WHO ASSIST (The Alcohol, Smoking and Substance Involvement Screening Test) Project*. Geneva: World Health Organization.

Rencher, A. C. (2003). *Methods of Multivariate Analysis* (2nd. Edition), *Wiley Series in Probability and Statistics*. USA: John Wiley & Sons, Inc.

Rith-Najarian, L. R., Boustani, M. M., & Chorpita, B. F. (2019). A systematic review of prevention programs targeting depression, anxiety, and stress in university students. *Journal of Affective Disorders*, *257*, 568-584. doi: 10.1016/j.jad.2019.06.035

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. doi: 10.1093/bioinformatics/btm344

Sajid, A., Ahmad, T., & Khalid, T. (2015). Stress in medical undergraduates; its association with academic performance. *Bangladesh Journal of Medical Science*, *14*(2), 135-141. doi: 10.3329/bjms.v14i2.21815

Sebena, R., El Ansari, W., Stock, C., Orosova, O., & Mikolajczyk, R. T. (2012). Are perceived stress, depressive symptoms and religiosity associated with alcohol consumption? A survey of freshmen university students across five European countries. *Substance Abuse Treatment, Prevention, and Policy*, *7*(1), 21. doi: 10.1186/1747-597X-7-21

Skidmore, C. R., Kaufman, E. A., & Crowell, S. E. (2016). Substance use among college students. *Child and Adolescent Psychiatric Clinics of North America*, *25*(4), 735-753. doi: 10.1016/j.chc.2016.06.004

Speiser, J. L. (2021). A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. *Journal of Biomedical Informatics*, *117*, 103763. doi: 10.1016/j.jbi.2021.103763

Stallman, H. M. (2008). *University Stress Scale*. Queensland University of Technology.

Stallman, H. M. (2010). Psychological distress in university students: A comparison with general population data. *Australian Psychologist*, *45*(4), 249-257. doi: 10.1080/00050067.2010.482109

Stallman, H. M., & Hurst, C. P. (2016). The University Stress Scale: Measuring Domains and Extent of Stress in University Students. *Australian Psychologist*, *51*(2), 128-134. doi: 10.1111/AP.12127

Stowell, D., Lewis, R. K., & Brooks, K. (2019). Perceived stress, substance use, and mental health issues among college students in the Midwest. *Journal of Prevention & Intervention in the Community*, *49*(3), 221-234. doi: 10.1080/10852352.2019.1654263

Tembo, C., Burns, S., & Kalembo, F. (2017). The association between levels of alcohol consumption and mental health problems and academic performance among young university students. *PLoS One*, *12*(6), e0178142. doi: 10.1371/journal.pone.0178142

World Health Organization. (2016). *Suicide data*.

SUPPLEMENTARY MATERIALS

Supervised methods

Classification And Regression Trees (CART)

Regression trees are one of the most popular supervised machine learning algorithms that belong to the family of decision trees, that can be used for both classification and regression purpose. The representation for the CART model is a binary tree. In our specific case we performed a regression tree (RT) in which the dependent variable (labels variable that has to be predicted) is continuous and the independent ones (covariates) can be categorical or quantitative.

RTs are directed graphs in which there is an initial node that branches to many. Each node represents an independent variable, each edge corresponds to a decision rule and each leaf represents an outcome (a value of the predicted variable). The top node contains all the sample that is consequently divided into different subsets. If the covariates are quantitative splits are created on the basis of some cut-offs on a scale; if the covariates are categorical, splits are based on the different categories (Wilkinson, 1992).

After computing the entire tree (with all the independent variables) some techniques have to be used to reduce tree dimension and to improve the tree predictive power, reducing overfitting. Among these, one of the most used is pruning (Breiman, Friedman, Olshen, & Stone, 1984), a method that allows to remove the variables that do not contribute (are not significantly associated) to the final outcome, considering a penalty for the increase of parameters in the model.

Therefore, the final tree shows only the independent variables that are significant predictors of the dependent one (outcome) and, differently from the traditional regression models, those that are not predictors do not influence the final result.

The RT can be built by a recursive partitioning program using a two-stage procedure (Therneau, Atkinson, & Mayo Foundation, 2018):

1. The variable which best splits the data into groups (i.e. with the greatest association with the dependent variable) is found. The subjects are divided and this process is repeated separately to each subject subgroup recursively until the subgroups either reach a minimum size or until no improvement (in terms of predictive performance) can be made (stopping criteria);
2. A cross-validation (pruning) will be performed to trim the full tree in order to overcome the possible overfitting problem and to improve the readability and interpretability of results by reducing the RT complexity.

RT fitting was carried out by performing two different growth algorithms: CHAID (a multi-way tree algorithm that builds segments and profiles with respect to the desired outcome) and CRT (a binary tree algorithm that partitions data and produces homogeneous subsets) available in SPSS software.

The input variables were the 11 variables resulted significantly associated to both GHQ-12 and USS scales as reported in Table 1S.

The choice of the best algorithm and corresponding fit was based on the predictive accuracy estimated by validation risk (within node variance). Regarding the stopping criteria, the depth of the tree was set by default (3 levels for CHAID algorithm and 5 levels for CRT algorithm); similarly, considering our large sample, we retained valid the default setting for the minimum cases per node: equal to 50 for child and equal to 100 for parent nodes. For CHIAD algorithm, the splitting nodes and merging categories parameters were set as default SPSS values. For CRT algorithm, the pruning criteria was based on maximum difference in risk set at 0.3.

