## Artificial Intelligence

## and Machine Learning

Project Report

Semester-IV (Batch-2022)

INNER ECHO

A red and white sign

Description automatically generated with low confidence

**Submitted To: Submitted By:**

Ms. Shagun Sharma Harshit Kukreja, 2210990387 (G9)

Harshit Singla, 2210990391 (G9)

Harshpreet Singh, 2210990393 (G9)

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

**INDEX**

|  |  |  |
| --- | --- | --- |
| Serial No. | Title | Page No. |
| 1 | Introduction | 1-6 |
| 2 | Problem Definition and Requirements | 6 |
| 3 | Materials and Methods | 8-13 |
| 4 | Data Analysis and Methodology | 14-35 |
| 5 | Evaluation Metrics | 35-37 |
| 6 | Inner Echo Frontend Result | 38-39 |

1. **Introduction:**

**Introduction to Depression:**

Depression is a common and serious mental health disorder characterized by persistent sadness, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, feelings of tiredness, and poor concentration. It is a significant public health concern, affecting millions of people worldwide. According to the World Health Organization (WHO), depression is one of the leading causes of disability globally, impacting individuals' quality of life, productivity, and overall well-being.

Nearly three in ten adults (29%) have been diagnosed with depression at some point in their lives and about 18% are currently experiencing depression, according to a 2023 [national survey](https://news.gallup.com/poll/505745/depression-rates-reach-new-highs.aspx). Women are more likely than men and younger adults are more likely than older adults to [experience depression](https://www.nimh.nih.gov/health/statistics/major-depression). While depression can occur at any time and at any age, on average it can first appear during one’s late teens to mid-20s.

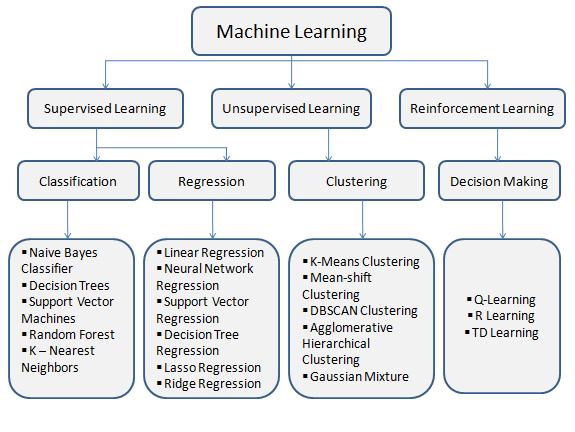
Early detection and intervention are crucial for managing depression effectively and mitigating its adverse consequences. However, traditional methods of diagnosing depression often rely on subjective assessments and self-reporting, which can be influenced by various factors, such as social stigma, lack of awareness, or difficulty in recognizing and articulating symptoms.



**1.1 Background:**

The work starts its roots in understanding depression and its types which is prevailing in this world. Peoples are defying reality and consciously living in the unreal, imaginary, hallucinatory state of mind and lost hope and faith that leads to the identification of depression and ends with taking their imaginary life not to be true. This problem leads to suicide and is faced by one-third of the population in the world. Another state of the awareness might be true but from the analysis of the human mind this state of mind keeps on fluctuating, but these roots progress rapidly which in the case is the lack of reality of the external world which is frequently termed as depression in a medical encyclopedia.

Machines have been an integral part of life in today's world, especially computers. The Association of Computing Machinery [ACM] definition of computing is any goal-oriented activity, which requires, benefits from computers or creating computers. This definition is the accepted direction of machines and their computations. Machine Learning [ML] is a technique to analyze data that automates the building of the Analytical models. ML is born from the pattern recognition and with the idea that computers can learn from data, can identify patterns and make informed decisions by the theory stating, computers can learn anything without being explicitly being programmed to perform specific tasks such as Early risk of depression detection with improved performance over time. There are mainly two approaches in machine learning such as Supervised Learning [SL] and Unsupervised Learning [UL], but recently Semi-Supervised Learning [SSL] and Reinforcement learning have emerged.



