Prediction of depressive symptoms from socioeconomic data and DNA methylation signatures in depression

Laura Guerrero Simón
Master in Bioinformatics and Biostatistics
Genetics studies of human diseases

Dr Helena Brunel Montaner
Dr David Merino Arranz
8 January 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Prediction of depressive symptoms from socioeconomic data and DNA methylation signatures in depression*
Nombre del autor:	*Laura Guerrero Simón*
Nombre del consultor/a:	*Dr Helena Brunel Montaner*
Nombre del PRA:	*Dr David Merino Arranz*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estudios genéticos de enfermedades humanas*
Idioma del trabajo:	*Inglés*
Palabras clave	*epigenetics, DNA methylation, depression*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

La depresión es un trastorno mental cada vez más común asociado con déficits sustanciales en la calidad de vida del paciente y un mayor riesgo de mortalidad. Se han realizado varios estudios de asociación genómica (GWAS) para identificar genes asociados con la depresión, pero su heredabilidad parcial entre otras características sugiere la existencia de cambios epigenéticos en el origen de la enfermedad. Dado que se ha demostrado una discrepancia entre la cognición objetiva y subjetiva en pacientes con trastorno depresivo mayor (MDD), es necesario encontrar un sistema para detectar la enfermedad desde una perspectiva biológica.

Utilizando una cohorte basada en una comunidad canadiense (n = 94) que contiene datos de metilación del ADN, estratificada para el estado socioeconómico temprano y evaluada para síntomas depresivos con la escala para Depresión del Centro para Estudios Epidemiológicos (CES-D), se realizó un análisis de metilación diferencial. A partir de este análisis, se identificaron 31 dinucleótidos citosina-guanina (CpG) como metilados diferencialmente en pacientes que muestran síntomas depresivos contra pacientes que no muestran esos síntomas. El análisis se realizó separando a los pacientes por género y tomando la variable *edad* como una covariable.

A partir de las variables socioeconómicas y biomoleculares, y los sitios de CpG identificados, se desarrolló un clasificador Random Forest con el objetivo de crear una herramienta de predicción de síntomas depresivos. El algoritmo resultante tiene una precisión del 73,74% (validación 15 veces cruzada con 3 repeticiones).

La aplicación web *Desypre* (http://desypre.000webhostapp.com/) se creó para

permitir el uso público del clasificador.

Abstract (in English, 250 words or less):

Depression is an increasingly common mental disorder associated with substantial deficits in the quality of life of the patient and increased mortality risk. Several Genomic-Wide Association Studies (GWAS) have been performed to identify genes associated with depression, but its partial heritability among other characteristics suggests the involvement of epigenetic changes in the origin of the disease. Since has been proven a discrepancy between objective and subjective cognition in Major Depressive Disorder (MDD) patients, it is necessary to find a system to detect the disease from a biological perspective.

Using a Canadian community-based cohort (n=94) containing DNA methylation data, stratified for early-life socioeconomic status, and assessed for depressive symptoms with the Center for Epidemiologic Studies Depression (CES-D) scale, a differential methylation analysis was performed. From this analysis, 31 cytosine guanine dinucleotides (CpG) were identified as differentially methylated in patients showing depressive symptoms from patients not showing those symptoms. The analysis was performed separating patients by gender and taking the variable *age* as a covariate.

From the socioeconomic and biomolecular variables, and identified CpG sites, a random forest classifier was developed to create a depressive symptoms prediction tool. The resulting algorithm has an accuracy of 73.74% (repeated 15-fold cross-validation, with 3 repeats).

The web application *Desypre* (http://desypre.000webhostapp.com/) was created to allow the public use of the classifier.

Acknowledgements

Elaborating this project has not been easy, but fortunately I have had awesome people helping me to be successful.

First of all, I have been really lucky with my consultant Dr Helena Brunel. Since our first contact, she has been fast and extensive in her answers to all my questions, helped me to find the perfect subject and guided me along the way.

Next to it, it would have never been possible to have the opportunity to dedicate all my energies to the project without all the emotional support from my life partner Maarten Hakvoort, and doing all the household chores to keep me allergy-free and fed these last few months. Thank you for believing in me and supporting me in all my decisions.

And finally, a sweet mention to my naughty cat Siurana, who came a few times to lie down on my laptop keyboard while I was working in this project, because she prefers when I cuddle her than when I'm working.

Index

List of figures

List of tables

1. Introduction

1.1 Context and justification of the project

1.1.1 Depression

Depression is an increasingly common mental disorder, with more than 320 million patients of all ages in the world by 2015 [1]. The total estimated number of people living with the disorder increased by 18.4% between 2005 and 2015 [2].

MDD is associated with substantial deficits in the quality of life of the patient [3], suffering a combination of several symptoms (Table 1). Besides, it has a major impact on mortality risk: depressed men and women are 20.9 and 27 times, respectively, more likely to commit suicide than the general population. But it also increases the risk of other kinds of decease, such as cardiovascular death. Because of being a disorder that profoundly affects different levels in the life of the patient (functional, social, economic), it is predicted to become a major health burden worldwide [4].

Major depressive disorder
5 or more of the following symptoms (including at least 1 of either depressed mood or loss of interest) in the same 2-week period • Depressed mood or persistent sadness • Loss of interest or pleasure in activities • Change in weight or appetite • Insomnia or hypersomnia • Psychomotor retardation or agitation • Loss of energy or fatigue • Feelings of worthlessness or guilt • Impaired concentration or indecisiveness • Thoughts of death or suicidal ideation or suicide attempt

Table 1. American Psychiatric Association Diagnostic and Statistical Manual (DSM-V) criteria for major depressive disorder [5].

1.1.2 Epigenetics in depression

MDD exhibits numerous non-Mendelian features that can be reviewed from an epigenetic perspective, such as partial heritability (approximately 37%), the discordance of monozygotic (MZ) twins and sexual dimorphism with more incidence in women than men [6]. Moreover, it has been suggested that environmental factors, e.g. low socioeconomic status (SES) in early life, trigger epigenetic changes that may increase the risk for future depression [7].

What are those epigenetic changes?

The most accessible and characterized form which epigenetic information in humans can be transmitted is via DNA methylation (DNAm, Figure 1) [8]. In somatic cells, DNAm occurs almost exclusively on cytosine residues in the context of CpGs. There are about 28 million CpG motifs non-randomly distributed across the haploid human genome [9]. The promoter regions of 60-70% of all human genes contain CpG islands, which are areas with a higher concentration of CpG motifs, and their methylation regulates the gene

expression [10].

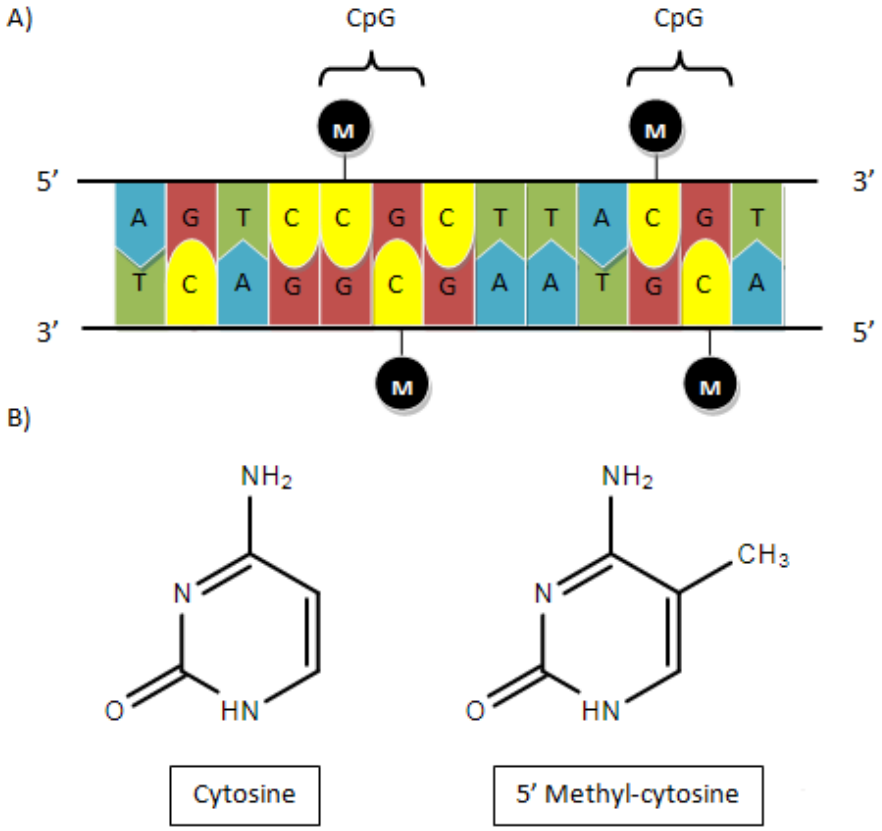

Figure 1: A) Fragment of DNA with methylated cytosines in the CpG sites. B) Left: non-methylated cytosine. Right: methylated cytosine.

State of the art

Numerous studies have found an association between DNA methylation modifications and depression. Epigenome-wide Association Studies (EWAS) located numerous CpG sites presenting significantly differentiated methylation in depressed patients, including hyper- and hypomethylations. BDNF, NR3C1, SLC6A4 and OXTR genes were the most frequently studied genes. Mostly, promoter regions and CpG islands were commonly targeted [11]. In addition, it has been suggested that many of the methylation differences in patients showing depressive symptoms are specially annotated to genes linked to a G-protein coupled receptor protein signalling pathway [12].

1.1.3 Problem to solve

DNAm in MDD is associated with socioeconomic factors but has not yet been used to discriminate depressive individuals from individuals without depressive symptoms. Since has been proven a discrepancy between objective and subjective cognition in MDD patients [13], it is necessary to find a system to detect the disease from a biological perspective. This project intends to create an algorithm-based tool to assess whether patients show depressive symptoms.

1.2 Objectives

1.2.1 General objective

Design, development and implementation of an algorithm to assess whether patients show depressive symptoms based on socioeconomic data and DNA methylation signatures in depression found in literature.

1.2.2 Specific objectives

1. To describe the subject and the state of the art.
2. To extract and prepare the data from a published cohort that contains information in socioeconomic features and DNA methylation profiles.
3. To perform a descriptive statistical analysis of the data.
4. To find differentially methylated CpGs between samples presenting depressive symptoms versus samples not presenting those symptoms.
5. To design, to develop and to validate the algorithm.
6. To design a website where external data can be analyzed by the algorithm.

1.3 Approach and follow-up method

A possible approach in the development of this Project would be to take samples and perform socioeconomic and depression tests to a large random sample of the general population. Due to the lack of resources, an easier approach will be to find a published cohort that contains information in socioeconomic features (including the score of a validated depression scale) and methylation profiles.

Moreover, after the analysis of the state of the art and the descriptive statistics of the data, a specific methodology will be selected, including an appropriate algorithm type.

1.4 Work planning

1.4.1 Necessary resources

For the development of this project it is necessary to have a computer with internet connection and the following software installed:

- Statistical software R, with the packages mentioned in Table 3.
- Word processor.
- Web browser.

1.4.2 Tasks

In Table 2 are shown the planned tasks for each specific objective with their respective expected time period.

Objectives	Tasks	Period
1. To describe the subject and the state of the art.	• Task 1: Analyze bibliography and write a short introduction and summary of the state of the art.	15/10/19 to 22/10/19

2. To extract and prepare the data from a published cohort that contains information in socioeconomic features and DNA methylation profiles.	• Task 2: Extract data from a published cohort.	23/10/19 to 25/10/19
	• Task 3: Pre-process the data.	26/10/19 to 29/10/19
3. To perform a descriptive statistical analysis of the data.	• Task 4: Perform descriptive statistics of the data.	26/10/19 to 29/10/19
4. To find differentially methylated CpGs between samples presenting depressive symptoms versus samples not presenting those symptoms.	• Task 6: Perform differential DNA methylation analysis.	30/10/19 to 02/11/19
5. To design, to develop and to validate the algorithm.	• Task 7: Design and implement the algorithm.	03/11/19 to 18/11/19
	• Task 8: Improve the algorithm.	19/11/19 to 05/12/19
6. To design a web application where external data can be analyzed by the algorithm.	• Task 9: Design the web application.	06/12/19 to 16/12/19

Table 2. Specific objectives of the project with their respective tasks and timing for each task.

