Information & Service Systems

UNIVERSITÄT DES SAARLANDES

Data Science

Summer Semester 2022

Depression forecast for supporting students at university

Information & Service Systems

DFKI

SIC Saarland Informatics Campus

Problem Statement

- Depression is a common health problem, ranking third after cardiac and respiratory diseases as a major cause of disability. There is evidence to suggest that university students are at higher risk of depression, despite being a socially advantaged population, but the reported rates have shown wide variability across settings. For instance, a study in Germany has shown that 23 % of students had depressive symptoms during the lockdown, compared to 13 % before the pandemic. Thus, to analyze the depression level in university students might help supporting individuals in preventing serious mental problems.

Task

- Specification of a service use case for university students regarding depression forecast.
- Preprocessing and descriptive analysis of given dataset (data of students of different study fields and semesters).
- Analyzing data with ML (forecast of depression score).
- Visualization of results and service.

Dataset

- Provided “Depression” dataset, 754 samples and 34 features

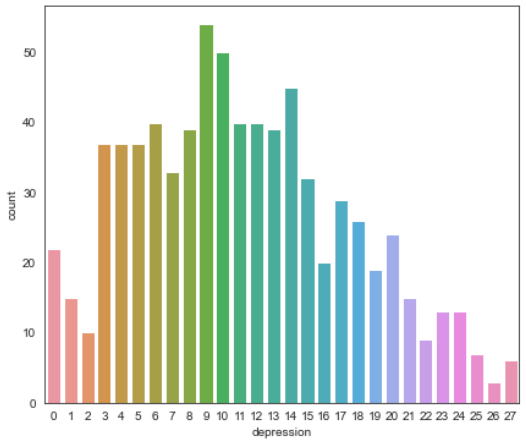
Overview

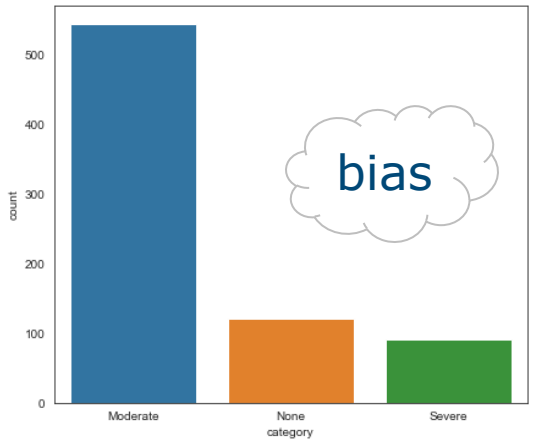
32 features
X

34 features

2 features Y

Ratio scale	Interval scale	Nominal scale		Ordinal scale
Age	Wakeup	Mood	Smoker	Hangout
Semester	Breakfast	Gender	Alcoholic	Social_event
Siblings	Launch	Subject	Drug_addicted	Pray
Study_hours	Dinner	Living	medication	Sports
Moives_per_week	Sleep	Employment		Exercises
Music hours		Personality		Video_games
Friends,		Relationship		Meditation
		Hobbies		Phone_hours