For both algorithms, the validation procedure involved split datasets: training (70%) and test (30%): the results on training and test samples resulted comparable in terms of predictive accuracy measured by risk estimate (i.e. within-node variance) and coherent with the fit-all results (i.e. results obtained by using the whole dataset). However, it is worth to note that our main purpose was to provide a comprehensive profiles of student well-being and distress by the identification of valuable predictors of psychophysical well-being and mental distress. With this regard, the application of machine learning approaches has to be evaluated not for prediction capability but (mainly) in terms of detection of best predictors in a context of multiple variables analysis (see e.g., Saeys, Inza, & Larrañaga, 2007; Loh, He, & Man, 2015; Speiser, 2021)

Table 1S
Accuracy and validation parameters of Regression trees

Regression tree (Outcome)	*Tree growing criteria*	*Accuracy*	*Risk estimate*	*Best predictors*
GHQ-12	Non binary Chaid method	RMSE = 6.1	Training: 7.3 (SE = 0.2)	1. Physical health
	Split decided by Pearson's chi-squared	MAE = 1.8	Test: 7.9 (SE = 0.2)	2.1 Psychoactive substance
	Max depth = 3			2.2 Sex
				2.3 Field of study
				3.1 Grades
				3.2 Risky use of alcohol
USS	Min parent	RMSE = 11.8	Training: 51.7 (SE = 1.95)	1. Physical health
	Size = 100	MAE = 2.4	Test: 53.9 (SE = 2.03)	2.1 Living status
	Min child size = 50			2.4 Sexual orientation
	Signif. level for split = .05			2.3 Registration
				3.1 Religion
				3.2 Risky use of alcohol
				3.3 University status

Note: RMSE = Root Mean Square Error; MAE = Mean Absolute Error; SE = standard error.

Unsupervised methods

Multiple Correspondence Analysis (MCA)

Multiple correspondence analysis is a generalization of correspondence analysis. It is a multivariate technique used when there are more than two categorical variables, with the purpose to study the association between the different categories of all the variables involved in the study to identify individuals with similar profiles (i.e. with the highest number of common categories). The final outcome is a plot that shows the relationships among categories, among subjects and among categories and subjects in a two-dimensional space in order to display the geometric configuration of the variable categories. Categories that are in the same quadrant or that are close enough suggest an association (Rencher, 2003). Substantially, the aim of the MCA is to obtain a measure of the association in terms of geometric distance so that associated categories are closely displayed in the output plot. The geometric distances are based on the definition of row and column profiles as defined in Ferrari, Macis, Rossi, and Cameletti (2018). Since MCA involves individuals and variable, two kind of distances can be evaluated: between row profiles (i.e. between individuals) and between column profiles (i.e. between categories of variables). The row-profiles distance will be equal to zero if the individuals have the same categories and it will increase when the number of distinct categories presented by the two subjects increases. Similarly, the column–profiles distance will be the same when these are shown by the same subjects and the distance will increase with the number of individuals that show different categories.

These distances will be displayed in a common unique plot (named Biplot; Greenacre, 1993) such that the distance between any row profile or column profile gives the measure of their similarity (or dissimilarity).

In the MCA implementation for this study, we firstly included all the variables associated to both USS and GHQ-12, (i.e. sex, religion, university status, living status, field of study, registration, grades, physical health, risk of use tobacco, risk of use alcohol, risk of use psychoactive drug); then we selected those that contributed the most in explaining the total variability of the data set by using a specific function of MCA (fviz_contrib() of factoextra R package). This function allowed to evaluate the contribution of each variable categories (in %) to the definition of the two-dimensional space plot (Biplot).

The selection provided the following list: physical health, field of study, risk of use alcohol, risk of use psychoactive drug, sex, grades. This selection was based on a specific fuction of MCA (fviz_contrib() of factoextra R package) which allowed to evaluate the contribution of each variable categories (in %) to the definition of the two-dimensional space plot (Biplot).

REFERENCES

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Ferrari, C, Macis, A., Rossi, R., & Cameletti M. (2018). Multivariate Statistical Techniques to Manage Multiple data in Psychology. *Open Access Journal of Behavioural Science & Psychology*, *1*(2).

Greenacre, M. J. (1993). Biplot in correspondence analysis. *Journal of Applied Statistics*, *20*(2), 251-269. doi: 10.1080/02664769300000021

Loh, W.-Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, *34*(11), 1818-1833. doi: 10.1002/sim.6454

Rencher, A. C. (2003). *Methods of Multivariate Analysis*, (2nd. Ed). USA: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. doi: 10.1093/bioinformatics/btm344

Speiser, J. L. (2021). A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. *Journal of Biomedical Informatics*, *117*, 103763. doi: 10.1016/j.jbi.2021.103763

Therneau, T. M., Atkinson, E., & Mayo Foundation. (2018). *An introduction to Recursive Partitioning Using the RPART Routines*. Available at https://cran.rproject.org/web/packages/rpart/vignettes/longintro.pdf

Wilkinson, L. (1992). *Tree Structured Data Analysis: AID, CHAID and CART*. Sun Valley: Sawtooth/SYSTAT Joint Software Conference, ID.