The initial work plan had an extra task for specific objective 1: to find the DNA methylation signatures in depression identified in literature. The intention was to develop the algorithm from a database including, for each sample, socioeconomic and biomolecular data together with the DNA methylation values (*M values*)[1] from the found DNA methylation signatures in depression. But the problem was that the resolution of the array used in the methylation data extraction from the selected published cohort is different than other studies. It was the *Illumina Infinium HumanMethylation27 Beadchip v1.2*, which covers 27,578 CpG sites per sample, while the published studies from last years, used *Illumina Infinium HumanMethylation450 Beadchip*, which covers 482,421 CpG sites. For this reason, the majority of the significant found CpG sites were not available in the cohort data.

A possible solution to this problem would have been to find another published cohort, but it was not possible to find another dataset meeting all the requirements to perform the planned algorithm. Thus, the new approach was to add a new task during the first phase of the project development: to perform a differential methylation analysis in the selected dataset to find the significant CpGs differentially methylated between the samples presenting depression

[1] Estimate of the methylation level, calculating the log2 ratio of the intensities of methylated probe versus unmethylated probe [50]).

symptoms versus the samples not presenting those symptoms according to the CES-D scale score. Further information about the CES-D scale is given in the materials and methods section.

1.4.3 Calendar

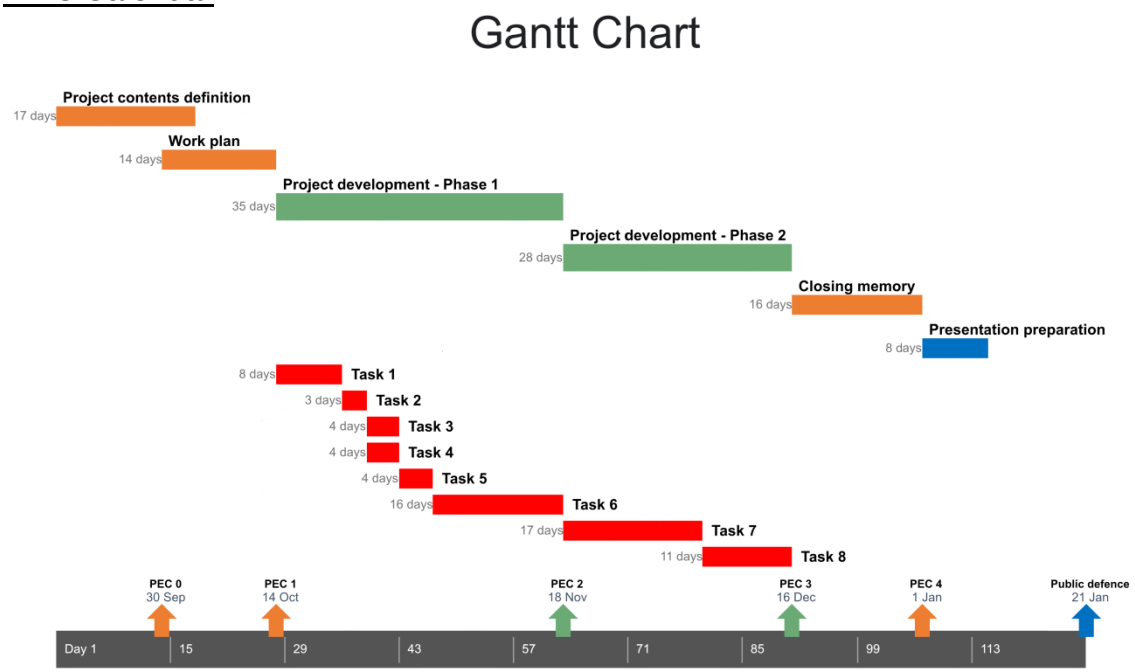

Figure 2. Phases, tasks and milestones of the project collected on a Gantt Chart.

1.4.4 Milestones

PEC 0:

- Project contents report

PEC 1:

- Work plan

PEC 2:

- Summary of the subject and state of the art
- Descriptive statistics report
- Differential DNA methylation analysis
- First version of the algorithm

PEC 3:

- Last and improved version of the algorithm
- Web application

PEC 4:

- Final report

1.5 Brief summary of the obtained products

1. Work plan
2. Algorithm to assess whether patients show depressive symptoms
3. Web application to perform the algorithm
4. Final report
5. Virtual presentation
6. Auto evaluation

1.6 Brief description of the other chapters of the report

Other chapters in the memory have the following contents:

- Chapter 2: Materials and methods
  In this section, the selected published dataset and the computational tools used in this project are detailed. Also, the necessary methodology to fulfil the goals of the project and the followed pipeline is described.

- Chapter 3: Results
  Here, the methodology outcomes are described and presented: statistical analysis of the dataset, differential DNA methylation analysis, machine learning algorithm and web application.

- Chapter 4: Discussion
  An interpretation of the results is presented, highlighting the most important findings and a comparison of those with literature knowledge.

- Chapter 5: Conclusions
  A small summary of the project outcome, collecting the most important facts extracted from the results and discussion.

2. Materials and Methods

2.1 Software

The software used to analyze the data is the statistical software R (R-Studio interface) in its latest version at the time of this project, which is 3.5.3.

The packages used are presented in the following table:

Package	Description	Reference
Biobase	This package contains standardized data structures to represent genomic data that are used by other R packages.	[14]
caret	The caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models.	[15]
GEOquery	GEOquery is the bridge between Bioconductor and the public repository of microarray data Gene Expression Omnibus (GEO).	[16]
ggplot2	ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics.	[17]
glmnet	Efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression.	[18]
grid	A rewrite of the graphics layout capabilities, plus some support for interaction.	[19]
gridExtra	Provides a number of user-level functions to work with "grid" graphics, notably to arrange multiple grid-based plots on a page, and draw tables.	[20]
limma	Data analysis, linear models and differential expression for microarray data.	[21]
MASS	Functions and datasets to support Venables and Ripley "Modern Applied Statistics with S" (4th edition, 2002).	[22]
moments	Functions to calculate: moments, Pearson's kurtosis, Geary's kurtosis and skewness.	[23]
randomForest	Classification and regression based on a forest of trees using random inputs.	[24]
reshape2	Facilitates to transform data between wide and long formats.	[25]

Table 3: R packages used during the performance of the project.

2.2 Data

The applied dataset was extracted from a published study [26] and was uploaded at the GEO database [27], an international public repository that archives and distributes functional genomics datasets free of charge. The dataset is registered under the access number: **GSE37008**.

In the original study, the data was collected from a community-based cohort stratified for early-life SES. The genomic DNA of 94 individuals was extracted from periferial blood mononuclear cells, bisulphite converted and hybridized, along with 5 technical replicates to the Illumina Infinium HumanMethylation27 Beadchip v1.2 for genome wide DNA methylation profiling. This array measures the DNA methylation in the promoter regions of more than 14.000 human genes.

The existence of depressive symptoms in the individuals of the cohort is screened by the CES-D scale. This scale measures symptoms defined by the DSM-V for a major depressive episode (Table 1). The possible range of scores is 0-60, with the higher scores indicating the presence of more symptomatology [28]. The common cut-off to assess the existence of depressive symptoms is 16.

2.3 Algorithm

Random Forest (RF) algorithm was used to build the model. RF is an ensemble tool based on decision trees which allows to obtain a fast and reproducible classification predictor [24]. A defined number of trees using selected features (in this case, the phenotype variables and the selected CpG markers) lead to the prediction of the existence of depression symptoms based on the average estimation of each tree. Random sub-setting of samples within the training set for construction of each tree avoids overfitting to the training set samples. Here the RF algorithm was implemented with the RandomForest package (v.4.6-14).

2.4 Web development

The web application was written in HTML, PHP and CSS code and hosted at 000webhost web server.

2.5 Pipeline

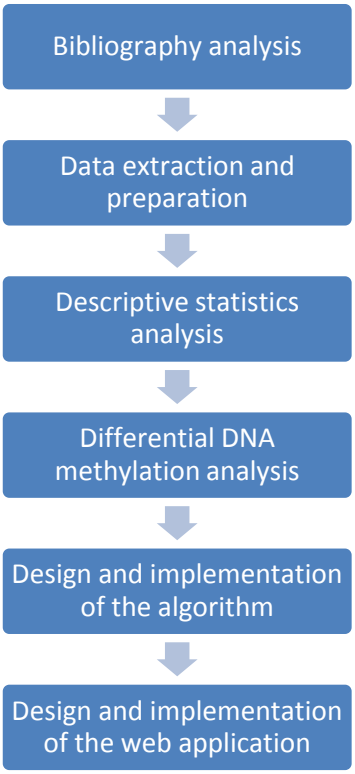

Figure 3. Workflow of the project.

3. Results

3.1 Descriptive statistics

3.1.1 Extraction of the data

The data was downloaded with the function *getGEO()* from the package *GEOquery,* using the GEO accession number (GSE37008). The *M values* were already normalized. The phenotype data was extracted as a data frame with the function *pData()* from the package *Biobase*.

3.1.2 Data analysis

The phenotype dataset contains a total of 99 observations (samples) and 39 variables (features). The variables can be divided in socioeconomic (Table 4) and biomolecular (Table 5).

Socioeconomic variables	Unit
age	years
gender	Female / Male
smoker	Yes / No
SES in early life	High / Low
SES at present time	High / Low
alcohol	drinks per week
body mass index (BMI)	kg/m^2
exercise	minutes per week
stress	Perceived Stress Scale questionnaire
depression	Center for Epidemiologic Studies Depression Scale
sleep	perceived quality; Pittsburgh Sleep Quality Index
birth control	current use of oral contraceptives
ethnicity	Caucasian or other ethnicity
pbifw	measures of recalled paternal warmth based on the questionnaire Parental Bonding Index
pbimw	measures of recalled maternal warmth based on the questionnaire Parental Bonding Index

Table 4. Socioeconomic variables of the phenotype data from the GSE37008 dataset.

Biomolecular variables	Unit
total cortisol	total cortisol release over 3 days in arbitrary units
dslope	diurnal rhythm of cortisol in arbitrary units
pam3csk4	Interleukin-6 (IL-6) response to pam3csk4 in pg/ml
pgn	IL-6 response to peptidoglycan in pg/ml
pic	IL-6 response to ppoly i:c in pg/ml
lps	IL-6 response to lipopolysaccharide in pg/ml
flagellin	IL-6 response to flagelin in pg/ml

zymosan	IL-6 response to zymosan in pg/ml
ssr	IL-6 response to single-stranded RNA in pg/ml
imiquimod	IL-6 response to imiquimod in pg/ml
odn	IL-6 response to oligodeoxyribonucleotide in pg/ml
il1b	IL-6 response to interleukin-1b in pg/ml
pma	IL-6 response to phorbol-12-myristate-13-acetatein in pg/ml
wbc count	total white blood cells x 10^9 / liter blood
neutrophils	neutrophils x 10^9 / liter blood
lymphocytes	lymphocytes x 10^9 / liter blood
monocytes	monocytes x 10^9 / liter blood
basophils	basophils x 10^9 / liter blood
eosinophils	eosinophils x 10^9 / liter blood
neutrophil%	percent of wbc that are neutrophils
lymphocyte%	percent of wbc that are lymphocytes
monocyte%	percent of wbc that are monocytes
basophil%	percent of wbc that are basophils
eosinophil%	percent of wbc that are eosinophils

Table 5. Biomolecular variables of the phenotype data from the GSE37008 dataset.

The biomolecular variables *total.cortisol* and *dslope* were discarded due to a high amount of missing values, which reduced the number of incomplete samples from 27 to 9. The remaining 9 samples with missing values (GSM908398, GSM908418, GSM908419, GSM908429, GSM908430, GSM908446, GSM908455, GSM908456 and GSM908474), together with the 4 samples not containing methylation data (GSM908432, GSM908438, GSM908440 and GSM908461), were also discarded. The final dimensions of the phenotype dataset were 86 observations and 37 variables.

3.1.3 Graphic analysis

The variables were divided in 4 groups to perform their individual graphic representation: quantitative socioeconomic variables (Figure 4), categorical socioeconomic variables (Figure 5), quantitative biomolecular variables (Figure 6) and categorical biomolecular variables (Figure 7). In the first group, the predictive variables showing a correlation with the response variable depression were stress (p-val = $8.12 \cdot 10^{-20}$) and pbifw (p-val = 0.0194). In the second group, the biggest correlations are for the variables smoker, current SES and sleep. In the biomolecular variables group, the correlated predictive variables with the response variable were pam3csk4 (p-val = 0.0217) and *odn* (p-val = 0.0301).

