Predictive analysis of depression in students using Machine Learning Algorithms

Master’s degree in engineering and computer science Machine Learning & Deep Learning – A. Y. 2024/2025
Filippo Ridolfi VR502263

Problem definition and main objective

Depression among students is an increasingly prevalent phenomenon, negatively impacting their psychological well-being, academic performance, and social relationships. However, the signs of depression are often overlooked due to a lack of strategies for early identification and prevention.

The main objective of the project is to develop predictive models that can identify potential students at risk of depression.

Dataset

- Format: CSV format
- Rows: Each row represents an individual student.
- Columns: Each column represents a specific feature or attribute.

It is composed of 27,898 observations and 18 columns.

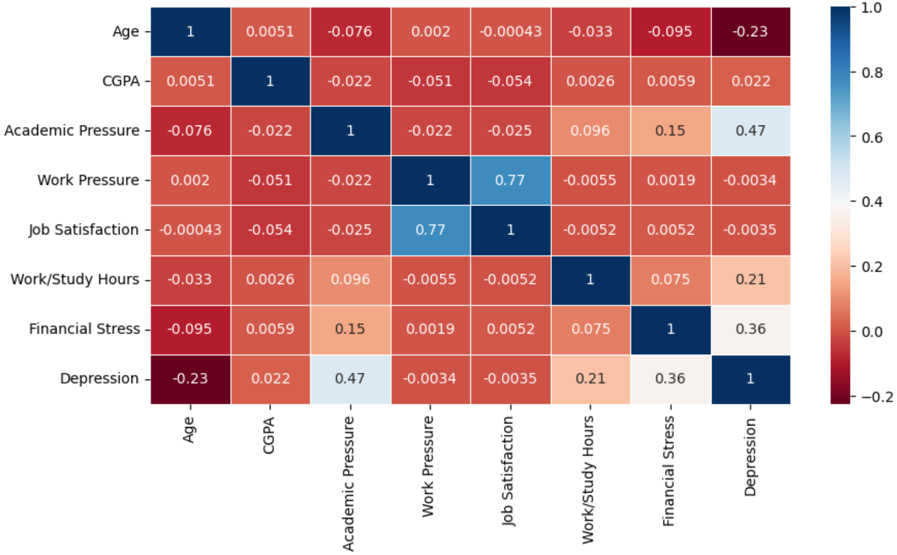
Dataset_link

Schedule of work

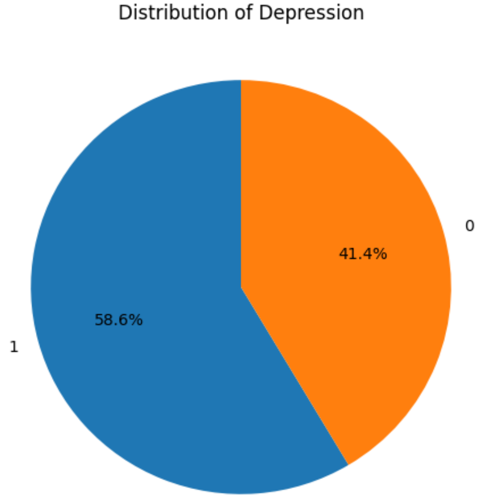
- Dataset analysis, features distribution, and extraction.
- Dimensionality reduction with PCA.
- Classification models: KNN, Random Forest, SVM.
- Performance evaluations.
- Conclusions.

Dataset analysis and features distribution

Some interesting points found during the analysis:

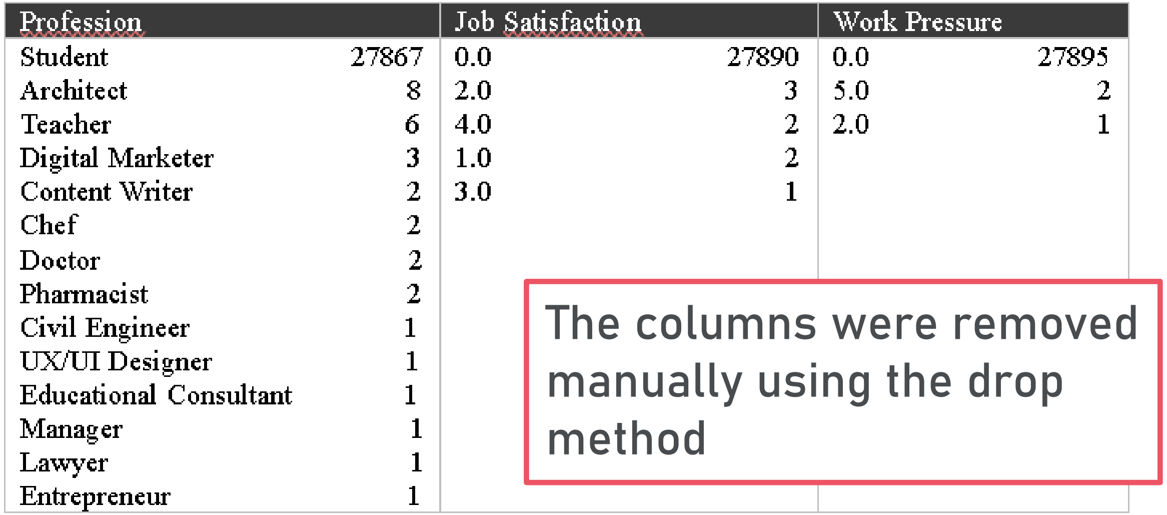


	Age	CGPA	Academic Pressure	Work Pressure	Job Satisfaction	Work/Study Hours	Financial Stress	Depression
Age	1	0.0051	-0.076	0.002	-0.00043	-0.033	-0.095	-0.23
CGPA	0.0051	1	-0.022	-0.051	-0.054	0.0026	0.0059	0.022
Academic Pressure	-0.076	-0.022	1	-0.022	-0.025	0.096	0.15	0.47
Work Pressure	0.002	-0.051	-0.022	1	0.77	-0.0055	0.0019	-0.0034
Job Satisfaction	-0.00043	-0.054	-0.025	0.77	1	-0.0052	0.0052	-0.0035
Work/Study Hours	-0.033	0.0026	0.096	-0.0055	-0.0052	1	0.075	0.21
Financial Stress	-0.095	0.0059	0.15	0.0019	0.0052	0.075	1	0.36
Depression	-0.23	0.022	0.47	-0.0034	-0.0035	0.21	0.36	1



Features distribution

Observing the distribution data, it was noted that the Job Satisfaction, Work Pressure, and Profession features do not provide any relevant information for the purpose, given their highly unbalanced distribution.



Profession		Job Satisfaction		Work Pressure	
Student	27867	0.0	27890	0.0	27895
Architect	8	2.0	3	5.0	2
Teacher	6	4.0	2	2.0	1
Digital Marketer	3	1.0	2		
Content Writer	2	3.0	1		
Chef	2				
Doctor	2				
Pharmacist	2				
Civil Engineer	1				
UX/UI Designer	1				
Educational Consultant	1				
Manager	1				
Lawyer	1				
Entrepreneur	1				

The columns were removed manually using the drop method

Features distribution and extraction

I transformed the categorical variables into a numeric format to be able to use them in the models through both One-hot encoding and Label encoding.

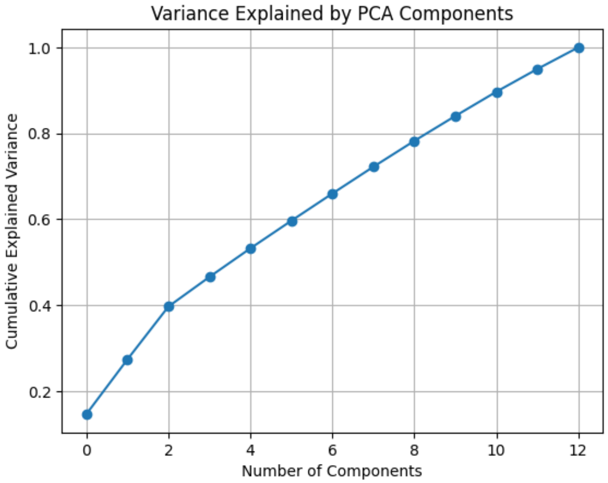
The variables chosen for Label encoding were: Degree, Sleep Duration, Dietary Habits and City.

Categorical variables such as City and Degree do not have a natural order but tests have been carried out both regarding their conversion via Label Encoding and One-Hot encoding, observing that the changes are not significant.

Dimensionality reduction

Objective: increase efficiency and model performance. The technique chosen was PCA (Principal Component Analysis), which transforms the data into a new base of smaller dimensions.

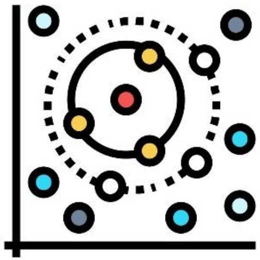
The variance was set to 95% to reduce the dimensionality of the dataset but still maintain a significant part of the variance.

