PROJECT WORK IN MACHINE LEARNING:
OSMI MENTAL HEALTH IN TECH PREDICTION

MASTER'S DEGREE IN ARTIFICIAL INTELLIGENCE

UNIVERSITY OF BOLOGNA

STEFANO POGGI CAVALLETTI

0000982439

DATASET

- OSMI Mental Health in Tech
  - collected by the Open Sourcing Mental Health corporation and available on Kaggle
  - measures the attitude and frequency towards mental health disorders in the context of tech workplace
  - aimed to understand whether any factor can affect the employee to get treatment or not

TASK

- Analyze the data and predict individual's mental health seek of treatment based on different features (e.g. age, gender, country and a variety of answers about their mental health related with work) through the deployment of machine learning models

CONTENT: 24 FEATURES (1)

- **Timestamp**
- **Age**
- **Gender**
- **Country state**: If you live in the United States, which state or territory do you live in?
- **self_employed**: Are you self-employed?
- **family_history**: Do you have a family history of mental illness?
- **treatment**: Have you sought treatment for a mental health condition? (Yes/No)
- **work_interfere**: If you have a mental health condition, do you feel that it interferes with your work?
- **no_employees**: How many employees does your company or organization have?
- **remote_work**: Do you work remotely (outside of an office) at least 50% of the time?
- **tech_company**: Is your employer primarily a tech company/organization?
- **benefits**: Does your employer provide mental health benefits?

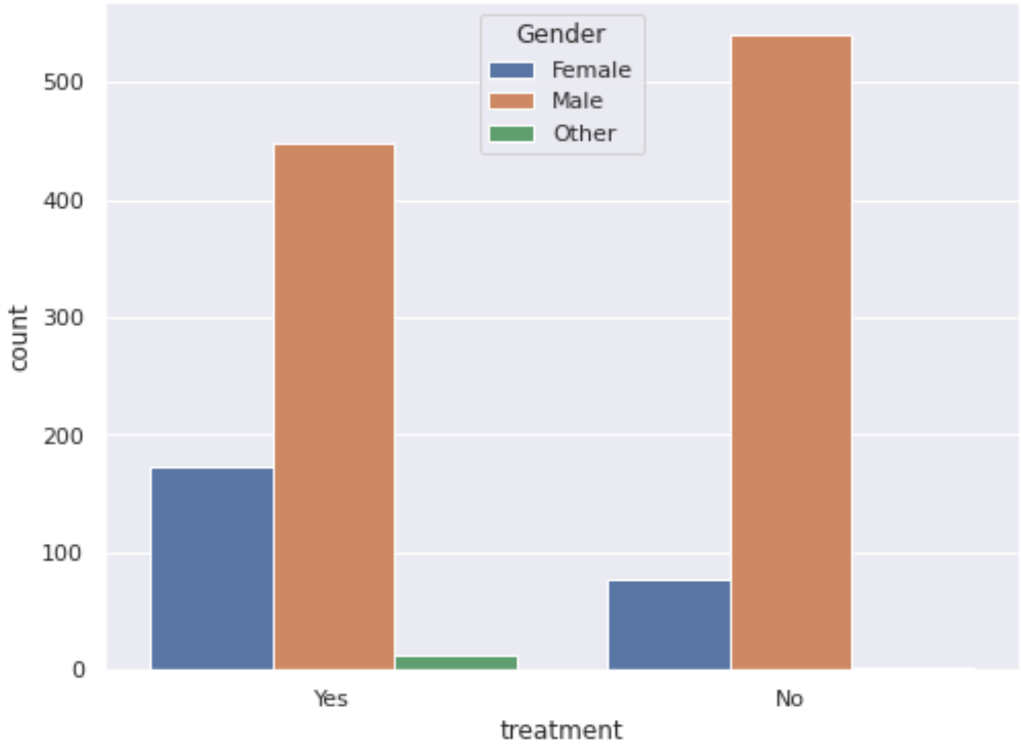
CONTENT: 24 FEATURES (II)