Source: https://drive.google.com/drive/folders/1Yj1RReg1mxZboc6XGtu2IkHdn66Qv-dA

Data Pre-processing for independent variables

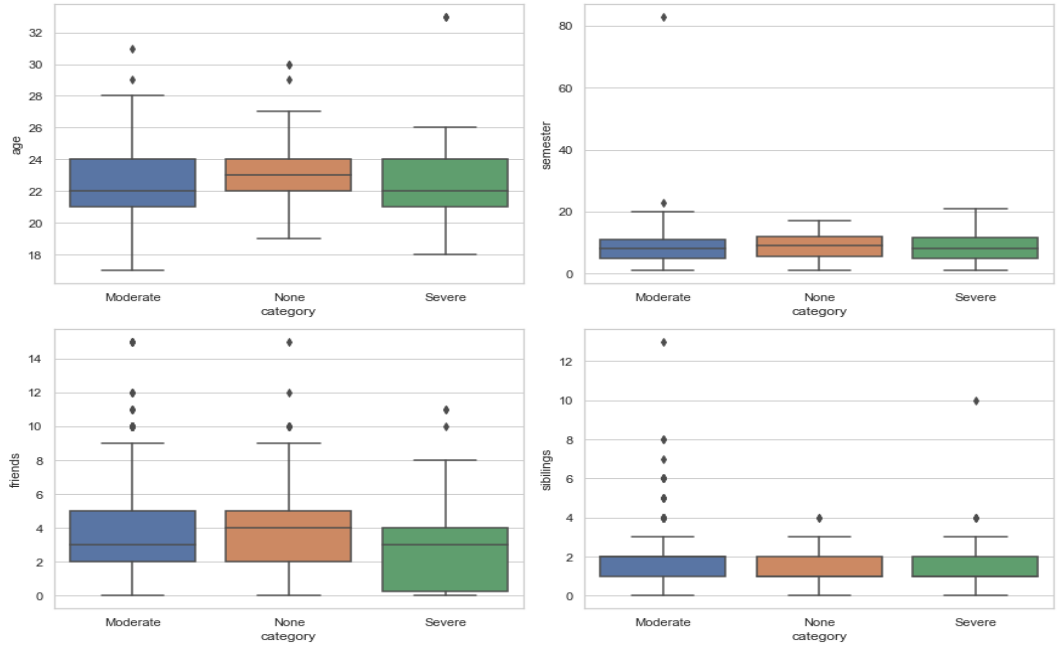
- Data Cleaning
  - remove items with NaN value in medication column
  - remove items with unreasonable value in ‘movies_per_week’ column
  - outlier detection

- Sleep Duration
