Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	738214
Project Title	Predicting Mental Health Illness Of Working Professionals Using Machine Learning.
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

- Implement the chosen resolution plans to address data quality issues.
- Explore data distributions (histograms, boxplots) to understand variable ranges and potential skewness. This can help identify outliers and inform normalization decisions.
- Perform feature engineering if needed (e.g., creating new features from existing ones, encoding categorical variables). Consider one-hot encoding for categorical features with many categories.
- Standardize or normalize numerical features (if necessary) to ensure all features contribute equally to the model. Standardization scales features to have a mean of 0 and standard deviation of 1, while normalization scales features to a range of 0 to 1. The choice depends on the algorithm's assumptions.
- Split the data into training and testing sets for model development and evaluation. A common split I is 80% for training and 20% for testing, but this can vary depending on the dataset size.

Goal: Build a system to predict mental health needs based on user input using machine learning.

Data: Analyze survey data on demographics, work environment, and mental health experiences.

Tools: Python libraries (scikit-learn, pandas, NumPy, Matplotlib/Seaborn ,Flask)

Section	Description
Data Overview	**Structure:** • Likely stored in a tabular format (CSV) named "survey.csv". • Each row represents a participant in the survey. • Each column represents a specific question or variable asked in the survey.

	Additional Notes: • The data likely includes a mix of data types (string, boolean, integer). • The dataset size is relatively small (around 303.68 KB).
Univariate Analysis	• **Access the Data:** If you have the "survey.csv" file, you can use Python libraries like pandas to load the data and perform univariate analysis. • **Explore Existing Analysis:** Look for resources that might have already analyzed the data. The original source (Open Sourcing Mental Illness) or related research papers might provide some insights into individual variable explorations. • **Focus on Descriptive Statistics:** You can analyze the data types from the description and identify categorical vs. numerical features. This offers a basic understanding of the data structure, even without specific values.
Bivariate Analysis	· No exact info on number of participants or basic statistics (mean, median) available. · Data likely stored in a CSV file with around 27 features (questions) per person. · Features likely include a mix of text answers (e.g., country), yes/no options (e.g., self-employed), and numbers (e.g., age). · You can't do a full analysis of individual variables (mean, median) or relationships between two variables (correlation) without the actual data.
Multivariate Analysis	• Full analysis involving multiple variables at once (multivariate analysis) requires the actual data. • Multivariate analysis is valuable because mental health is influenced by many factors, not just one or two. • Machine learning algorithms can be used for this analysis once you have the data. • For now, you can explore existing research on similar data analysis or brainstorm potential relationships between multiple variables based on the data description.

<table>
<tr><td>Outliers and Anomalies</td><td>

- Identification and treatment of outliers.**Skewed Models:** Outliers can significantly influence machine learning models, leading to inaccurate predictions.
- **Data Errors:** They might indicate data entry errors or inconsistencies requiring investigation.
- **Genuine Rarities:** Sometimes, outliers represent genuine but rare cases that shouldn't be removed without justification.

Identifying Outliers:

- **Visualizations:** Techniques like boxplots and scatter plots can help visually identify data points that fall far from the main cluster.
- **Statistical Methods:** Techniques like calculating standard deviation or using Interquartile Range (IQR) can help define thresholds for outliers.