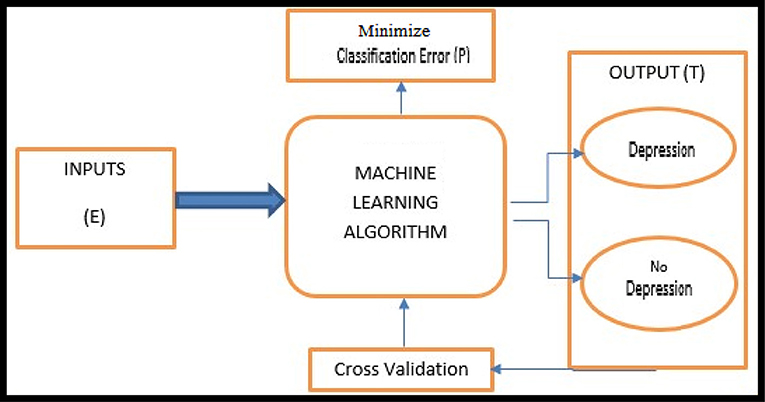
**Traditional Methods of Depression Diagnosis:**

Traditionally, depression has been diagnosed through clinical interviews and self-report questionnaires, such as the Beck Depression Inventory (BDI) or the Patient Health Questionnaire (PHQ-9). These methods heavily rely on individuals' subjective reports and clinicians' expertise in interpreting the responses. While valuable, these approaches have limitations, including potential biases, inconsistencies, and limited scalability for widespread screening and monitoring.

Motivation for Machine Learning Approach

Machine learning techniques have shown great potential in addressing complex problems across various domains, including healthcare. In the context of depression detection, machine learning offers several advantages over traditional methods:

1. **Objectivity**: Machine learning models can analyze data objectively, reducing potential biases and inconsistencies associated with human assessments.
2. **Scalability**: These models can process large amounts of data efficiently, enabling widespread screening and monitoring of depression at a population level.
3. **Multimodal Data Integration**: Machine learning algorithms can integrate and analyze diverse data sources, such as text, audio, visual cues, and physiological signals, potentially improving the accuracy and robustness of depression detection.
4. **Personalization**: By leveraging individual-level data, machine learning models can enable personalized depression detection and monitoring, tailoring interventions to specific needs and characteristics.



* 1. **Objectives:**

1. **Data Acquisition and Preprocessing**: To acquire and preprocess a comprehensive multimodal dataset for depression detection, addressing challenges such as data quality, label noise, and potential biases. Robust data preprocessing techniques will be employed to handle missing data, remove noise, and ensure consistency across modalities.
2. **Accurate Depression Detection**: The primary objective is to develop a highly accurate machine learning model capable of detecting depression in individuals based on various data inputs. The model should exhibit high sensitivity and specificity in identifying individuals who may be experiencing depression.
3. **Early Identification and Intervention**: Model aims to enable early identification of depression symptoms, allowing for timely intervention and support. Early detection can play a crucial role in mitigating the potential negative impacts of depression and facilitating effective treatment and management.
4. **Personalized Assessment and Monitoring**: The model aims to adapt to individual characteristics, patterns, and contexts, providing tailored insights and recommendations for better understanding and managing the condition.
5. **User-friendly and Accessible Interface**: Design an intuitive and user-friendly interface that encourages engagement and adoption among users. The software should be accessible and user-friendly, ensuring a seamless experience for individuals seeking depression screening or monitoring.
6. **Interpretability and Transparency**: Model must be interpretable and transparent that can provide insights into the factors contributing to depression detection. This transparency can help build trust with users and healthcare professionals, enabling better understanding and informed decision-making.
7. **Continuous Improvement and Adaptation**: Implement mechanisms for continuous learning and adaptation of the machine learning models. As more data becomes available or new patterns emerge, the software should be capable of updating and improving its performance over time.
8. **Comprehensive Evaluation and Benchmarking**: To conduct rigorous evaluation and benchmarking of the developed models using appropriate metrics, benchmarks, and cross-validation techniques. This objective involves assessing the performance of the models on held-out test sets, comparing them with existing state-of-the-art approaches, and validating the models' performance across diverse demographic and clinical subgroups.

