PROJECT PROPOSAL

Group 10

1. Category of the project

Our project category falls into the application category as we will study an ongoing issue of healthcare.

2. Specific application area or solution methodology you intend to study.

Anxiety is one of the most common types of mental health disorders affecting 34% of the adult population in the USA (Szuhani & Simon, 2022). Every person may experience it in different ways and with different levels of severity which may be correlated to different factors. The busy and overwhelming lifestyle of a full time worker or student makes us forget the importance of taking care of ourselves and our mental health. Our goal is to find the most accurate, supervised machine learning model potentially using linear regressions and other methods, predicting the severity of anxiety attacks based on factors such as demographics, lifestyle, health indicators, and psychological aspects of each individual.

3. Plan for using or collecting data.

The data for our project has been extracted from Kaggle, a web platform that is widely used for machine learning projects due to its large collection of public datasets, competitions, and notebooks. The specific dataset we have borrowed for our project is titled “Anxiety Attack : Factors, Symptoms, and Severity”, published by Ashay Choudhary on 19/01/2025. This dataset contains 12.000 samples which is a large sample size that will help us make predictions with our machine learning model. It offers a comprehensive collection of features that impact on anxiety attacks and its possible severity.

4. Features of the data and Potential Risks

The chosen input data set contains a variety of features that are crucial for analyzing the severity of anxiety attacks. These features can be categorized into groups for future analysis, and will be examined to identify correlation and predictive patterns. The features are:

- Demographic Factors: includes attributes such as ag, gender and occupation. These attributes help us understand how the severity of anxiety attacks varies across different population segments.
- Symptoms: captures common symptoms associated with anxiety attacks such as rapid breathing rate, sweating. These features are essential for our model.
- Contributing Factors: encompasses lifestyle and environmental factors like stress levels, sleeping hours, social support or diet habits.

While working on the project, several potential risks may arise, which could determine the quality and outcomes of our analysis. Some of this risks are:

- Data quality issue: the input dataset may contain missing values, outliers or inconsistent samples. We can mitigate it by analyzing the input and preprocessing it.
- Overfitting / underfitting: overfitting may occur if the model learns noise or irrelevant data. Underfitting happens if the model is too simple to capture underlying patterns.
- Bias: the data set may be biased in terms of demographics. We must ensure a balanced input data.

By identifying all risks, analyzing them and implementing appropriate mitigation strategies, we improve the result of our project.

5. Plan of activities, including timelines.

For structuring our group's plan of activities and timeline, we looked at the tentative course schedule, ensuring that we use the most important lectures and tutorials to guide ourselves through each phase of the project. Below is our planned schedule:

1. **Early February (Weeks 1 and 2, after the proposal)**: We will retrieve our dataset from Kaggle, and begin cleaning, preprocessing and organizing the variables that will be later important. Regarding meetings, we will meet after our Thursday sessions to discuss the strategies to follow since we had just seen and studied the techniques.
2. **Mid February (Weeks 3 and 4)**: Using techniques introduced in class and tutorial sessions, we will investigate data distributions, correlations, and potential outliers. Although classes pause during reading week (Feb 17-21), we plan to maintain minimal progress remotely, preparing possible questions that could arise.
3. **Late February and Early March (Weeks 5 and 6)**: We will start introducing the supervised learning methods covered in lectures. Training of some preliminary models will start in order to see how the data performs. During this period, the midterm will take place. For this reason, the focus of the week will be on the midterm.
4. **Mid to late March (Weeks 7 and 8)**: During this period, we will continue using the learned techniques during class. For example, we will study the possibility of using cross-validation, tree-based methods or neural networks. We will also try to perform an evaluation of the models that we have by that time.
5. **Late March and early April (Weeks 9 and 10)**: During these last weeks, we will start finishing our studies, finding the best visualizations and preparing our presentation. Two or three days before the deadline, we will attempt to finish and polish the project so that we have time for any last-minute changes.

Throughout the project, we will continue having group meetings every week or two weeks. These meetings will take place on Thursdays after the tutorials. This schedule ensures we can apply the techniques learned during those classes (held on Thursdays) and lectures (held also on Tuesdays). Besides, we will fairly divide the tasks among the four team members. During the meetings, we will exchange feedback and gather ideas for each part of the project. With that, we aim to keep the entire project on track and promote collaboration.

6. Validation and comparisons

For the validation of the machine learning model, we have decided to follow the 80%-20% split in the dataset to create training and testing subsets. In addition, we will use k-fold cross-validation to prevent overfitting and reduce randomness from a single train-test split.

To ensure that we choose the best model, we will evaluate them using performance metrics such as accuracy, precision, recall, F1-score, the receiver operating characteristic (ROC), and the area under the curve (AUC). Once we calculate these metrics, we will conduct a comparative analysis to determine which model best predicts the severity of anxiety attacks.

References

- Szuhany, K. L., & Simon, N. M. (2022). Anxiety disorders. *JAMA*, *328*(24), 2431. Anxiety Disorders: A Review
- Anxiety Attack : Factors, Symptoms, and Severity