  - combine the ‘wakeup’ and ‘sleep’ column to calculate the sleep duration and then replace

- Subject
  - common wrongly written subjects and similar subjects; remove subjects which are occurring less than 5 times
  - one hot encoding for the subjects left

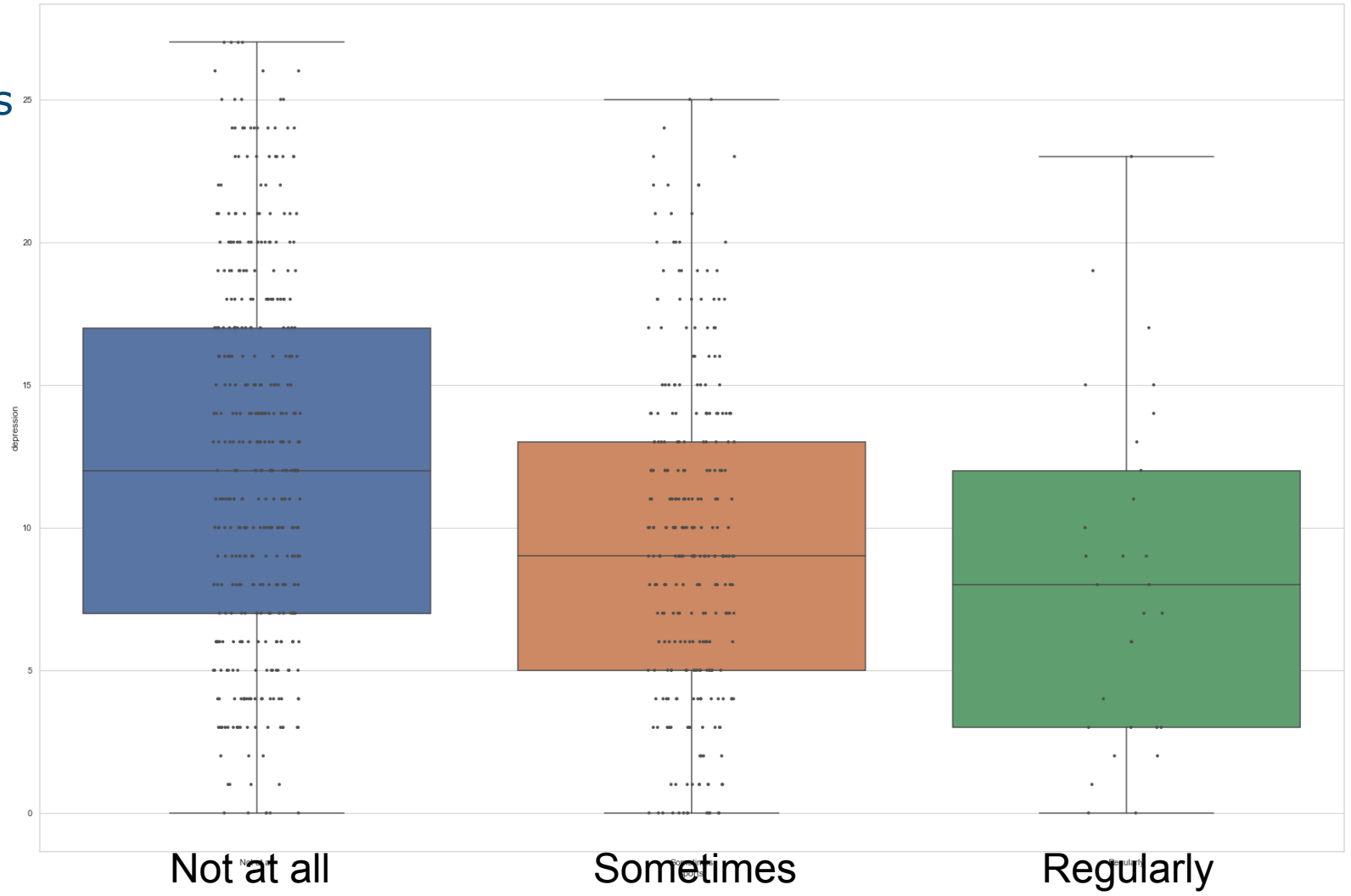
- Hobbies
  - choose 8 most common hobbies as one hot encoding and ignore others