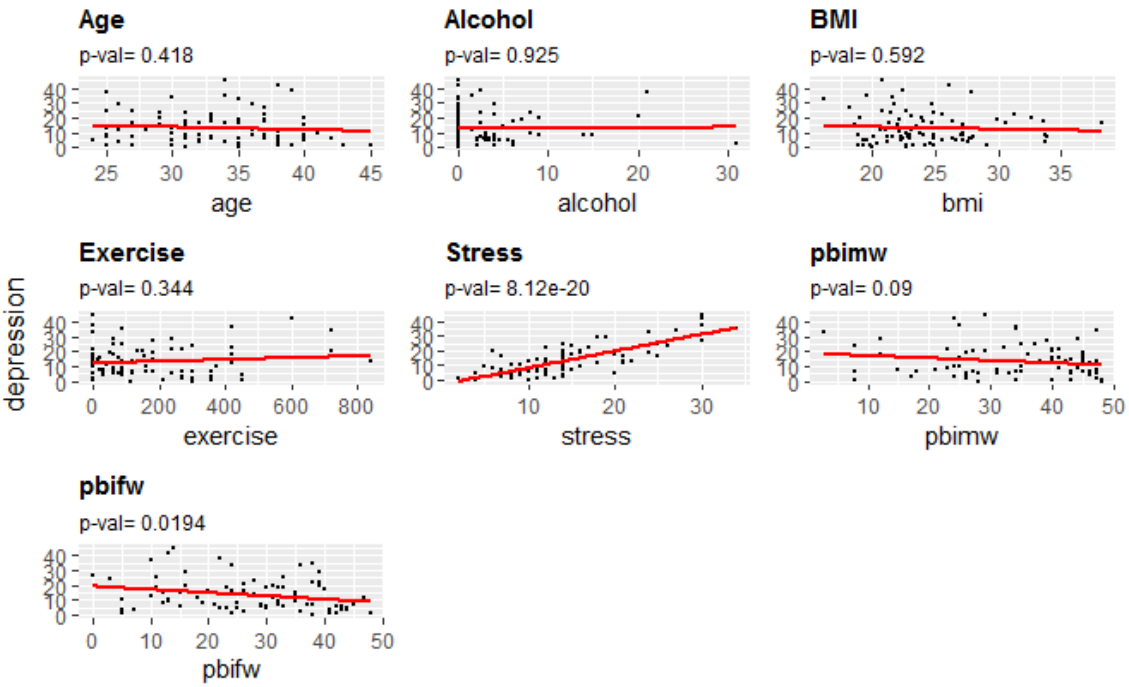

Figure 4. Linear regression plots of quantitative socioeconomic predictor variables against response variable depression.

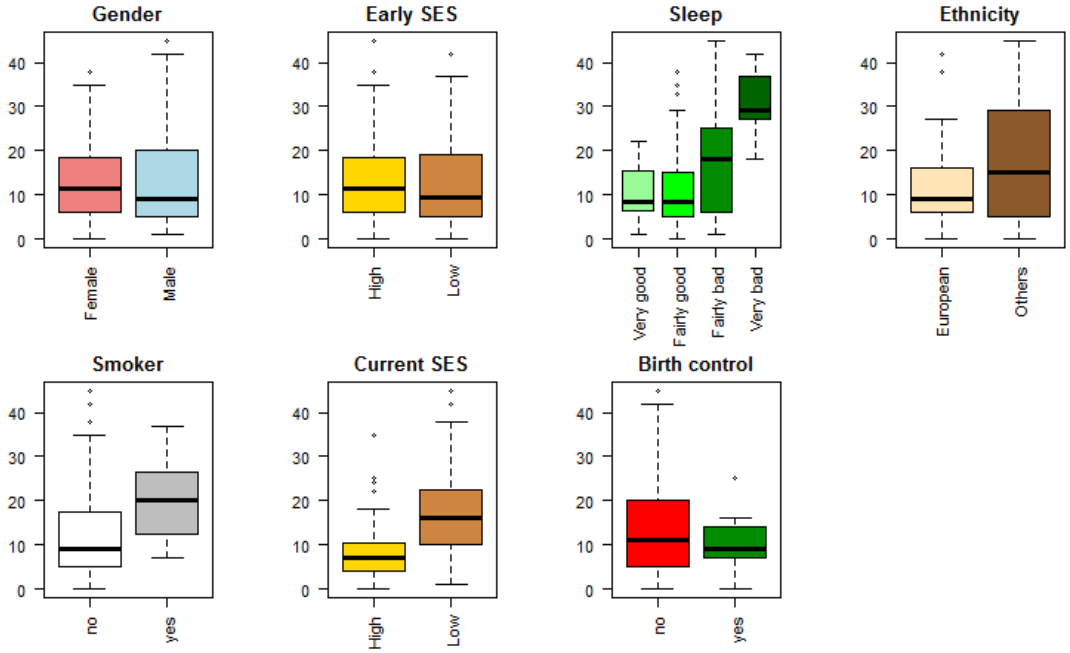

Figure 5. Boxplots of categorical socioeconomic predictor variables against response variable depression.

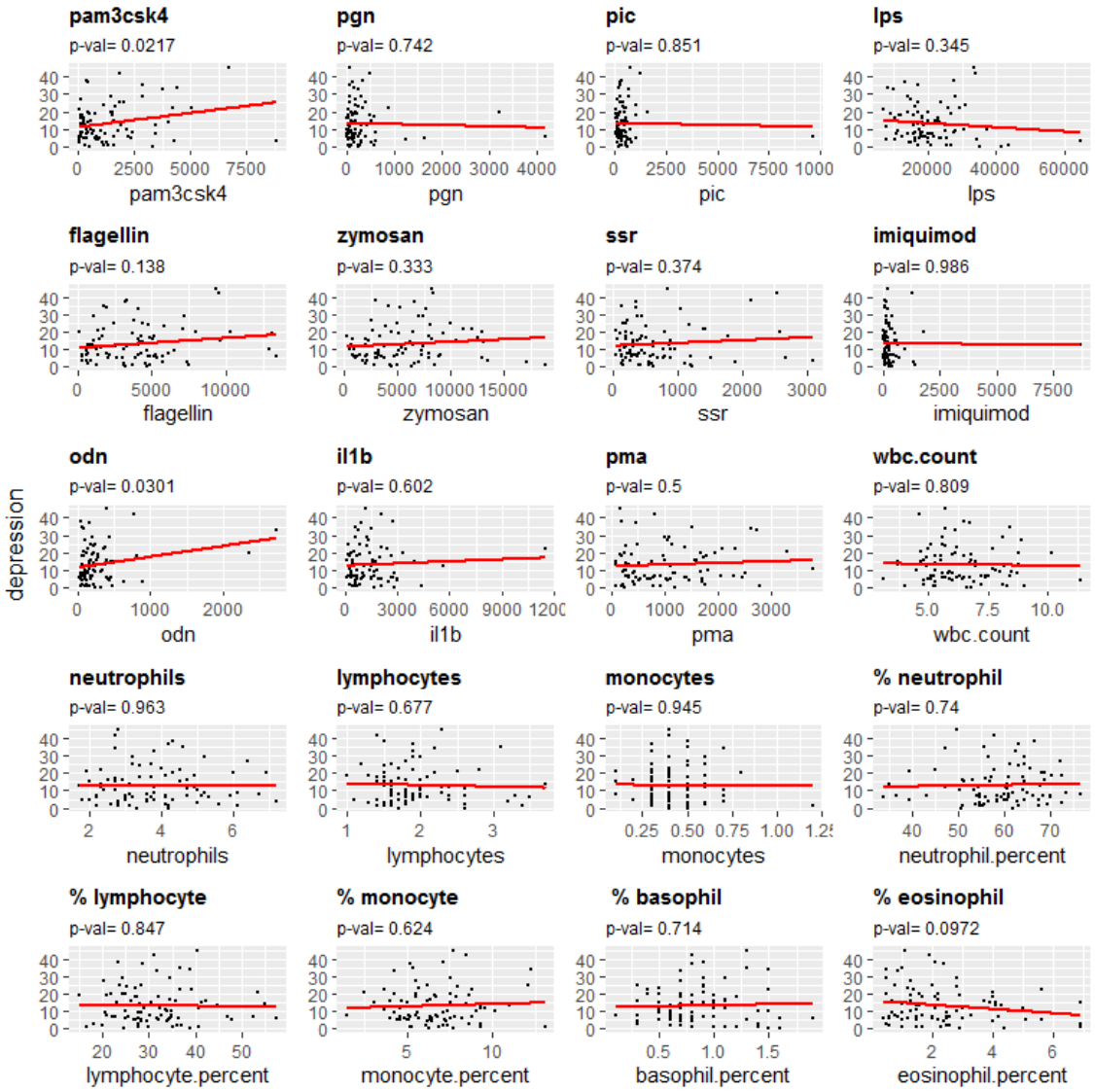

Figure 6. Linear regression plots of biomolecular predictor variables against response variable depression.

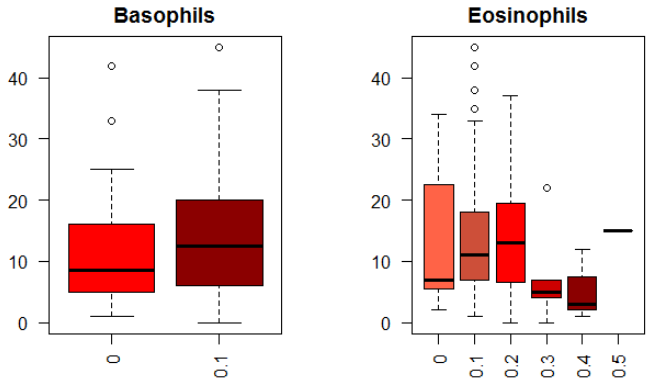

Figure 7. Boxplots of categorical biomolecular predictor variables against response variable depression.

3.1.4 Analysis of the response variable *depression*

The response variable is a quantitative variable representing the CES-D score. In this dataset it ranges from 0 to 45 points, with an average of 13.17 and a median of 10.00. Since average and median are different, it was necessary to study this variable to assess its normality.

Figure 9-A shows the density of the response variable, and it seems that the distribution is clearly asymmetric, with a skewness of 1.077. To correct the asymmetry, the Box-Cox correction method was applied with the package *MASS*. It was necessary to add a small constant ($1 \cdot 10^{-5}$) to the response variable before calculating the lambda value due to the presence of zeros in the variable.

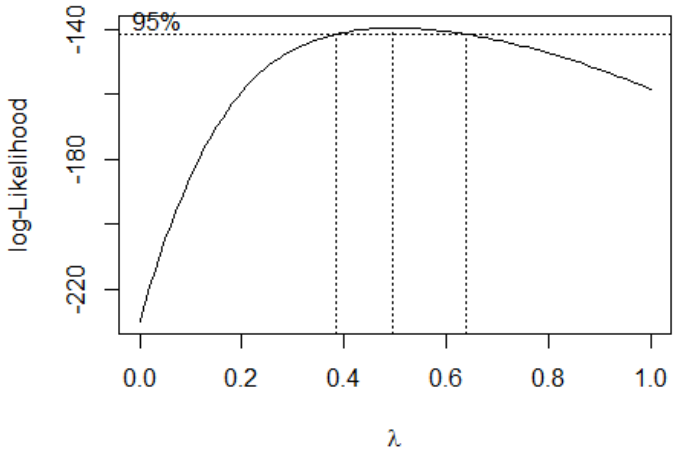

Figure 8. Log-likelihood plot for Box-Cox response variable depression transformation.

The variable was corrected with the parameter λ=0.5 (Figure 8).

