Project Title: Predicting Depression Risk Using Machine Learning
Team Members: Karl-Ustav Kõlar, Oskar Pukk

Task 2. Business understanding (0.5 points)

Identifying Business Goals

Background: Depression is a growing concern globally, affecting millions of individuals across all demographics. Despite increasing awareness, many people at risk remain undiagnosed or untreated due to limited access to healthcare or delayed interventions. This project leverages a dataset collected through an anonymous survey conducted across various cities in January and June 2023, capturing responses from adults aged 18 to 60. The survey included diverse factors such as job and study satisfaction, family history of mental health issues, and lifestyle attributes. This non-clinical dataset enables us to explore correlations between everyday factors and mental health risks, providing a foundation for predictive modeling.

Business Goals: The primary goal of this project is to develop a machine learning-based predictive model that identifies individuals at risk of depression using demographic and lifestyle data. By identifying key contributing factors, this project aims to facilitate early awareness and intervention programs. The end users of this project are primarily researchers, policymakers, and advocacy organizations focused on mental health, rather than a profit-driven business. Despite that, the model can be used in a profit-driven business to determine depression risk in employees and trying to mitigate it, thus improving their ability to work and the wellbeing of themselves.

Business Success Criteria: The success of this project will be evaluated by the accuracy, reliability, and interpretability of the model. Quantitatively, the model must achieve at least 93% prediction accuracy on validation datasets. Qualitatively, the project must provide clear and actionable insights into the key drivers of depression risk, enabling stakeholders to target interventions effectively.

Assessing Your Situation

Inventory of Resources: The project is supported by a comprehensive dataset collected during a survey in 2023. Tools available include Python libraries such as Scikit-learn, TensorFlow, and XGBoost, along with the computing power of personal laptops. The team consists of two data science students with skills in preprocessing, model building, and result interpretation.

Requirements, Assumptions, and Constraints: The project requires ethical adherence to data use guidelines, ensuring that participants' anonymity is preserved. It assumes that self-reported survey data accurately reflects underlying mental health risks. Constraints include the absence of clinical diagnostic data, reliance on non-clinical indicators.

Risks and Contingencies: Potential risks include biased data due to self-reporting, overfitting in machine learning models, and misinterpretation of results. To address these, the project incorporates cross-validation and interpretable models to enhance generalizability and reliability.

Terminology: A glossary will be maintained to ensure consistent understanding of terms such as "depression risk" (binary target variable) and "predictive features" (factors influencing outcomes).

Task 3. Data understanding (1 points)

Our project is based on a Kaggle competition so we got our data from there. That means that the data was already gathered for us. On Kaggle, there's a short overview of the dataset:

“”” *This section is copied from the competition's Data tab*
This dataset was collected as part of a comprehensive survey aimed at understanding the factors contributing to depression risk among adults. It was collected during an anonymous survey conducted between January and June 2023. The survey was conducted across various cities, targeting individuals from diverse backgrounds and professions. Participants, ranging from 18 to 60 years old, voluntarily provided inputs on factors such as age, gender, city, degree, job satisfaction, study satisfaction, study/work hours, and family history among others. Participants were asked to provide inputs without requiring any professional mental health assessments or diagnostic test scores.
“””

In total, data was gathered of 140700 people. The dataset consisted of nineteen attributes and one label:

“Name”: Name of the participant. Categorical value

“Gender”: Gender of the participant. Two possible values: “Male” and “Female”. “Male” was slightly more common.

“Age”: Age of the participant. Numerical value between 18 and 60

“City”: A categorical value that shows which city the participant lives in. 98 different values, out of which 30 had above 100 participants

“Working Professional or Student”: Shows whether the participant is a working professional or a student. Two possible values: “Working professi:onal” and “Student”. “Working professional” is much more common. There was no other option so people who weren't either were not accounted for, there were no NaN values either.

“Profession”: Profession of the participant. It has many different values, also many missing and erroneous values.

“Academic Pressure”: Answered by students. Participants self-assessed their academic pressure from 1 to 5. A lot of missing values, because most participants are not students.