- **care_options**: Do you know the options for mental health care your employer provides?
- **wellness_program**: Has your employer ever discussed mental health as part of an employee wellness program?
- **seek_help**: Does your employer provide resources to learn more about mental health issues and how to seek help?
- **anonymity**: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
- **leave**: How easy is it for you to take medical leave for a mental health condition?
- **mental_health_consequence**: Do you think that discussing a mental health issue with your employer would have negative consequences?
- **phys_health_consequence**: Do you think that discussing a physical health issue with your employer would have negative consequences?
- **coworkers**: Would you be willing to discuss a mental health issue with your coworkers?
- **phys_health_interview**: Would you bring up a physical health issue with a potential employer in an interview?
- **mental_vs_physical**: Do you feel that your employer takes mental health as seriously as physical health?
- **obs_consequence**: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
- **comments**: Any additional notes or comments

DATA PRE-PROCESSING

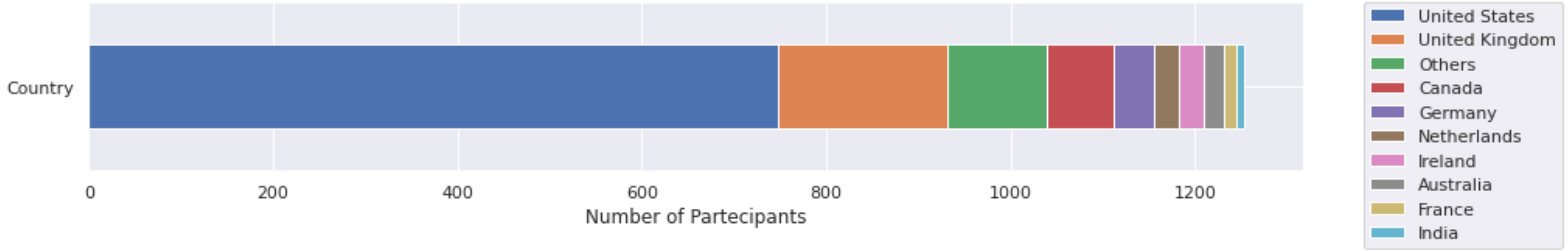
- Removal of NaN values and not useful features (comments, State, Timestamp) and replacement of empty values
- Data cleaning:
  - Encoding of gender values
  - Removal of meaningless age values (negative or too high)