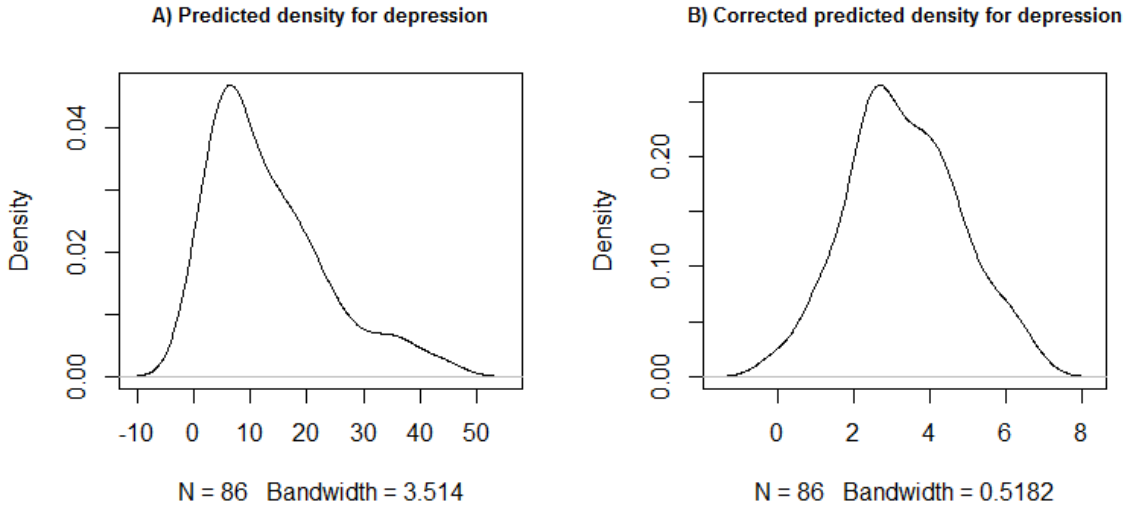

Figure 9. Comparision between response variable depression before (A) and after (B) Box-Cox transformation.

After the application of the Box-Cox correction, the skewness decreased to 0.126.

3.1.5 LASSO linear regression to identify most significant phenotype variables

Although all the socioeconomic and biomolecular features will be used to the development of the algorithm, a LASSO model was constructed to pinpoint the significant variables. First of all, model matrix of the predictor variables was created. The function *gv.glmnet()* was used to find the best lambda for the model, identifying λ=0.2. The function *glmnet()* was then used to build the model. The predictive variables selected as significant by the LASSO model were: *smoker*, *current SES*, *stress* and *lps*.

3.2 Differential DNA methylation analysis

For this analysis the Linear Models for Microarrays method, implemented in the *limma* package [21] is used to select differentially methylated CpGs.

3.2.1 Data preparation

The methylation values were extracted with the function *exprs()* from the package *Biobase*. The dataset contains a total of 27,578 observations (*M value*s) and 99 variables (samples). The incomplete samples mentioned in the previous section (9 samples with missing values within the phenotype data and 4 samples without methylation data) were removed. Afterwards, there were identified empty 4,656 observations, which were as well eliminated. The final dimensions of the methylation dataset were 22,578 observations and 86 variables.

The samples were divided in four groups: *control female* (Group 0) for females with a CES-D score under 16, *control male* (Group 1) for males with a CES-D score under 16, *depressed male* (Group 2) for males with a CES-D score equal or above 16, and *depressed female* (Group 3) for females with a CES-D equal or above 16.

3.2.2 Principal Components Analysis (PCA)

In order to do a first exploratory analysis of the data, a PCA was performed.

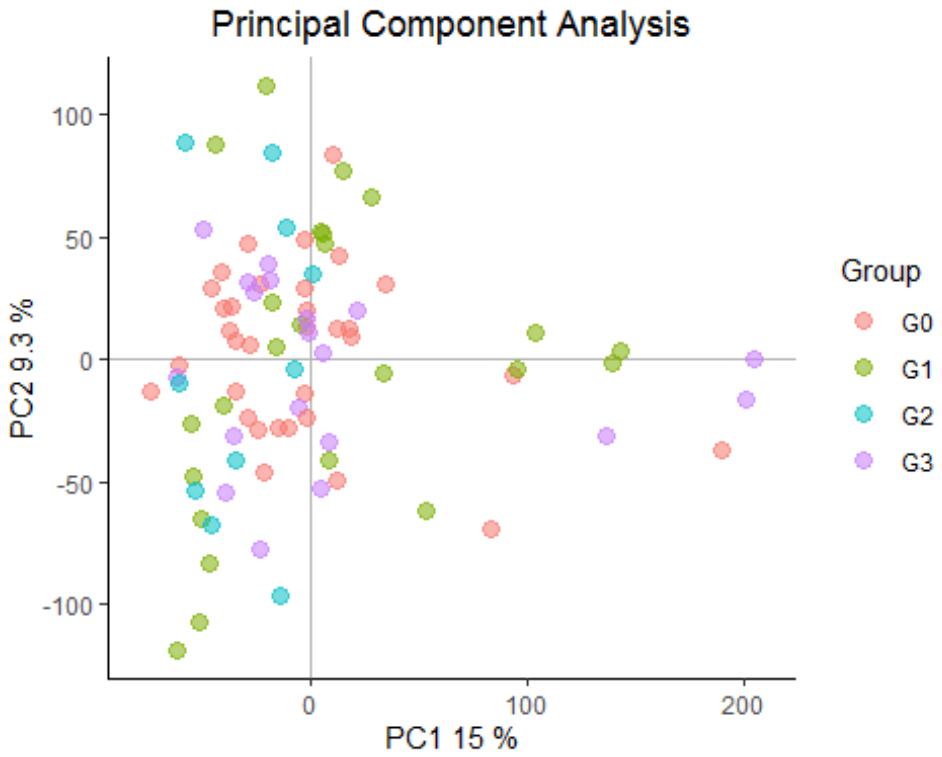

Figure 10. PCA plot for the first two principal components (PC).

The conclusion was that the accountability for the total variability of the samples was widely distributed. First component of the PCA accounts for 15% of the total variability, and second component accounts for a 9.3%. For this reason, at the plot of the first two components it was not possible to observe big differences between groups.

3.2.3 Design matrix

A design matrix was created, in order to describe the allocation of each sample to its group. The variable age was added as a covariate to the model due to the known chronological age-dependent changes in methylation [29]. Table 6 shows the first 10 observations of the design matrix for this analysis.

	G0	G1	G2	G3	age
GSM908383	1	0	0	0	38
GSM908384	0	1	0	0	26
GSM908385	1	0	0	0	33
GSM908386	0	1	0	0	38
GSM908387	0	1	0	0	37
GSM908388	0	0	0	1	31
GSM908389	1	0	0	0	38
GSM908390	0	0	0	1	37
GSM908391	0	1	0	0	31
GSM908392	1	0	0	0	30

Table 6. First 10 samples of the design matrix for the differential methylation analysis.

3.2.4 Contrasts Matrix

The comparisons between groups were defined by the contrasts matrix (Table 7). In this case, the goal was to check the effect of depression ("depressed vs control") separately for both genders.

	G3 - G0	G2 - G1	G1 - G0	G3 - G2
G0	-1	0	-1	0
G1	0	-1	1	0
G2	0	1	0	-1
G3	1	0	0	1
age	0	0	0	0

Table 7. Contrasts matrix for the differential methylation analysis.

The contrast matrix was defined to perform four comparisons: effect of depression in females (G3 - G0), effect of depression in males (G2 - G1), effect of gender in controls (G1 - G0) and effect of gender in depressed (G3 - G2).

3.2.5 Model estimation and CpG selection

Once the model and the contrasts were estimated, and the significance tests were performed, a list of the differentially methylated CpG motifs was obtained. The list was ordered from smallest to biggest p-value, and the observations with the p-value over 0.05 were eliminated. The total number of selected CpG was 31, which were annotated with the available data in the GEO *ExpressionSet* applying the function *fData()* to the initial ExpressionSet object (Table 8).

ID	adj.P.Val	Chr	Gene_Strand	Gene_ID	Symbol
cg23758485	5.53E-15	16	-	GeneID:55512	SMPD3
cg18485485	5.53E-15	8	+	GeneID:1666	DECR1
cg16718678	5.53E-15	11	-	GeneID:9379	NRXN2
cg09067967	8.57E-09	4	-	GeneID:7358	UGDH
cg13913917	0.000118	3	+	GeneID:23200	ATP11B
cg02058918	0.0009	9	-	GeneID:9568	GABBR2
cg20050113	0.001585	2	+	GeneID:6549	SLC9A2
cg00798206	0.001585	3	+	GeneID:9943	OXSR1
cg14912575	0.001953	14	-	GeneID:56936	C14orf162
cg08590939	0.003394	5	-	GeneID:5019	OXCT1
cg23552468	0.003714	9	+	GeneID:138724	MGC41945
cg14464791	0.005554	11	-	GeneID:1656	DDX6
cg00234616	0.006183	2	+	GeneID:3196	TLX2
cg18329349	0.008854	7	+	GeneID:857	CAV1
cg22584138	0.017207	17	-	GeneID:6532	SLC6A4
cg08532057	0.017207	13	+	GeneID:9818	NUPL1
cg10164137	0.017928	1	-	GeneID:1855	DVL1
cg08124399	0.022535	6	+	GeneID:55510	DDX43
cg08085267	0.022535	17	+	GeneID:124989	C17orf57
cg08832227	0.022625	12	+	GeneID:3736	KCNA1
cg09354267	0.025742	6	-	GeneID:9474	ATG5
cg11787522	0.025742	15	-	GeneID:64220	STRA6

cg25664034	0.027517	7	+	GeneID:2318	FLNC
cg22631938	0.041012	3	+	GeneID:2850	GPR27
cg03311036	0.041012	7	+	GeneID:835	CASP2
cg18022322	0.042035	14	+	GeneID:22990	PCNX
cg00425710	0.042035	16	+	GeneID:4489	MT1A
cg15446391	0.04664	11	-	GeneID:7490	WT1
cg09886641	0.04664	15	+	GeneID:246777	SPESP1
cg14576824	0.04664	1	+	GeneID:26750	RPS6KC1
cg27167601	0.048736	15	-	GeneID:6095	RORA

Table 8. List of the differentially methylated CpG sites in depression with a p-value under 0.05.

3.2.6 Graphic representation of the differentially methylated CpG sites

In order to visualize the *M values* of the CpG sites which methylation levels are significantly differentiated in depression, they were plotted (Figure 10).

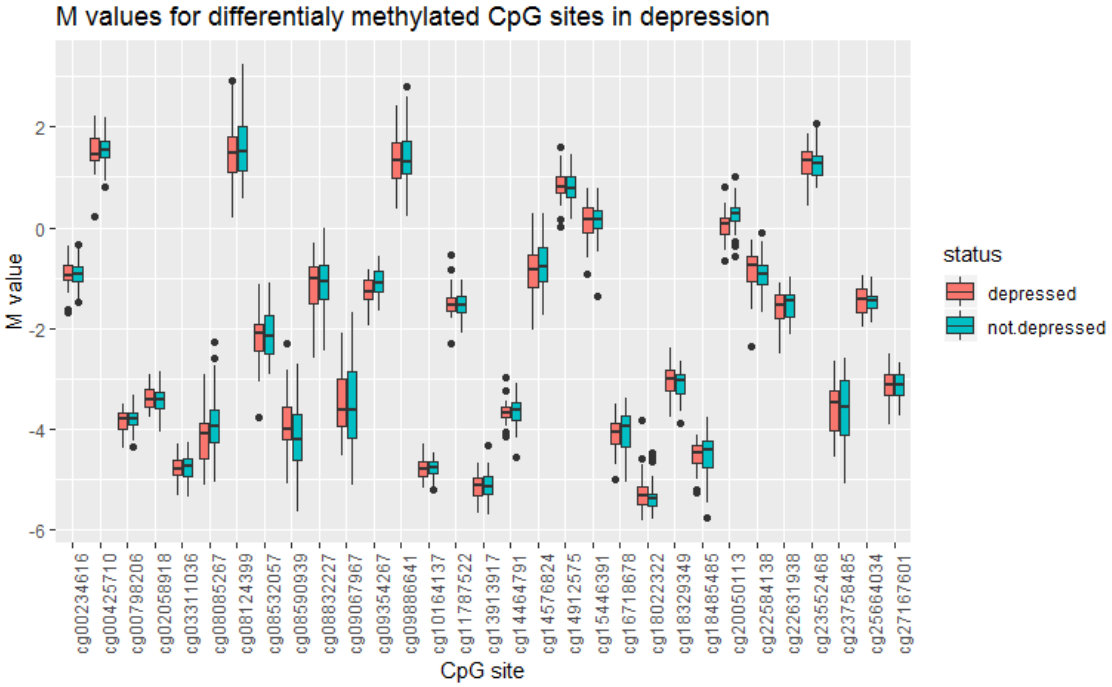

Figure 11. M values of the selected 31 CpG sites with a p-value < 0.05 for the differential analysis, differentiating between depressed and not depressed samples.