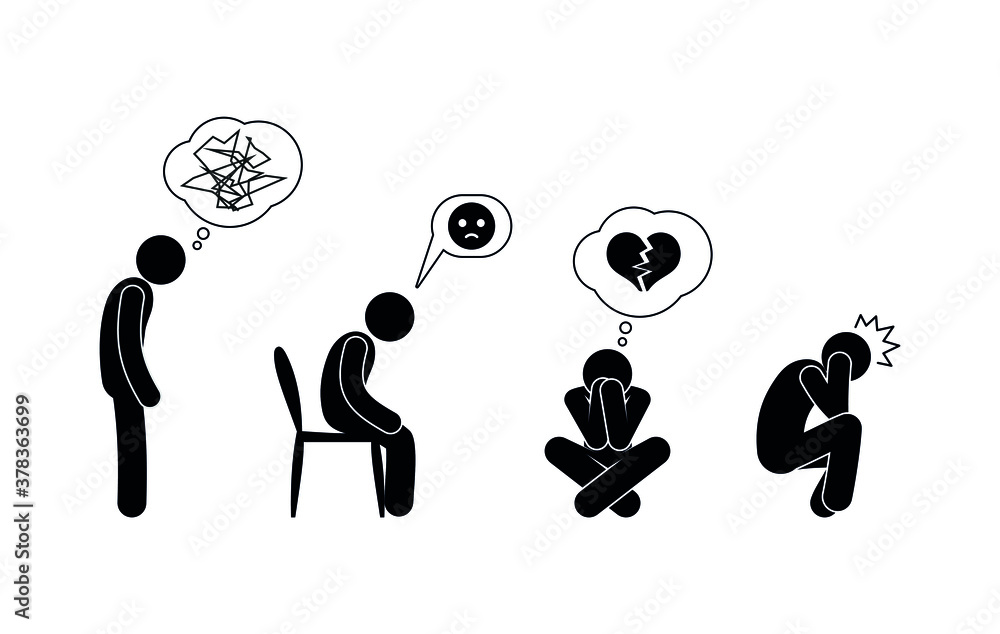
**1.3 Significance:**

The development of a machine learning depression detection software holds significant importance and potential impact in various domains. Here are some key points highlighting the significance of such software:

1. **Improved Mental Health Outcomes**: Early and accurate detection of depression can lead to timely intervention and treatment, which is crucial for managing the condition effectively. The software can play a pivotal role in identifying individuals at risk or experiencing depression, enabling them to seek appropriate support and care, potentially mitigating the adverse effects of depression on their well-being and daily functioning.
2. **Destigmatization and Accessibility**: The software can provide a discreet and accessible way for individuals to assess their mental health without the stigma or barriers often associated with traditional diagnostic methods. By offering a user-friendly and non-judgmental platform, the software can encourage more people to seek help, overcoming the reluctance or hesitation that may arise from societal stigma or personal reservations.
3. **Personalized and Tailored Support**: By leveraging individual-level data and adapting to personal characteristics and patterns, the software can provide personalized assessment and monitoring of depression. This tailored approach can lead to more effective interventions and support strategies tailored to individual needs and circumstances.
4. **Research and Data-driven Insights**: The data collected and analyzed by the software can contribute to advancing research in the field of depression and mental health. Researchers can leverage this rich dataset to gain insights into patterns, risk factors, and potential interventions, ultimately informing more effective strategies for prevention, diagnosis, and treatment.
5. **Scalability and Cost-effectiveness**: Machine learning models can process large amounts of data efficiently, enabling widespread screening and monitoring of depression at a population level. This scalability can be particularly beneficial in regions or communities with limited access to mental health professionals, making depression detection more accessible and cost-effective.
6. **Addressing Mental Health Challenges**: Depression is a significant public health concern with far-reaching social and economic implications. By addressing this challenge through innovative technological solutions like machine learning software, society can work towards mitigating the burden of depression and promoting overall well-being.