EXPLORATORY DATA ANALYSIS



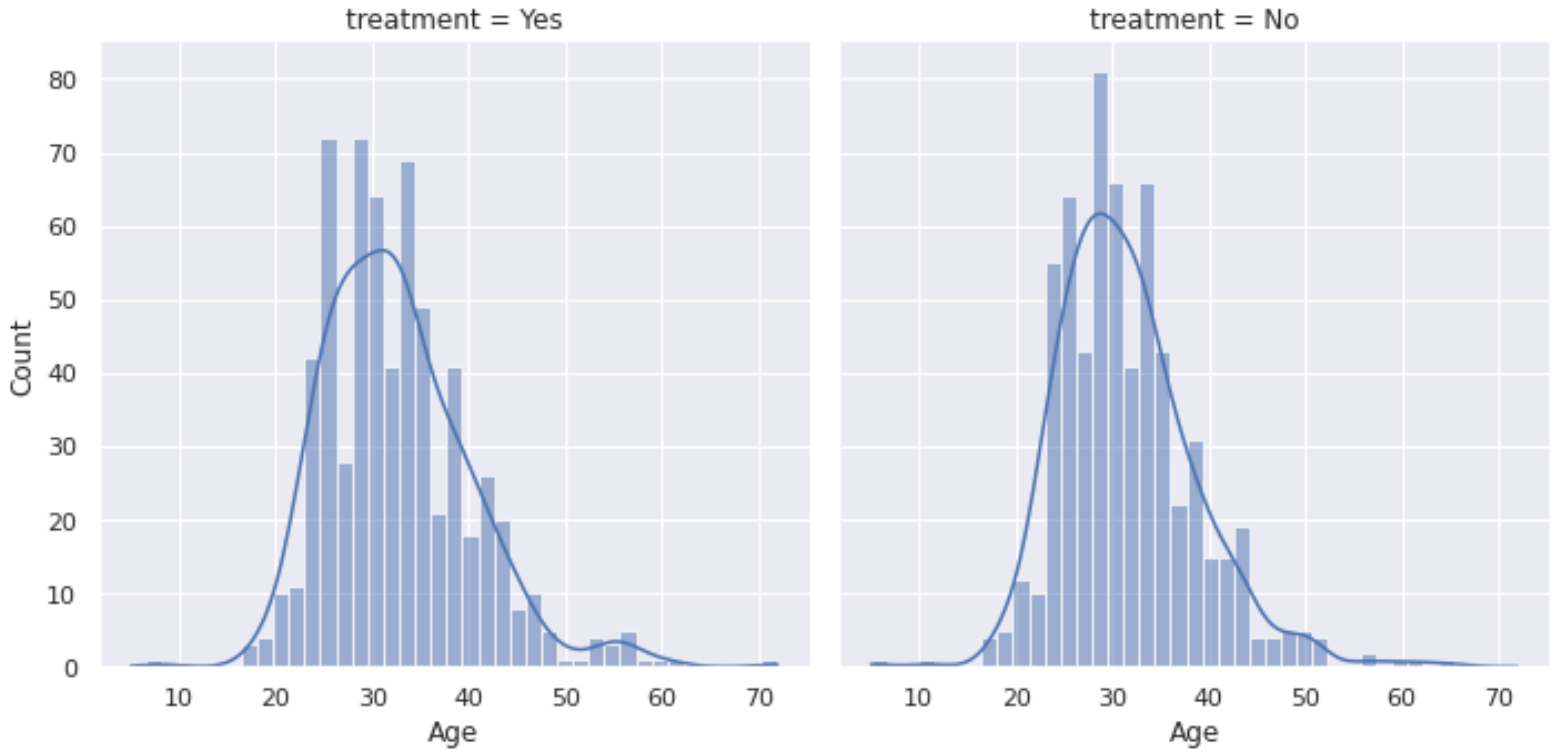
Relationship between gender and treatment: in the dataset, the number of male individuals is much higher compared to the other genders