3.3 Machine learning algorithm

3.3.1 Initial model

A new dataset was created, joining for each sample the phenotype variables with the *M values* for the differentially methylated CpG sites. It contained a total of 86 observations (samples) and 68 variables (15 socioeconomic, 22 biomolecular and 31 CpG *M values*). The variable depression containing the CES-D scores was substituted by a categorical variable with the group classification, and the categorical variables (*smoker*, *current SES*, *birth control* and *ethnicity*) were changed to binary factors.

The dataset was randomly divided into a train dataset with 57 samples (67%) which was used to train the model, and a test dataset with 29 samples (33%) which was used to evaluate the model.

Model training with the data

The model was trained using the function *randomForest()* from the package with the same name. For the initial model, the parameters *ntree* (number of trees produced by the algorithm) and *mtry* (number of randomly selected variables as candidates in each division) were the function default *ntree* = 500, and *mtry* = 8. Figure 12 shows the plot of the obtained model.

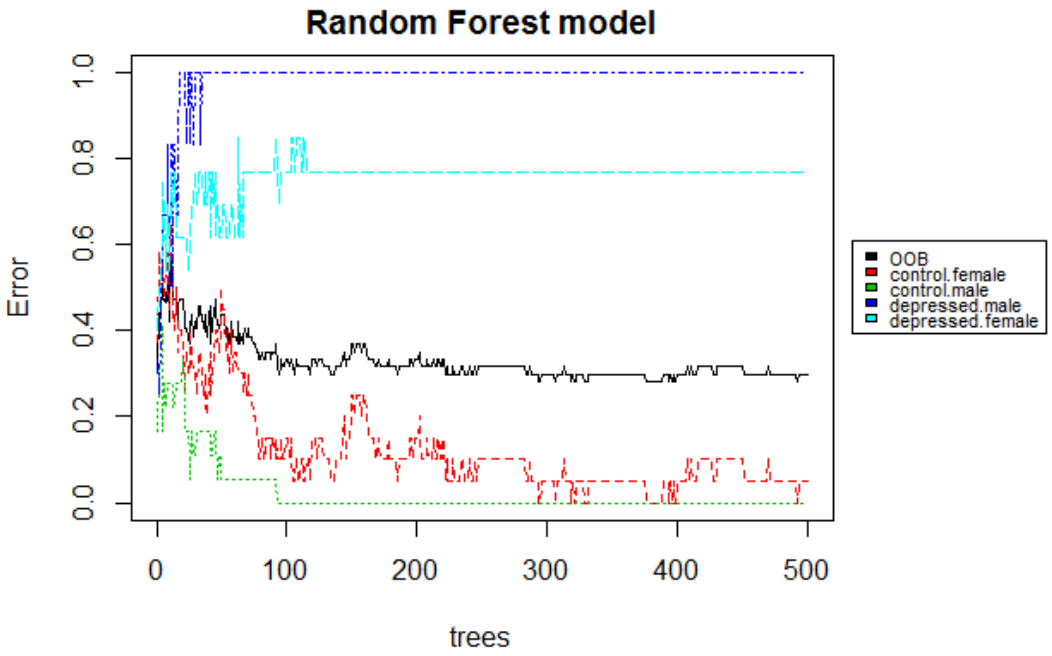

Figure 12. Error evolution of the model created by Random Forest algorithm for ntree=500 and mtry=8.

Model performance evaluation

With the aim of evaluating the performance of the model, the test dataset samples were classified in one of the four groups using the function *predict()*. The predictions were compared to the real labels of each sample through a confusion matrix (Table 9) using the function *confusionMatrix()* from the *caret* package.

		Reference			
		control female	control male	depressed male	depressed female
Prediction	control female	13	1	0	1
	control male	0	5	3	0
	depressed male	0	0	1	0
	depressed female	0	0	0	5

Table 9. Confusion matrix of the predicted classification versus the real labels of the test samples.

For the group *control female*, all the 13 samples were classified correctly. For the group *control male*, five samples were classified correctly but one of the samples was wrongly identified as control female. This means that there were no false positives regarding the disease condition, only one wrongly identified gender. On the other hand, for the group *depressed male*, only one sample was classified correctly, and three samples were classified as *control male* (false negative). For the group *depressed female*, five samples were classified correctly and one sample was wrongly classified as *control female* (false negative). Thus, the model has a very good specificity but an insufficient sensitivity differentiating patients showing depressive symptoms.

Statistics:
The model had an **accuracy of 82.76%**, with a confidence interval at 95% between 64.23 and 94.15% (p-value = $3.119 \cdot 10^{-5}$) and Kappa = 74.29%, which indicates on average a good agreement between the model predictions and the true values.

	Class			
	control female	control male	depressed male	depressed female
Sensitivity	1.0000	0.8333	0.25000	0.8333
Specificity	0.8750	0.8696	1.00000	1.0000

Table 10. Sensitivity and specificity of the model by class.

3.4 Algorithm tuning

Once the model is built, it is necessary to tune the algorithm in order to achieve the best performance for the specific model but avoiding overfitting. There are two main parameters to be tuned in the RF algorithm: the *mtry* parameter, which is the number of variables randomly sampled as candidates at each split and the *ntree* parameter, which is the number of trees to create. In this project, different tuning methods involving the adjustment of the mentioned parameters were tried [30].

3.4.1 Tuning *mtry* while holding *ntree* constant

The *caret* package provides the function *train()* for algorithm tuning, although only *mtry* parameter adjustment is available for the RF algorithm. The *ntree*

value will stay constant at its default *ntree* = 500. Both applied search methods to find the best value for *mtry* with caret were adjusted to do a 15-fold cross-validation and 3 repeats to limit and reduce overfitting on the training set. The accuracy value was used to select the optimal model.

Random search

For the random search, performing a total of 30 trials, the best *mtry* value was 29, with an accuracy of 72.64% and Kappa = 59.36%. Figure 13 shows a plot of the different obtained accuracies for the randomly used *mtry* values.

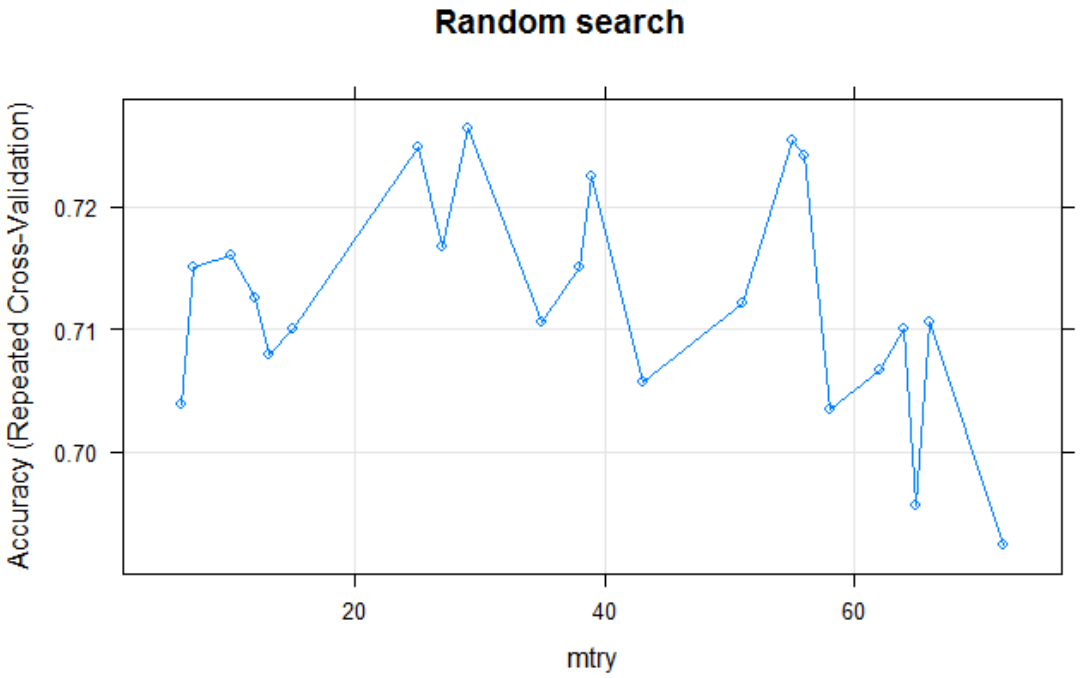

Figure 13. Results for the RF tuning with random search from caret package.

Grid search

For the grid search, trying a total of 30 *mtry* values (from 1 to 30), the best *mtry* was 24 with an accuracy of 73.97% and Kappa = 61.28%. Figure 14 shows a plot of the different obtained accuracies for the 30 tried *mtry* values.

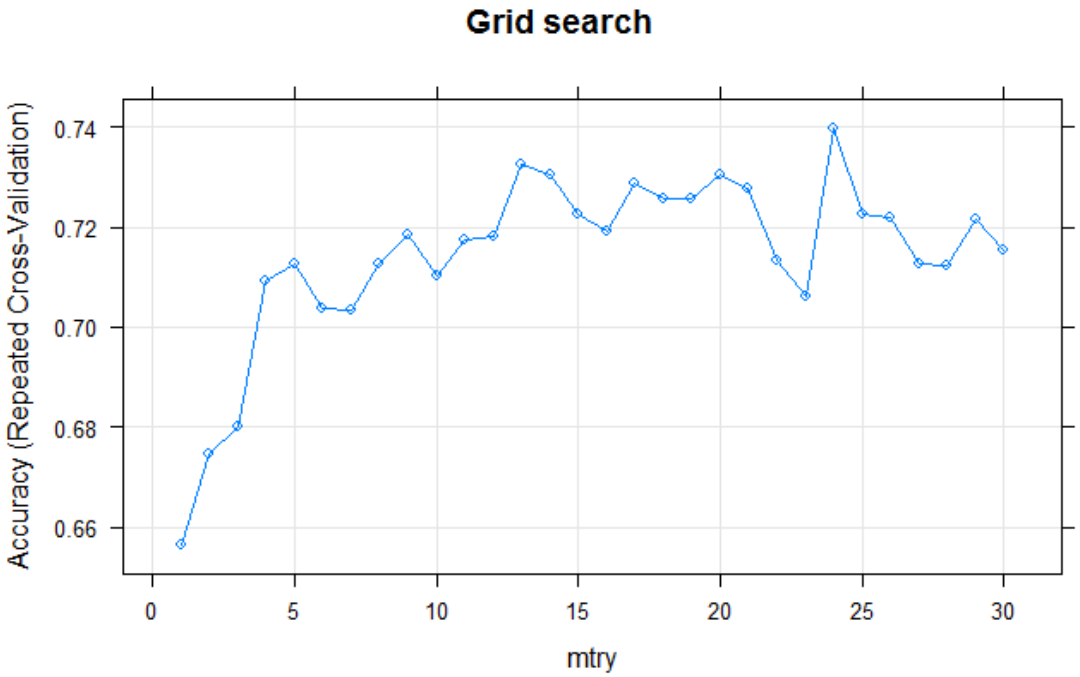

Figure 14. Results for the RF tuning with grid search from caret package.

3.4.2 Tuning *ntree* while holding *mtry* constant

To overcome the disadvantage of the caret package regarding the *ntree* parameter tuning, several caret models were created and compared. For this approach, the *mtry* parameter is maintained constant at the square root of the number of variables. In this case, *mtry* = 8. A total of 6 models with different values for *ntree* were analyzed (500, 1000, 1500, 2000, 2500 and 3000).

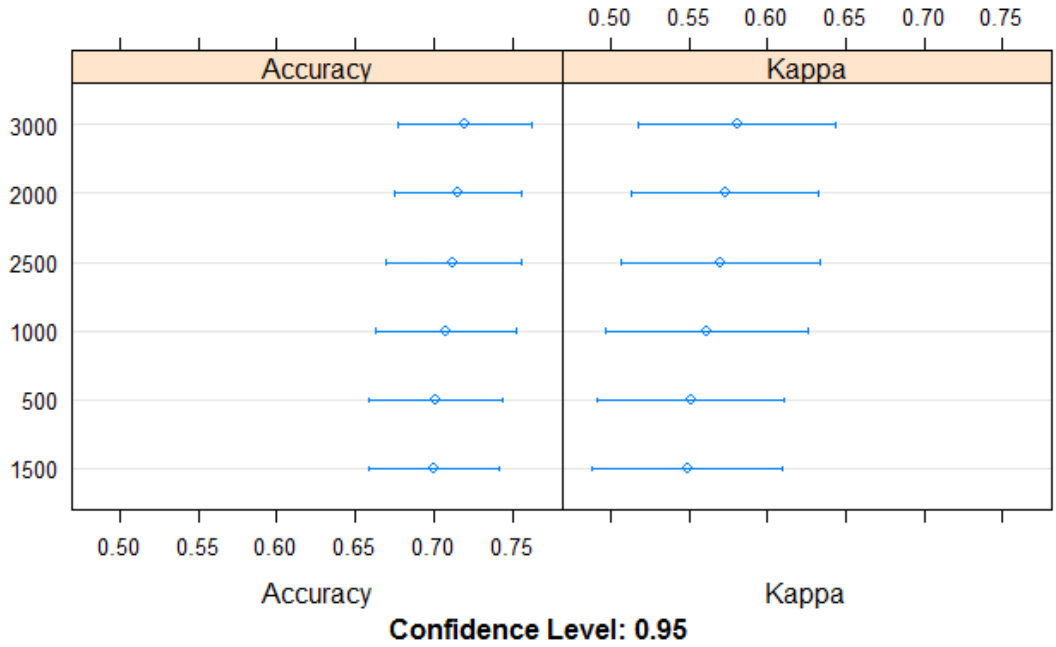

Figure 15. Representation of Accuracy and Kappa scores for the caret models with different values for the parameter ntree.