- Categorical to numeric and one hot encoding for other features



Descriptive Analysis

- This plot shows the relationship between Sports and Depression.
- Therefore you can see that even doing sports sometimes decreases your depression. This statement is also backed up by the Jewett et.al 2014



https://www.sciencedirect.com/science/article/pii/S1054139X14001967?casa_token=M7zHgxUAyB4AAAAA:aA2xHdVOTGsrceronbmvh5mff-1-

Feature Selection

- Remove Improper Features
  - unreasonable: breakfast, lauch, dinner
  - replaced: wakeup, sleep, hobbies
  - targets: depression, category

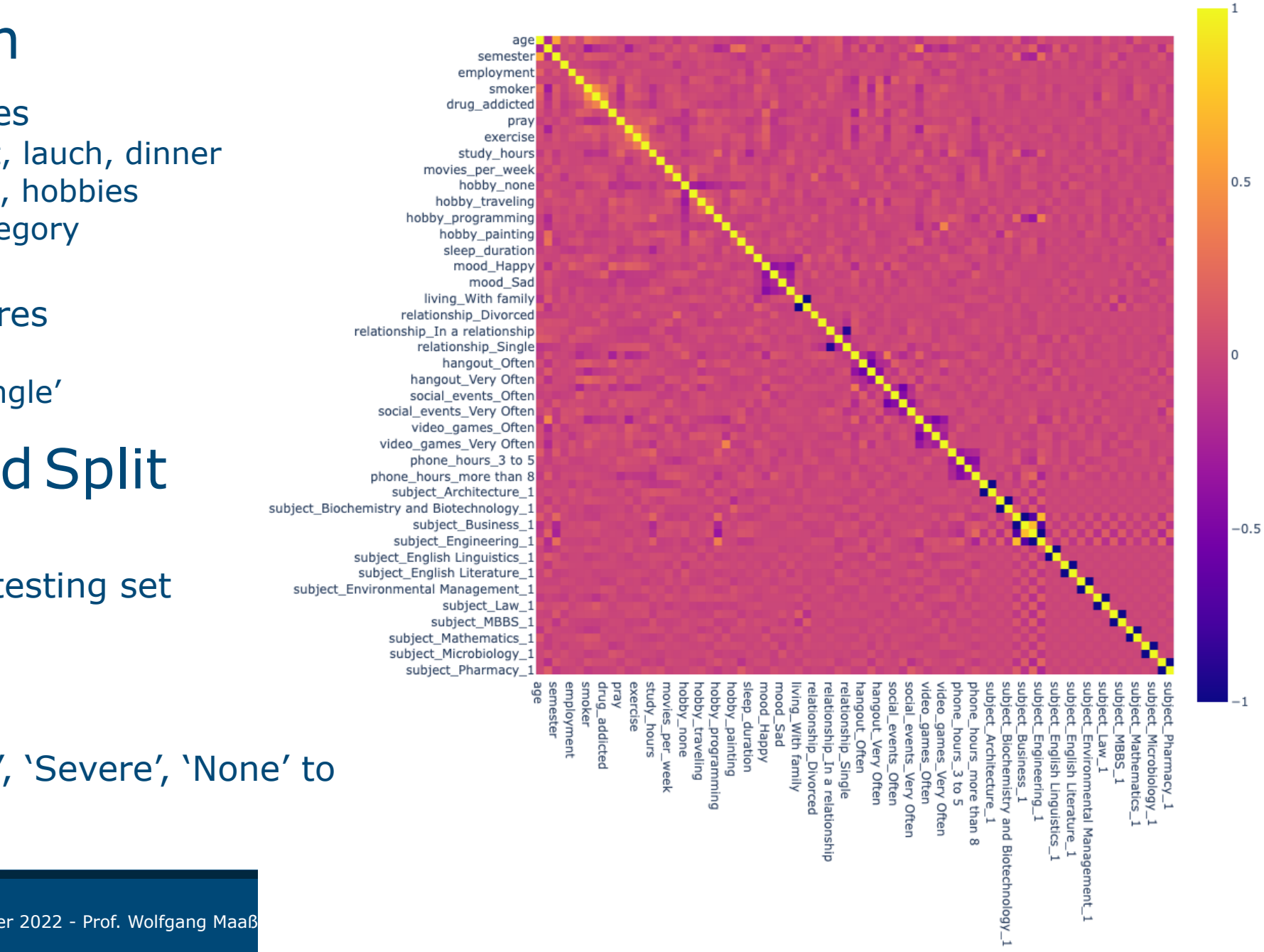
- Remove Correlated Features
  - threshold = 0.8
  - remove ‘relationship_Single’

Normalization and Split

- Scale features to 0-1
- Split for training set and testing set

Data Balancing

- SMOTE
- Oversampling ‘Moderate’, ‘Severe’, ‘None’ to be the same amount



Statistics Overview (OLS)

- To understand the dataset better and the correlation between the data and the depression score we are using OLS Regression to get R-squared

	Original Data	Oversampling	Undersampling	SMOTE
R-squared	0.398	0.620	0.651	0.541
Adj. R-squared	0.322	0.600	0.512	0.512

- The balanced dataset improves the overall correlation a lot, no matter which directions.
- We use the balanced dataset (1143 samples) with SMOTE for the following analysis.

Model Creation and Evaluation

	Random Forest	KNN	SVM
Test Accuracy	0.79	0.60	0.84
Train Accuracy	1.00	0.99	0.88

- Feature Importance for Random Forest
  - threshold = 0.02
  - 18 features left: age, gender, semester, siblings, employment, personality, friends, pray, sports, exercise, meditation, study_hours, movies_per_week, music hours, **sleep_duration,** mood_Happy, mood_Lost, mood_Tenend
- Hyper-parameters search based on Precision and Recall
- SVM performs best
  - data mess up in high-dimension but kernel works
  - Relationship beyond linear exists

Conclusion

- Both OLS and Feature Importance help us to analyze the important reasons (features) that cause depression to students in university
- Our machine learning models can predict depression category based on specific features in a high accuracy

Service

- Since we analyze the main reasons that cause depression to students in university so we could make a small service that students can test by themselves, then the output depression level could be used as a reference to help students decide if make an appointment in Psychology Department.

Outlook

- Since there are still some results not ideal, in the future, we could try to collect more data, use other feature selection strategies and other algorithms to get better performance, in order to generate more applicable models.

Information & Service Systems

UNIVERSITÄT DES SAARLANDES

Thank you