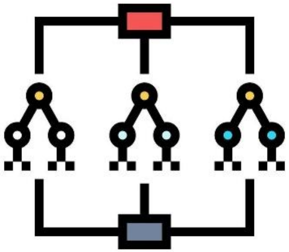


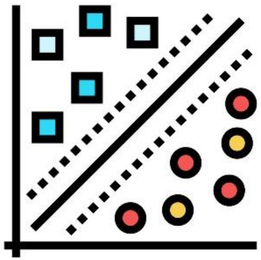
Machine Learning models

Used models:

- K-Nearest Neighbors (KNN): quick and easy, suitable for nonlinear data.
- Random Forest: robust, resistant to outliers, generally good performance.
- Support Vector Machine (SVM): effective for linear and nonlinear problems.



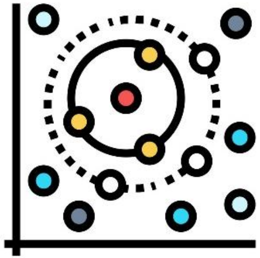


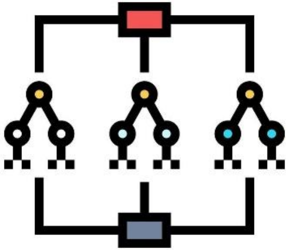


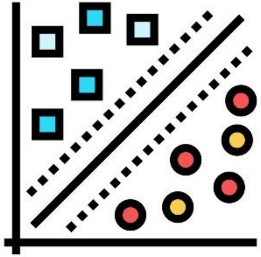
ML models: procedure

For each Machine Learning models I followed this steps:

- Try different values for constant variables (k for KNN, num of trees for Random Forest, C parameter for SVM).
- To train the model I performed 5-fold cross-validation.
- Having chosen the constant parameter with the best performance, I tested the model with both the validation data and the testing data.



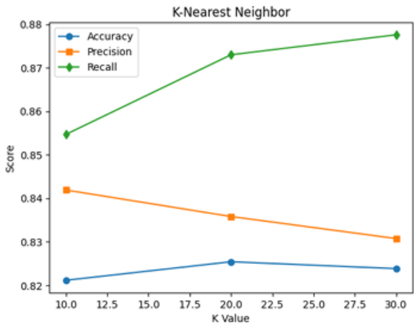


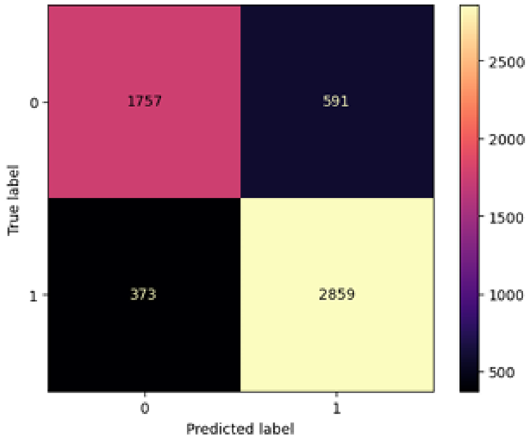


K-Nearest Neighbors (KNN)

Testing the model on twenty neighbors the results in the testing phase were:

- The correctly classified instances are on the total 82.72%.
- 82.86% of the examples classified as positive by the model were positive.
- The positive instances were correctly identified for 88.46%

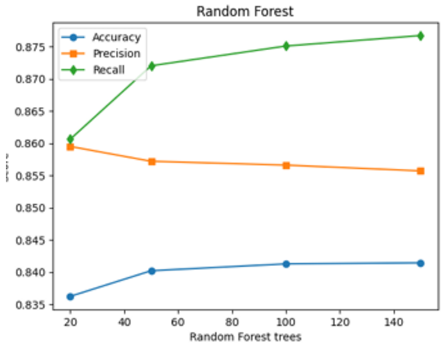


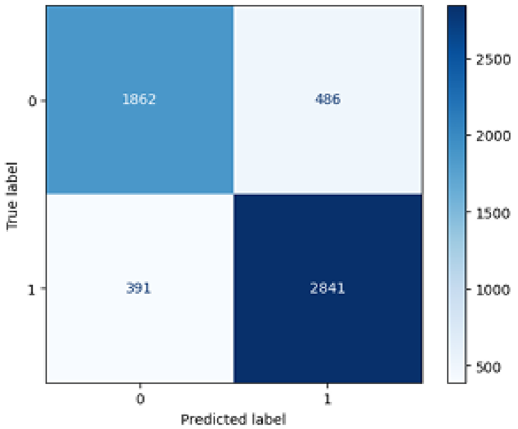


Random Forest

Testing the model on fifty trees the results in the testing phase were:

- The correctly classified instances are on the total 84.28%.
- 85.39% of the examples classified as positive by the model were positive.
- The positive instances were correctly identified for 87.90%

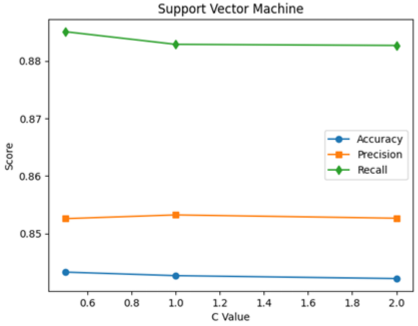


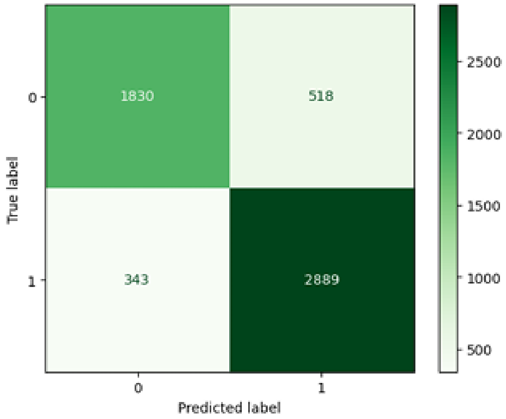


Support Vector Machine

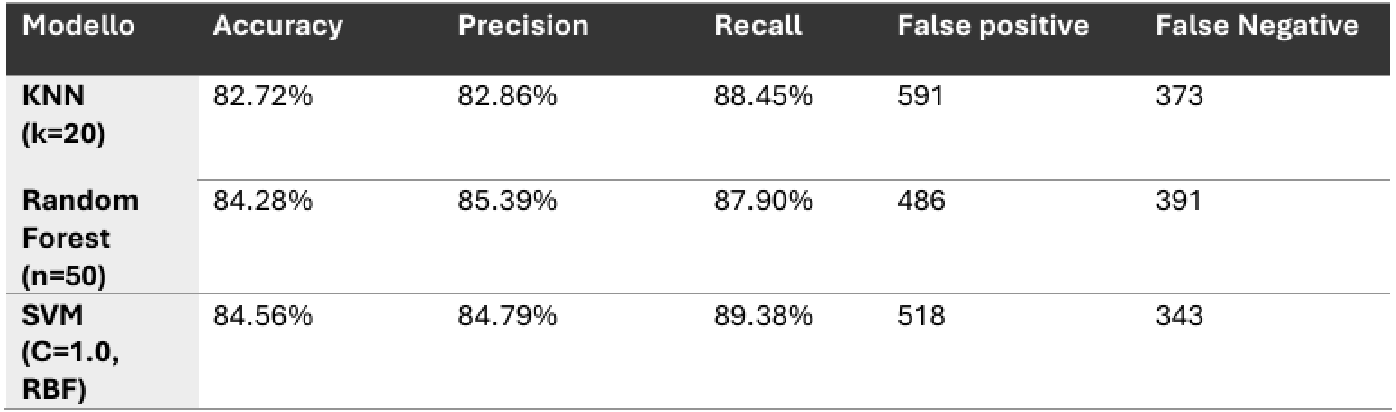
Testing the model on C = 1.0 the results in the testing phase were:

- The correctly classified instances are on the total 84.56%.
- 84.79% of the examples classified as positive by the model were positive.
- The positive instances were correctly identified for 89.38%





Performance summary



Modello	Accuracy	Precision	Recall	False positive	False Negative
KNN (k=20)	82.72%	82.86%	88.45%	591	373
Random Forest (n=50)	84.28%	85.39%	87.90%	486	391
SVM (C=1.0, RBF)	84.56%	84.79%	89.38%	518	343

It is observed that SVM achieved the best results in terms of accuracy, recall, and the number of false negatives, proving to be the most effective model.

Conclusions and future prospective

Among the models tested, SVM emerged as the most suitable choice, mainly due to its low number of false negatives.

This study highlights the potential of machine learning to detect early signs of depression among students, demonstrating how these technologies could be leveraged to develop more proactive and data-driven approaches to mental health care.

Machine learning could play a crucial role in improving psychological well-being and providing timely support to those who need it most.