3.4.3 Tuning both *ntree* and *mtry* parameters simultaneously

The last approach was customizing the *train()* function from the caret package used in the previous methods, to allow the simultaneous adjustment for both *ntree* and *mtry* parameters. Thus, instead of using the default “rf” method, new algorithm parameters and functions needed to be defined in an R list. The R list was later specified in the *method* parameter of the *train()* function to try different values of *ntree* and *mtry*.

The tried values were:

- *mtry*: from 1 to 66 (all possible values).
- *ntree*: 1000, 1500, 2000, 2500 and 3000.

The results can be visualized in Figure 16. The best values for the parameters mtry and ntree adjusted simultaneously by a customized method were *mtry* = 17 and *ntree* = 2500, and the model had an accuracy of 73.74% and Kappa = 60.82%

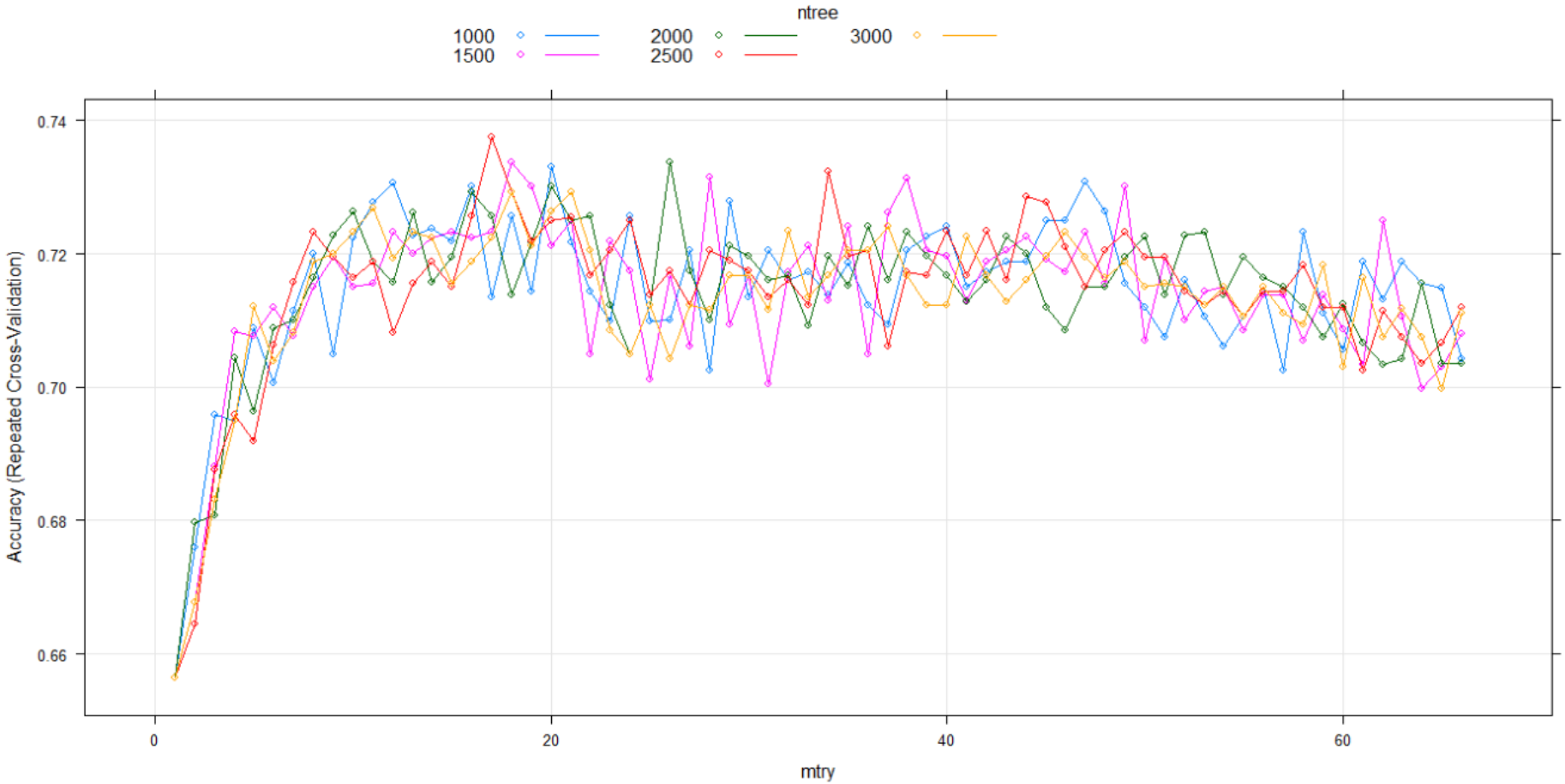

Figure 16. Results for the RF tuning with customized method for both ntree and mtry parameters adjustment simultaneously.

3.4.4 Final model

Finally, the best tuning values (*mtry* = 24, *ntree* = 500) were tried in a single model.

Model performance evaluation

With the aim of evaluating the performance of the model, the test dataset samples were classified in one of the four groups using the function *predict()*. The predictions were compared to the real labels of each sample through a confusion matrix (Table 11) using the function *confusionMatrix()* from the *caret* package.

		Reference			
		control female	control male	depressed male	depressed female
Prediction	control female	13	0	0	0
	control male	0	6	0	0
	depressed male	0	0	4	0
	depressed female	0	0	0	6

Table 11. Confusion matrix of the predicted classification versus the real labels of the test samples.

For this individual model, all the samples of the four groups were classified correctly.

Statistics:

The model had an **accuracy of 100%**, with a confidence interval at 95% between 88.06 and 100% (p-value = $7.849 \cdot 10^{-11}$) and Kappa = 100%, which indicates on average an outstanding agreement between the model predictions and the true values.

Variable importance for the final model

In order to analyze the most important variables in the final model, the variable importance was plotted (Figure 17) with the function *varImpPlot()* from the *randomForest* package. The plot represents the mean decrease in accuracy[2] for the variables with a biggest value for this mean. In the plot, it is possible to appreciate that the most important variables (with a mean decrease accuracy over 10) for the final model are *stress*, *cg16718678* (*NRXN2* gene), *cg18485485* (*DECR1* gene) and *cg23758485* (*SMPD3* gene).

[2] The mean decrease in accuracy refers to the number or proportion of observations that are incorrectly classified by removing the feature in question from the model.

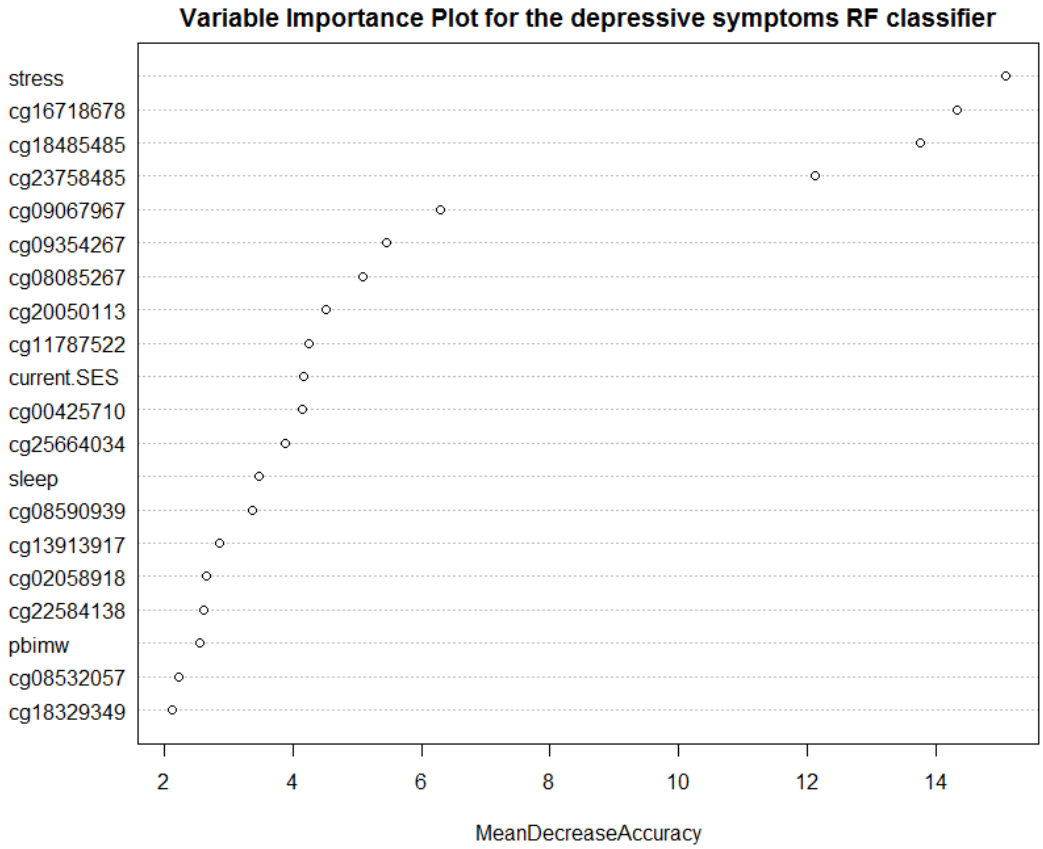

Figure 17. Variable importance plot for the depressive symptoms RF classifier for the 20 variables with highest mean decrease accuracy.

3.5 Web application

A web application was developed for public use of the algorithm. The web was programmed in HTML / PHP and has two options to introduce the necessary patient data to return a prediction, by filling a form or uploading a *csv* file. It was necessary to create the following files:

1. HTML files
   - 1.1. Index (Figure 18): contains the home page of the web application, with a short description of the algorithm and one link for each data introduction option.
   - 1.2. Form (Figure 19): if the selected method for introducing data is by a form, the user will be conducted to this page, which contains a formulary to be filled with the patient data. Once the data is introduced and submitted, the user will be directed to the PHP form results.
   - 1.3. Upload (Figure 20): if the selected method for introducing data is by uploading a *csv* file, the user will be conducted to this page and will be able to browse in their machine for the file containing the patient data and upload it. The *csv* file must contain all the variables for the patient data in a single row, ordered as specified in the web page and separated by commas. Once the file is submitted, the user will be directed to the PHP upload results.

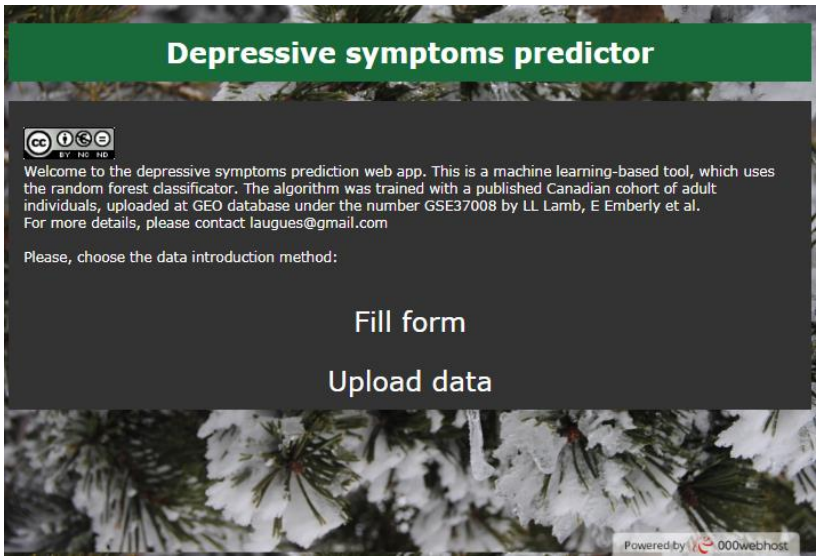

Figure 18. Home page of the web app developed to predict the existence of depressive symptoms.