“Work Pressure”: Same as “Academic Pressure”, but instead answered by working professionals.

“CGPA”: Cumulative Grade Point Average – a metric described to value students’ performance in school. In this dataset the values are between 5.03 and 10.0. Answered by students.

“Study Satisfaction”: Answered by students. Participants self assessed their study satisfaction from 1 to 5. Lots of missing values, since it was answered by students.

“Job Satisfaction”: Same as “Study satisfaction”, but answered by working professionals

Note: There were slight inconsistencies for “Academic Pressure”, “Work Pressure”, “Study Satisfaction” and “Job Satisfaction”, where rarely students and working professionals answered the wrong questions (meaning that for example students answered “Work Pressure”).

“Sleep Duration”: An attribute which shows how long the participant usually sleep at night. Has four main values: “Less than 5 hours”, “5-6 hours”, “7-8 hours” and “More than 8 hours”. Also has other values out of which some are clearly erroneous.

“Dietary Habits”: Dietary habits of the participant. Three main values: “Healthy”, “Moderate” and “Unhealthy”. Also has some noise.

“Degree”: The degree of the participant. Categorical attribute, has many different values and some noise as well.

“Have you ever had suicidal thoughts ?”: The attribute contains a “Yes/No” answer to this question. About equally distributed, no noise.

“Work/study hours”: A numeric attribute which shows how many hours the participant studies/works daily. All the answers were integers from 0 to 12. No noise.

“Financial stress”: Participants assessed how financially stressed they were on a scale of 1 to 5, 1 being the least and 5 the most stressed. No noise

“Family History of Mental Illness”: Participants reported whether there’s mental illness in their family. “Yes/No” answer, about equal distribution.

The attributes are relevant for our business goals.

It’s clear that there’s quite a lot of noise which needs to be cleaned. After that some attributes are already ready to be used for machine learning, some need to be converted to binary. Some

features should be one-hot encoded such as “Sleep Duration” and “Dietary Habits”. Other categorical attributes that had many different categorical values such as “Degree”, “Profession” and “City” we decided to target encode, because one-hot encoding them would lead to too many dimensions. It’s safe to assume that the “Name” attribute has no correlation with depression, so we decided to drop that immediately.

“”” *Copied from Kaggle*
The target variable, 'Depression', represents whether the individual is at risk of depression, marked as 'Yes' or 'No', based on their responses to lifestyle and demographic factors.
“””

It’s important to note that there’s much more people who are not at risk of depression, making this an unbalanced dataset.

After plotting every feature against “Depression” individually, it became clear that some attributes such as “Age” amd “Have you ever had suicidal thoughts ?” had a large correlation with “Depression” while others such as “Family History of Mental Illness” had almost no correlation. Even so, we decided to keep all remaining attributes.

Task 4. Planning your project (0.25 points)

1. **Business understanding.** Understanding why we are doing this task. Make it so that organisations would have a better understanding on how to fix the problem. **3h each**
2. **Data understanding.** There’s some existing information on Kaggle. Manually looking at the data and its features. Visualizing all attributes and plotting them against “Depression”. Making conclusions. **3h each**
3. **Data preparation.** Removing noise, encoding when necessary. **1h each**
4. **Early models.** Trying out the models covered in the course and googling some more that may be relevant for this binary classification task. **3h each**
5. **Optimizing models and data for Kaggle competition.** It quickly became clear that RFClassifier and XGBoost were the best models. We can use cross validation (and try other tactics) to tune hyperparameters and try to get the best public score. *By now, the competition is already over. **5h each**
6. **Understanding and using deep learning models.** Understanding the extra material we were given (and googling for more) and then trying to get a working model for our case. **5h each**
7. **Optimizing deep learning for Kaggle competition**. Trying some optimization tactics to try to get a good score in the Kaggle competition (late submissions). **5h each**
8. **Conclude and visualize results**. Find out what the most relevant features were that actually contributed to depression. See whether we accomplished our business goals. **4h each**
9. **Prepare for poster session.** Prepare a beautiful poster. **1h each**