EXPLORATORY DATA ANALYSIS



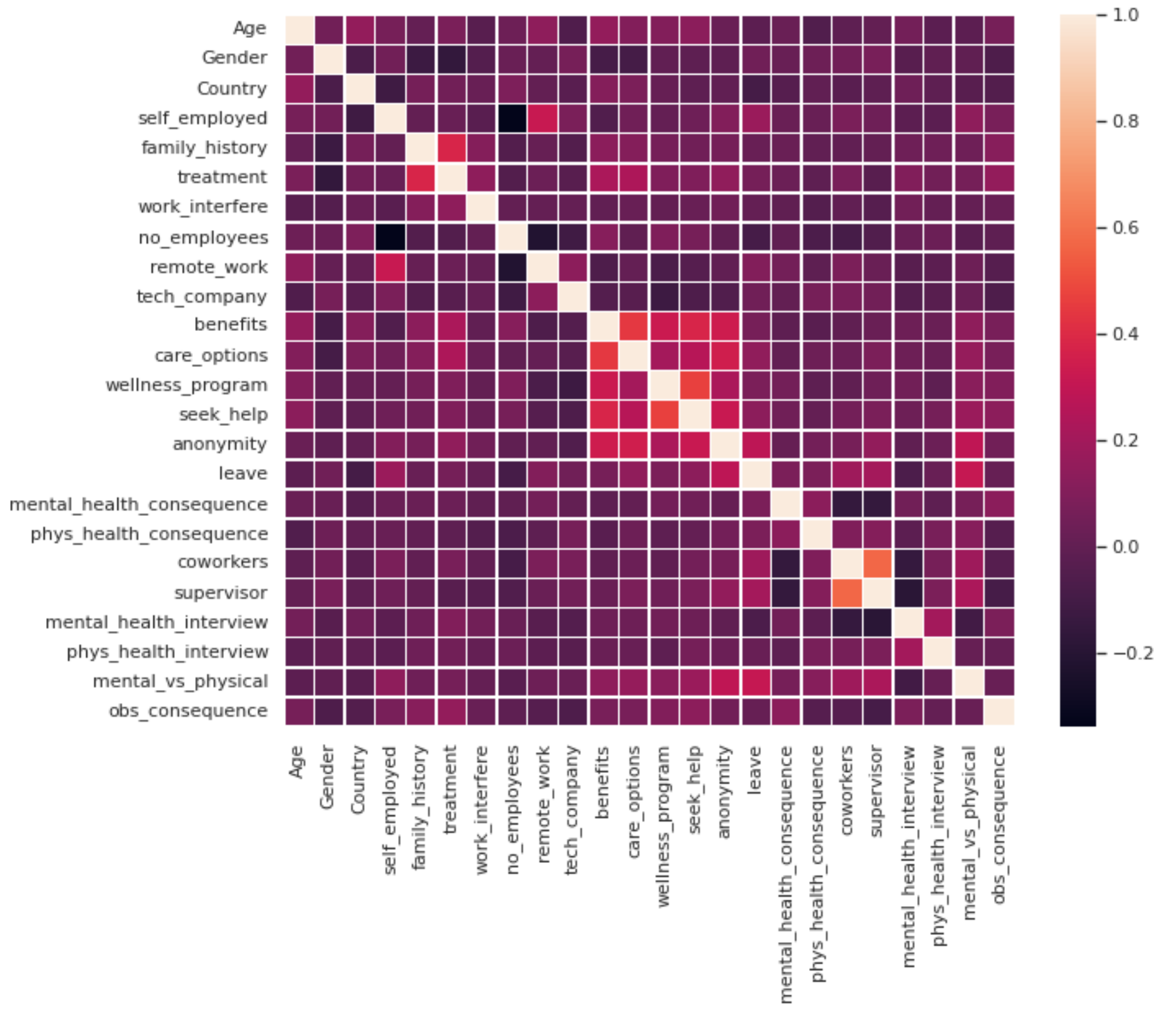
Distribution of individuals across the countries: the number of participants from the USA is much higher compared to any other country

EXPLORATORY DATA ANALYSIS



Relationship between age and treatment status among the respondants: peak around 30