**Target Users:**

The survey form is distributed to various categories of the users such as General Public including students, The Doctors, and Mental Health professionals and Information Technology Professionals to gain insights for future improvements as well as obtaining new reflections and know the best value use case of the system. The primary users would be the Doctors and Mental health professionals who can use for automating their productivity. The Secondary users are the General Public including students who can gather facts about their current mental state and the third category includes the Information technology professionals who can refer to this system as for learning and optimizing purposes. The data will be analyzed, and improvements will be made based on their feedback and opinions. The Questionnaire is to prepare among 50 persons who are interested in Computational Psychiatry.



1. **Problem Definition and Requirements:**

**Problem Statement:**

The goal of this project is to develop a machine learning model that can accurately predict depression levels in individuals based on their responses to the Depression Anxiety Stress Scales (DASS) survey, demographic information, personality traits, and other relevant features. The model should be able to analyze the complex relationships between factors such as DASS responses, age, gender, education level, marital status, personality characteristics, and any other potentially predictive variables available in the dataset. The ultimate objective is to create a reliable tool that can assist in identifying individuals who may be at risk for depression or experiencing depressive symptoms, enabling early intervention and support.

**Requirements:**

1. Python Programming Language: The project will be implemented using Python, which is a popular language for machine learning and data analysis tasks.
2. Python Integrated Development Environment (IDE):

* PyCharm: PyCharm is a powerful IDE developed by JetBrains, which provides excellent support for Python development, debugging, and integration with various libraries and frameworks.
* Jupyter Notebook: You also mentioned using Jupyter Notebook, which is a web-based interactive computing environment that allows you to combine code, visualizations, and narrative text.

1. Python Libraries and Frameworks:

* NumPy: For numerical computing operations.
* Pandas: For data manipulation and analysis.
* Scikit-learn: A machine learning library for Python, providing a wide range of algorithms and tools for model building, evaluation, and deployment.
* Matplotlib and Seaborn: For data visualization and plotting.

1. Version Control System: Git: A distributed version control system widely used for tracking changes in source code and collaborating on projects.
2. **Materials and Methods:**

This section outlines the methodology employed in this study to detect depression using machine learning techniques. The methodology encompasses data collection, preprocessing, feature extraction, and the selection and training of machine learning algorithms.

* 1. **About The Dataset:**

This data was collected with an on-line version of the Depression Anxiety Stress Scales (DASS), see http://www2.psy.unsw.edu.au/dass/

The survey was open to anyone and people were motivated to take it to get personalized results. At the end of the test they also were given the option to complete a short research survey. This datatset comes from those who agreed to complete the research survey and answered yes to the question "Have you given accurate answers and may they be used for research?" at the end.

This data was collected 2017 - 2019.

The following items were included in the survey:

Q1 I found myself getting upset by quite trivial things.

Q2 I was aware of dryness of my mouth.

Q3 I couldn't seem to experience any positive feeling at all.

Q4 I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).

Q5 I just couldn&#39;t seem to get going.