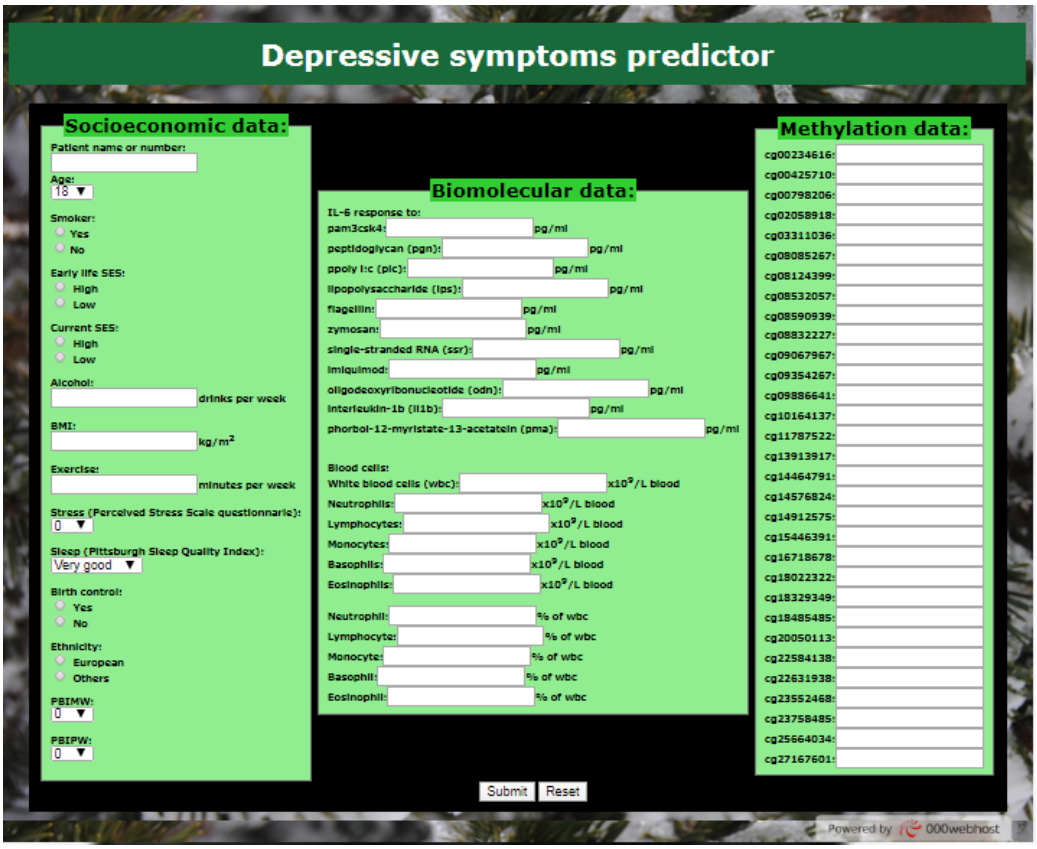

Figure 19. Form page of the web app where the user can fill in the patient data.

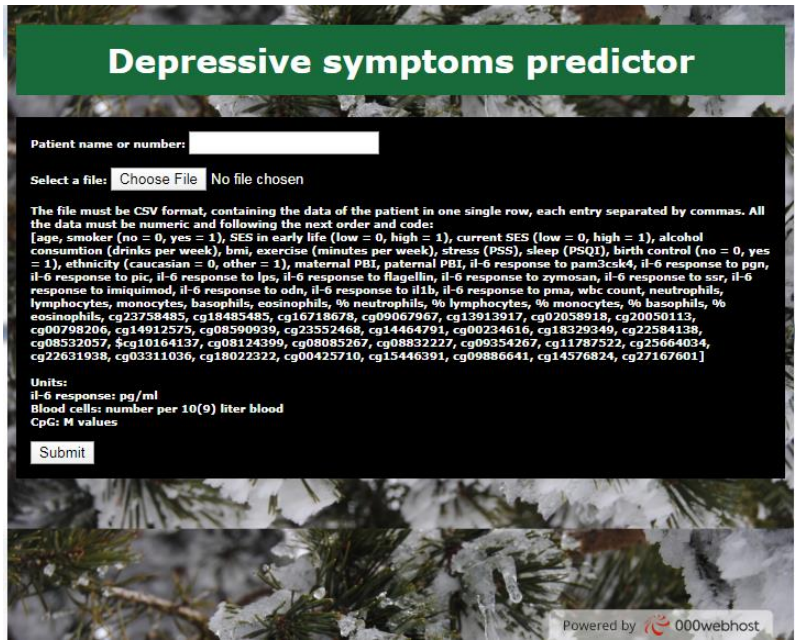

Figure 20. Uploading page of the web app where the user can upload a file containing the patient data.

2. CSS file: contains the styles of the web app (font size, background image, colour, ...)
3. CSV file: a *for* loop together with the function *getTree()* from the package *randomForest* in R was used to extract each tree from the final model forest. Each tree is a matrix containing the structure of one individual decision tree. All the matrixes were bound together and saved in a *csv* file.
4. PHP files
   4.1. Random forest algorithm: a file defining the classes for individual tree branches, for trees and for the entire forest, which will point the pathway to make decisions using the forest: while asking for a result, it will loop through all the trees in the forest making a prediction for each tree. The most frequent answer will be the prediction of the algorithm [31].
   4.2. Form results: after submitting the HTML form, this file will collect all the fields of the form into an array, load the random forest algorithm file, and run the prediction using the forest from the *csv* file. The process will direct to a results page (Figure 21).
   4.3. Upload results: after submitting the HTML upload, it will collect the data from the file into an array and it will follow the same steps than in the forms result file.

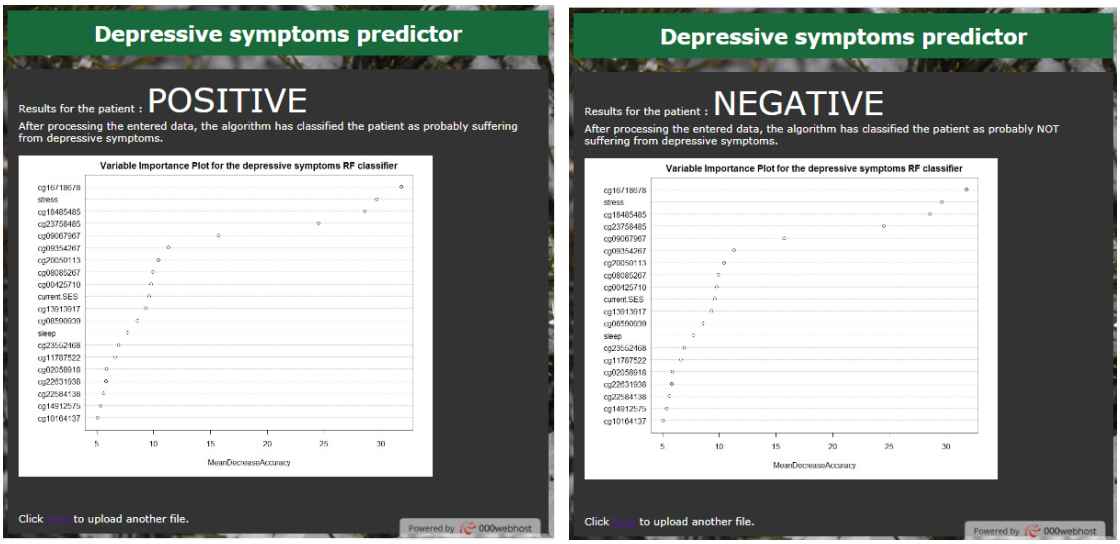

Figure 21. Results page. In the left, capture of the results page when the loaded sample is a probable case of depression. In the left, capture of the results page when the loaded sample is a probable healthy patient.

All the files contents are available in annexes.

During the development of the web app, the files were tested by a local host using XAMPP. Once all the files were ready, they were uploaded into the 000webhost server to create the website, and it was hosted under the address http://desypre.000webhostapp.com/.

4. Discussion

4.1 Stress, current SES, cigarette smoking and IL-6 response to lipopolysaccharid is correlated with depressive symptomatology

The individual linear regression models showed a strong positive correlation between depression and stress (p-val = $8.12 \cdot 10^{20}$), that was already evidenced in the literature [32]. There were two other relationships found in this project: low current SES with depressive symptomatology and cigarette smoking with MDD, both are also supported by previous studies [33] [34].

Regarding to IL-6 response to lipopolysaccharid, Moieni et al. stated that the change in IL-6 following LPS administration it is positively correlated with depression-related symptoms in women, but not in men [35]. This fact explains the correlation in the studied cohort, but further analysis would be needed to analyse the differences between genders.

Since there is a relation between socioeconomic and biomolecular factors and depression, new studies need to be done for DNA methylation in depression with a larger cohort and well defined socioeconomic factors.

4.2 Thirty-one differentialy methylated CpG loci associated with depression

The differential methylation associated with depression is widely spread throughout the genome and here it was possible to differentiate 31 CpG sites influencing the appearance of depressive symptoms in the studied cohort, being the most significant CpGs: *cg23758485* in gene *SMPD3*, *cg18485485* in gene *DECR1*, *cg16718678* in gene *NRXN2* and *cg09067967* in gene *UGDH*.

CpG *cg23758485* is known to be differentially methylated by sex in humans, with a higher methylation in females than in males [36]. There are no previous publications showing a direct association between *SMPD3* and depression, but a recent study found that SMPD3 deficiency perturbs neuronal proteostasis in mouse [37], which could be a starting point to search for other primary pathogenic mechanisms. Moreover, the Comparative Toxicogenomics Database Gene-Disease Associations contains a curated association between SMPD3 and depressive disorder mediated by several toxins (arsenic, bisphenol A, dietary fats, estradiol, ...) [38].

The dataset SNP-Phenotype Associations from GWASdb, a database for human genetic variants, identified by genome-wide association studies [39], associates *DECR1* as one of the 2.466 genes associated with the depression phenotype, but no more further information was found about this association in literature.

Regarding *NRXN2*, Born et al. validated the exhibition of behavioural abnormalities, characterized by social interaction deficits and increased anxiety-like behaviour, in mice lacking Nrxn2α [40], which corresponds with part of the depression symptomatology in humans.

No connexion could be found in literature between *UGDH* and depression. However, it has been mentioned a gender-based differential regulation in various diseases like breast cancer [41], and its expression is known to be stimulated by androgen, a male sex hormone [42]. Thus, it could be a candidate to performing further studies of the association of *UGDH* with the gender differences in depression.

Other connections were found for some of the other differentially methylated CpG sites in this analysis. A methylome-wide association study (MWAS) performed by Aberg et al. identified the *GABBR2* gene as a candidate to be associated with MDD [43]. A recent study found *SLC9A2* gene not only to be related with depression, but as well with suicide in patients with MDD [44]. The gene OXCT1 has been related with schizophrenia, bipolar disorder and depression [45]. The loci *cg22584138* is located at the *SLC6A4* gene, one of the most studied genes associated with depression [11], which is responsible for serotonin re-uptake, known to be often disrupted in depression [46]. TLX2 has been associated with differential regulatory mechanisms in untreated MDD patients [47]. Wilkinson et al. stated that brain tissue from both mice models and post-mortem depressed humans showed downregulation of the DVL1-3 genes [48]. And as a last example, GWAS showed a strong association for depression in *RORA* gene, which is involved in circadian rhythm [49].

Even though the results of this analysis gave new candidate CpG sites associated with depression and confirmed some others already known in literature, it could still be improved applying the pipeline designed from this differential methylation analysis to a larger dataset (with more observations) and using an array with higher resolution (higher number of CpG sites).

4.3 RF algorithm to identify patients showing depressive symptoms

The RF algorithm resulting from the training with the selected variables (socioeconomic, biomolecular and 31 CpGs) predicts if a patient presents depressive symptoms with an accuracy of 100% with a CI at 95% between 88.06 and 100% (p-value = $7.849 \cdot 10^{-11}$) for a test subsample of the studied cohort containing 29 samples, and 73.97% for a 15-fold cross-validated with three repeats model. The parameters of the algorithm were set at *mtry* = 24 and *ntree* = 500.