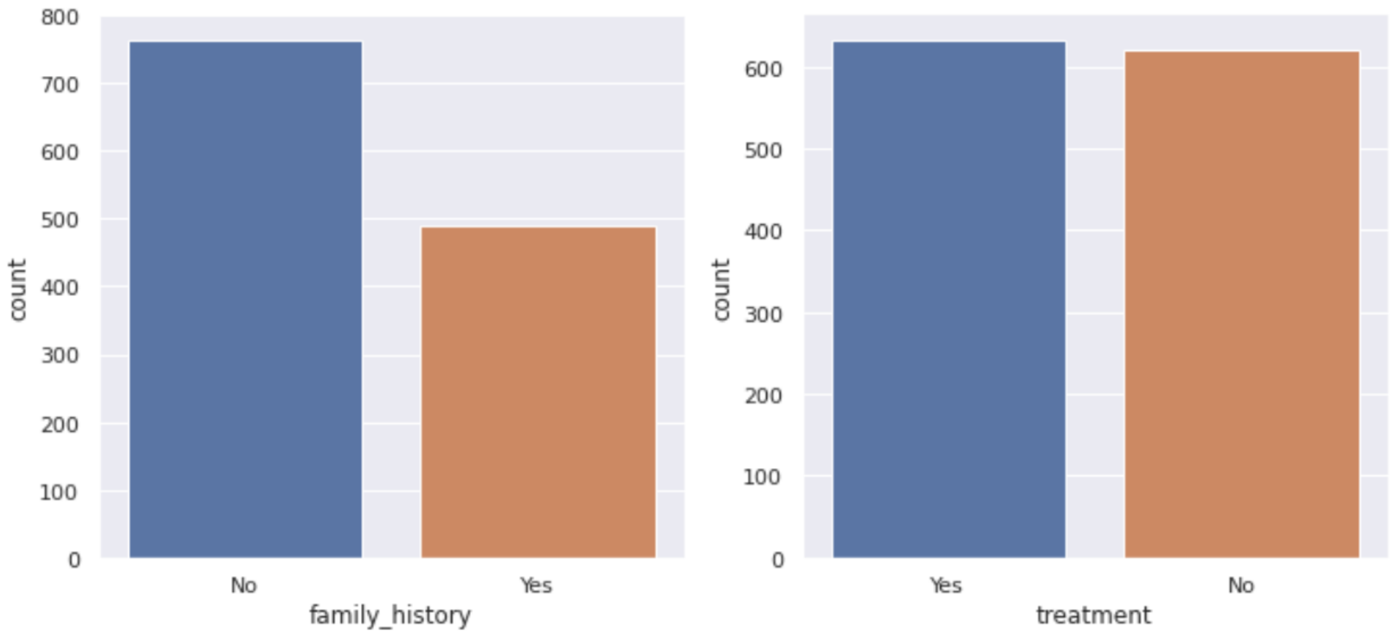
EXPLORATORY DATA ANALYSIS



- Correlation heatmap to see the correlation between variables and cause-effect relationships, computed using DataFrame.corr from Pandas library and sns.heatmap from seaborn

- Interesting correlation between family_history and seek of treatment

EXPLORATORY DATA ANALYSIS



- The percentages of respondents who want and do not want to get treatment are balanced and nearly equal to 50
- Fewer respondents claim to have a family history of mental illness but it is more likely that they want to get treatment compared to those without a family history, as they are more correlated

DATA ENCODING

- Features are encoded using LabelEncoder from sklearn.preprocessing, which encodes target labels with value between 0 and n_classes-1
- **Scaling**: the Age attribute is scaled to normalize values in the range [0,1] with the MinMax scaler
- Splitting dataset into train and validation set with train_test_split function

PREDICTION MODELS

Five models were used and tuned with their hyperparameters:

- Support Vector Machine
- Logistic Regressor
- K-Nearest Neighbor
- Random Forest
- XGB Classifier

PREDICTION MODELS

- Cross validation is applied on the defined models, tuning them for both accuracy and F1-score, while the best models are saved

- For each model, the best parameters found are printed in output, as well as their:
  - Accuracy
  - F1-score (harmonic mean of precision and recall)
  - ROC-Curve plot with AUC score (performance metric for classification problems: it gives an indication of how much the model is capable of distinguishing between classes: the higher the better the model is at predicting the right class)

RESULTS OF CLASSIFIERS (I)

1) Support Vector Machine

- Best parameters found: {'C': 10, 'cache_size': 8000, 'gamma': 'scale', 'kernel': 'rbf', 'probability': True}
- Accuracy: 0.71
- F1-score: 0.70
- ROC-AUC score: 0.78

2) Logistic Regressor

- Best parameters found: {'C': 3, 'max_iter': 100, 'solver': 'lbfgs'}
- Accuracy: 0.72
- F1-score: 0.71
- ROC-AUC score: 0.78

RESULTS OF CLASSIFIERS (II)

3) K-Nearest Neighbor

- Best parameters found: {'metric': 'manhattan', 'n_neighbors': 9}
- Accuracy: 0.68
- F1-score: 0.64
- ROC-AUC score: 0.73

4) Random Forest

- Best parameters found: {'max_depth': 25}
- Accuracy: 0.77
- F1-score: 0.77
- ROC-AUC score: 0.83

RESULTS OF CLASSIFIERS (III)

5) XGB Classifier

- Best parameters found: {'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 50}
- Accuracy: 0.77
- F1-score: 0.78
- ROC-AUC score: 0.84

CONCLUSIONS

- The results of the predictions show that the best performing models in terms of accuracy, F1-score and AUC score were:
  - **Random Forest**
  - **XGB Classifier**

- On the other hand, the model which showed the worst results was K-Nearest Neighbors