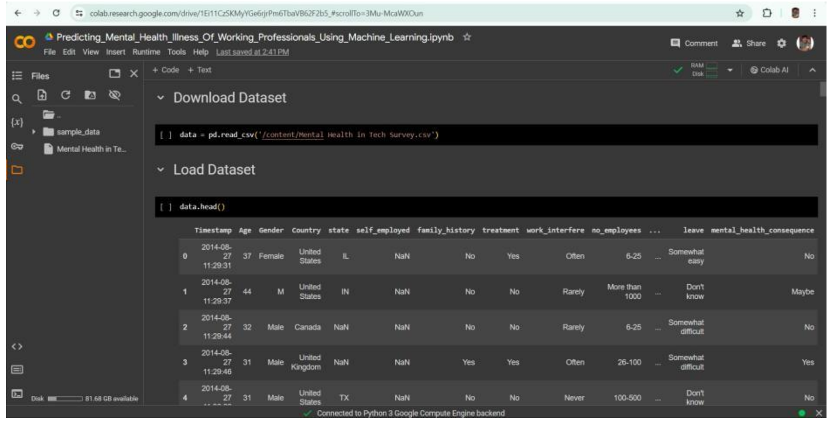
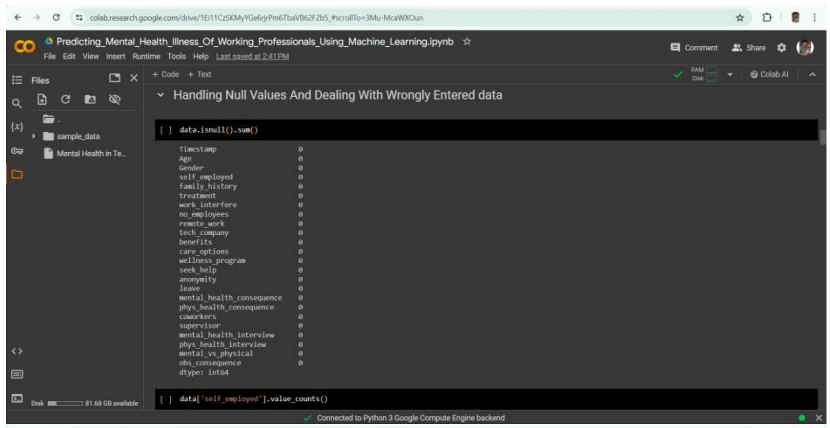
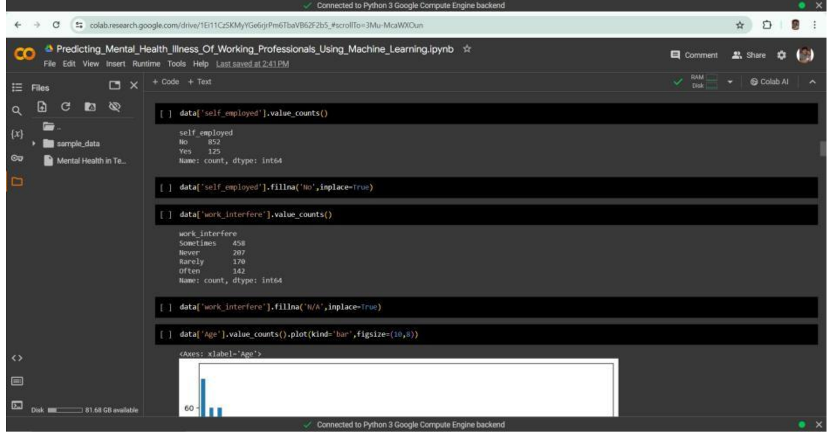
Treatment of Outliers:

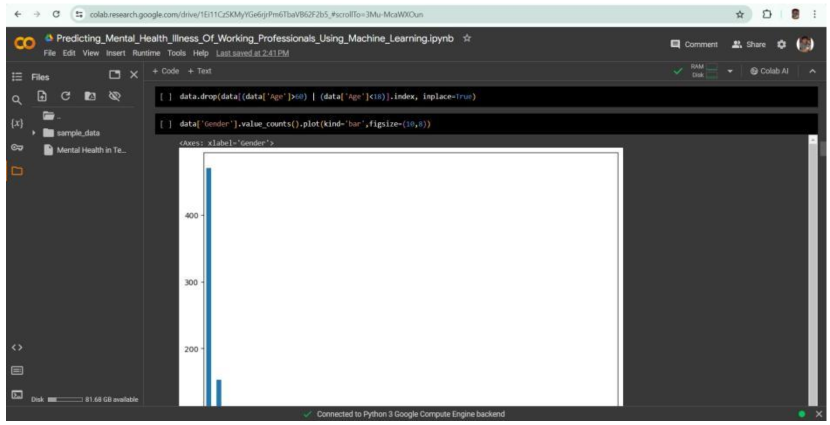
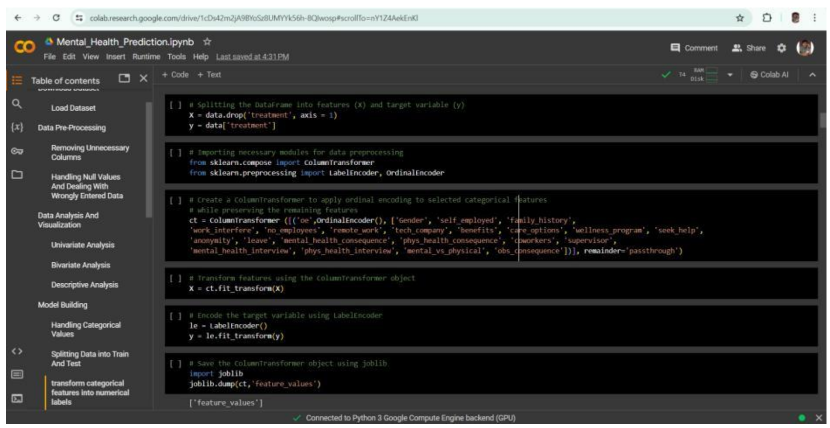
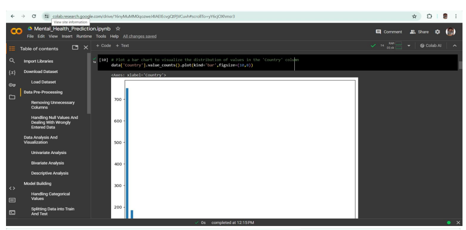
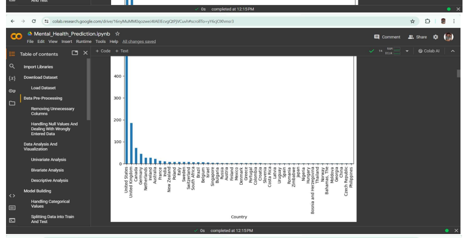
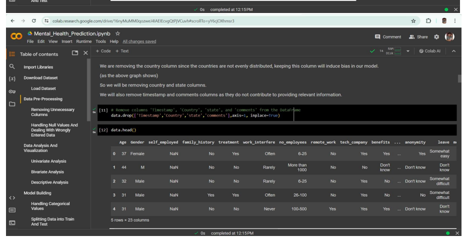
- **Investigate:** First, try to understand why the outlier exists. Is it a genuine case or a potential error?
- **Winsorization:** This technique caps outliers to a specific value within the distribution, preserving their influence but reducing their impact.
- **Removal (Caution):** Removing outliers should be a last resort, especially if they represent genuine but rare cases. Only remove them if you're confident they are errors.

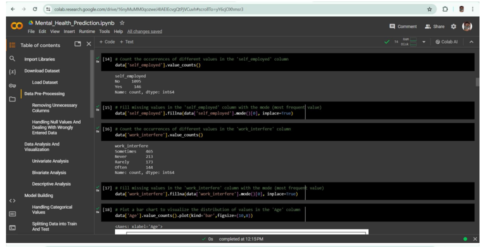
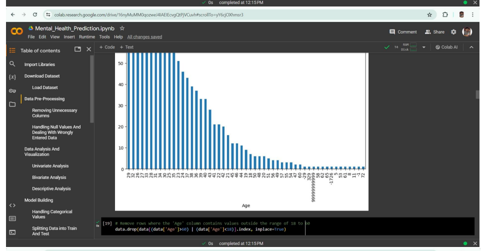
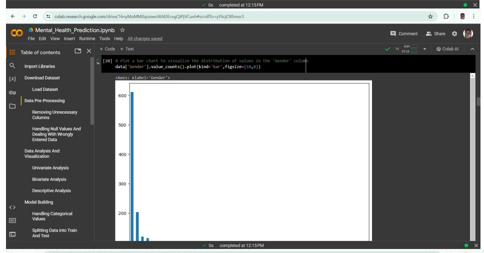
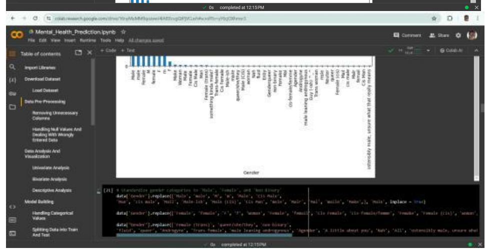
Considerations for Mental Health Data:

- **Sensitivity:** Mental health data can be sensitive. Removing outliers needs careful justification to avoid excluding potentially vulnerable populations.
- **Domain Knowledge:** Understanding the context and expected ranges for each variable is crucial for identifying meaningful outliers.

Remember: There's no one-size-fits-all approach for outliers. The strategy depends on the specific data and the chosen model.
</td></tr>
<tr><td colspan="2">Data Preprocessing Code Screenshots</td></tr>
</table>

Loading Data	
Handling Missing Data	

<table>
<tr>
<td></td>
<td>
</td>
</tr>
<tr>
<td>Data Transformation</td>
<td>
</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>

</td>
</tr>
</table>

	   
Save Processed Data