To reach the highest accuracy for the test model it was necessary to find the best combination of values for both *mtry* and *ntree*. Regarding the *mtry* tune with the default value for *ntree*, the random and grid search showed a variable accuracy while increasing *mtry* value, being lower in the extremes. For the *ntree* tune with the default value for *mtry*, the best accuracy is obtained in higher values of *ntree*. Tuning different combinations of *mtry* and *ntree* does not obtain an improved accuracy compared to the best found at the *mtry* tune with a default *ntree* (*mtry* = 24 and *ntree* = 500), which was used in the final test model.

The accuracy from the cross-validated model could be improved using a larger population, since the only found publically available dataset containing

methylation data together with socioeconomic data and including depression status or score only contains 94 observations. A further improvement of a future cohort would be the depression assessment by different depression scales to improve the previous classification of the patients. However, the pipeline designed for this project could be easily applied to future datasets to generate an improved algorithm.

4.4 *Desypre*, the new tool to detect depression symptoms

Desypre is an online tool addressed to medical doctors in clinical environments to be used as an objective depressive symptoms predictor. To use it, the doctor will need a methylation analysis for the 31 target CpG sites, a blood test analysis containing IL-6 response to different components and blood cells count, the Perceived Stress Scale and the Parental Bonding Index scores, and the socioeconomic data of the patient.

The clinical implications of this tool are that early screening of those at risk for MDD may be possible and from an objective approach, allowing an earlier direction of clinical treatment course.

5. Conclusions

This project has developed a RF-based prediction tool for depressive symptoms from socioeconomic, biomolecular and differential DNA methylation features. During the development process, it has been found that stress, current SES, cigarette smoking and IL-6 response to lipopolysaccharid are correlated to depressive symptomatology. Also, 31 CpG had been found to be differentially methylated.

Despite not being able to fully follow the initial planned pipeline, it has been possible to reach the final goal building the prediction tool with a different approach. Initially, the differentially methylated CpGs in depression were going to be identified in literature. But the cohort was analyzed with a low resolution array in which the main loci found in literature were not present. For this reason, it was necessary to perform a differential DNA methylation analysis, with which it was possible to identify 31 CpGs differentially methylated.

Even though this project has obtained its goal outcome, it would be possible to improve it by using a larger cohort with different depression symptoms assessment methods. The pipeline of this project has been designed to be easily applicable to future cohorts.

4. Glossary

BMI	Body Mass Index
CES-D	Center for Epidemiologic Studies Depression scale
CpG	Cytosine Guanine dinucleotid
DNAm	DNA methylation
DSM-V	American Psychiatric Association Diagnostic and Statistical Manual (version V)
EWAS	Epigenome-Wide Association Study
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Study
IL-6	Interleukin-6
MDD	Major Depressive Disorder
mtry	Number of randomly selected variables as candidates in each division in a Random Forest algorithm
MZ	Monozygotic
M values	DNA methylation values
ntree	Number of trees produced in a Random Forest algorithm
PC	Principal Components
PCA	Principal Components Analysis
RF	Random Forest
SES	Socioeconomic Status

5. Bibliography

[1] World Health Organization, *Depression and Other Common Mental Disorders: Global Health Estimates*. Geneva, 2017.

[2] T. Vos *et al.*, “Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015,” *Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016.

[3] M. H. Rapaport, C. Clary, R. Fayyad, and J. Endicott, “Quality-of-life impairment in depressive and anxiety disorders,” *Am. J. Psychiatry*, vol. 162, no. 6, pp. 1171–1178, 2005.

[4] J. P. Lépine and M. Briley, “The increasing burden of depression,” *Neuropsychiatr. Dis. Treat.*, vol. 7, no. SUPPL., pp. 3–7, 2011.

[5] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. 2013.

[6] J. Mill and A. Petronis, “Molecular studies of major depressive disorder: The epigenetic perspective,” *Mol. Psychiatry*, vol. 12, no. 9, pp. 799–814, 2007.

[7] J. R. Swartz, A. R. Hariri, and D. E. Williamson, “An epigenetic mechanism links socioeconomic status to changes in depression-related brain function in high-risk adolescents,” *Mol. Psychiatry*, vol. 22, no. 2, pp. 209–214, 2017.

[8] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck, “Epigenome-wide association studies for common human diseases,” *Nat. Rev. Genet.*, vol. 12, no. 8, pp. 529–541, 2011.

[9] Y. Li *et al.*, “The DNA methylome of human peripheral blood mononuclear cells,” *PLoS Biol.*, vol. 8, no. 11, 2010.

[10] G. Felsenfeld, C. D. Allis, T. Jenuwein, and D. Reinberg, *Epigenetics*. 2007.

[11] M. Li, C. D’Arcy, X. Li, T. Zhang, R. Joober, and X. Meng, “What do DNA methylation studies tell us about depression? A systematic review,” *Transl. Psychiatry*, vol. 9, no. 1, 2019.

[12] M. Shimada *et al.*, “An epigenome-wide methylation study of healthy individuals with or without depressive symptoms,” *J. Hum. Genet.*, vol. 63, no. 3, pp. 319–326, 2018.

[13] M. Srisurapanont, S. Suttajit, K. Eurviriyanukul, and P. Varnado, “Discrepancy between objective and subjective cognition in adults with major depressive disorder,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, 2017.

[14] W. Huber *et al.*, “Orchestrating high-throughput genomic analysis with Bioconductor,” *Nat. Methods*, vol. 12, no. 2, pp. 115–121, 2015.

[15] A. G. H. McCollum, “Introduction,” *Semin. Orthod.*, vol. 15, no. 3, pp. 159–160, 2009.

[16] D. Sean and P. S. Meltzer, “GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor,” *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[17] G. Wickham, *ggplot2: elegant graphics for data analysis*. 2016.

[18] J. Friedman and T. Hastie…, “Regularized paths for generalized linear models via coordinate descent (Technical Report,” *Citeseer*, vol. 33, no. 1, 2008.

[19] M. P, *R graphics*. 2005.

[20] B. Auguie and A. Antonov, “Package ‘gridExtra’. Miscellaneous Functions for ‘Grid’ Graphics,” 2017.
[21] G. K. Smyth, “limma: Linear Models for Microarray Data,” *Bioinforma. Comput. Biol. Solut. Using R Bioconductor*, no. 2005, pp. 397–420, 2005.
[22] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. 2002.
[23] L. Komsta and F. Novomestky, “Moments, cumulants, skewness, kurtosis and related tests,” 2015.
[24] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[25] H. Wickham, “Reshaping Data with the reshape Package,” *J. Stat. Softw.*, vol. 21, no. 12, pp. 1–20, 2007.
[26] L. L. Lam *et al.*, “Factors underlying variable DNA methylation in a human community cohort,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. SUPPL.2, pp. 17253–17260, 2012.
[27] R. Edgar, M. Domrachev, and A. E. Lash, “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, 2002.
[28] L. S. Radloff, “The CES-D scale: a self-report depression scale for research in the general population.,” *Appl. Psychol. Meas.*, vol. 1, no. 3, pp. 385–401, 1977.
[29] J. Naue *et al.*, “Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression,” *Forensic Science International: Genetics*, vol. 31. pp. 19–28, 2017.
[30] J. Brownlee, “Tune Machine Learning Algorithms in R (random forest case study),” 2016. [Online]. Available: https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/. [Accessed: 13-Dec-2019].
[31] D. Batten, “Random Forest Classifiers as a Web Service in PHP,” 2015. [Online]. Available: http://daynebatten.com/2015/06/random-forest-web-service-php/. [Accessed: 13-Dec-2019].
[32] H. C, “Stress and Depression,” *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 293–319, 2005.
[33] V. Lorant, D. Deliège, W. Eaton, A. Robert, P. Philippot, and M. Ansseau, “Socioeconomic inequalities in depression: A meta-analysis,” *Am. J. Epidemiol.*, vol. 157, no. 2, pp. 98–112, 2003.
[34] D. M. Fergusson, R. D. Goodwin, and L. J. Horwood, “Major depression and cigarette smoking: results of a 21-year longitudinal study,” *Psychol. Med.*, vol. 33, no. 8, pp. 1357–1367, 2003.
[35] M. Moieni, M. R. Irwin, I. Jevtic, R. Olmstead, E. C. Breen, and N. I. Eisenberg, “Sex differences in depressive and socioemotional responses to an inflammatory challenge: implications for sex differences in depression.,” *Neuropsychopharmacology*, vol. 40, pp. 1709–1716, 2015.
[36] J. D. Blair and E. M. Price, “Illuminating potential technical artifacts of DNA-methylation array probes,” *Am. J. Hum. Genet.*, vol. 91, no. 4, pp. 760–762, 2012.
[37] W. Stoffel, B. Jenke, I. Schmidt-Soltau, E. Binczek, S. Brodesser, and I. Hammels, “SMPD3 deficiency perturbs neuronal proteostasis and causes progressive cognitive impairment,” *Cell Death Dis.*, vol. 9, no. 507, 2018.
[38] A. P. Davis *et al.*, “The Comparative Toxicogenomics Database’s 10th year anniversary: Update 2015,” *Nucleic Acids Res.*, vol. 43, no. D1, pp.

D914–D920, 2015.
[39] M. J. Li *et al.*, “GWASdb: A database for human genetic variants identified by genome-wide association studies,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. 1047–1054, 2012.
[40] G. Born *et al.*, “Genetic targeting of NRXN2 in mice unveils role in excitatory cortical synapse function and social behaviors,” *Front. Synaptic Neurosci.*, vol. 7, no. FEB, pp. 1–16, 2015.
[41] M. Callari *et al.*, “Gene expression analysis reveals a different transcriptomic landscape in female and male breast cancer,” *Breast Cancer Res. Treat.*, vol. 127, no. 3, pp. 601–610, 2011.
[42] Q. Wei, R. Galbenus, A. Raza, R. L. Cerny, and M. A. Simpson, “Androgen-stimulated UDP-glucose dehydrogenase expression limits prostate androgen availability without impacting hyaluronan levels,” *Cancer Res.*, vol. 69, no. 6, pp. 2332–2339, 2009.
[43] K. Aberg *et al.*, “No TitleMethylome-wide association findings for major depressive disorder overlap in blood and brain and replicate in independent brain samples.,” *Mol Psychiatry*, 2018.
[44] Y.-K. Kim *et al.*, “Association between norepinephrine transporter gene (SLC6A2) polymorphisms and suicide in patients with major depressive disorder,” *J. Affect. Disord.*, vol. 158, pp. 127–132, 2014.
[45] X. Chen, F. Long, B. Cai, X. Chen, and G. Chen, “A novel relationship for schizophrenia, bipolar and major depressive disorder Part 3: Evidence from chromosome 3 high density association screen,” *J. Comp. Neurol.*, vol. 526, no. 1, pp. 59–79, 2018.
[46] D. Lam, M. L. Ancelin, K. Ritchie, R. Freak-Poli, R. Saffery, and J. Ryan, “Genotype-dependent associations between serotonin transporter gene (SLC6A4) DNA methylation and late-life depression,” *BMC Psychiatry*, vol. 18, no. 1, pp. 1–10, 2018.
[47] F. Xu *et al.*, “Differential co-expression and regulation analyses reveal different mechanisms underlying major depressive disorder and subsyndromal symptomatic depression,” *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–10, 2015.
[48] M. B. Wilkinson *et al.*, “A novel role of the WNT-dishevelled-GSK3β signaling cascade in the mouse nucleus accumbens in a social defeat model of depression,” *J. Neurosci.*, vol. 31, no. 25, pp. 9084–9092, 2011.
[49] A. Terracciano *et al.*, “Florida State University Libraries Genome-Wide Association Scan of Trait Depression,” *Biol. Psychiatry*, vol. 68, no. 9, pp. 811–817, 2010.
[50] R. A. Irizarry *et al.*, “Comprehensive high-throughput arrays for relative methylation (CHARM),” *Genome Res.*, vol. 18, no. 5, pp. 780–790, 2008.

6. Annexes

6.1 R Script

annexes\script.Rmd

6.2 HTML, PHP and CSS files content

annexes\index.html
annexes\form.html
annexes\upload.html
annexes\depression_forest.csv
annexes\random_forest.php
annexes\depression_forest_form.PHP
annexes\depression_forest_upload.PHP
annexes\styles.css