And many more…

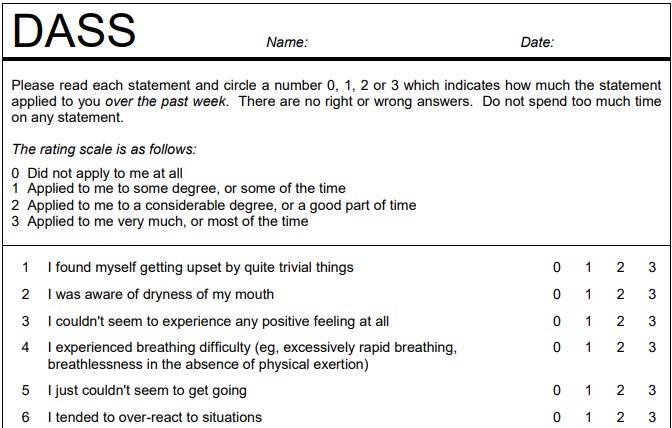
Each item was presented one at a time in a random order for each new participant along with a 4-point rating scale asking the user to indicate how often that had been true of them in the past week where

1 = Did not apply to me at all

2 = Applied to me to some degree, or some of the time

3 = Applied to me to a considerable degree, or a good part of the time

4 = Applied to me very much, or most of the time



* 1. **Handling Missing Values and Relabelling Categorical Features**

**Education Feature**:

* Mapped numerical codes (0, 1, 2, 3, 4) to new labels (1, 2, 3, 4).
* Converted numerical labels to descriptive strings (e.g., "Less than high school", "University degree").

**Urban Feature**:

* Mapped existing codes (0, 1, 2, 3) to new labels (3, 1, 2, 3), treating 0 as "Urban".
* Converted numerical labels to descriptive strings ("Rural", "Suburban", "Urban").

**Gender Feature**:

* Mapped codes (0, 1, 2, 3) to new labels (2, 1, 2, 3), treating 0 as "Female".
* Converted numerical labels to descriptive strings ("Male", "Female", "Other").

**Religion Feature**:

* Mapped 0 to 12 (treated as "Other").
* Converted numerical labels to descriptive strings for various religions.

**Race Feature**:

* Divided existing codes (10, 20, 30, ...) by 10 to create new labels (1, 2, 3, ...).
* Converted numerical labels to descriptive strings ("Asian", "Arab", "Black", ...).

**TIPI Features (TIPI1 to TIPI10)**:

* Replaced missing value (0) with the most common value in each TIPI feature.
* Converted numerical labels to descriptive strings for personality traits (e.g., "Disagree strongly", "Agree moderately").
  1. **Removing Irrelevant or Redundant Features**

Several features were identified as irrelevant or redundant for the depression detection task and were removed from the dataset. The following features were dropped:

* Response time features (Q\*E)
* Question order features (Q\*I)
* Vocabulary check features (VCL\*)
* Other extraneous features like source, elapse, engnat, hand, orientation, voted, country, screensize, uniquenetworklocation.

This step aimed to simplify the dataset by retaining only the most informative and relevant features for predicting depression levels, potentially improving model performance and interpretability.

* 1. **Feature Engineering:**

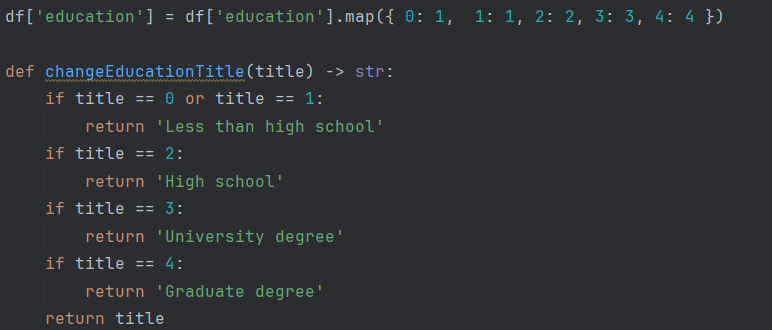
Feature Engineering is the process of creating new features or transforming existing features to improve the performance of a machine-learning model. It involves selecting relevant information from raw data and transforming it into a format that can be easily understood by a model. The goal is to improve model accuracy by providing more meaningful and relevant information.

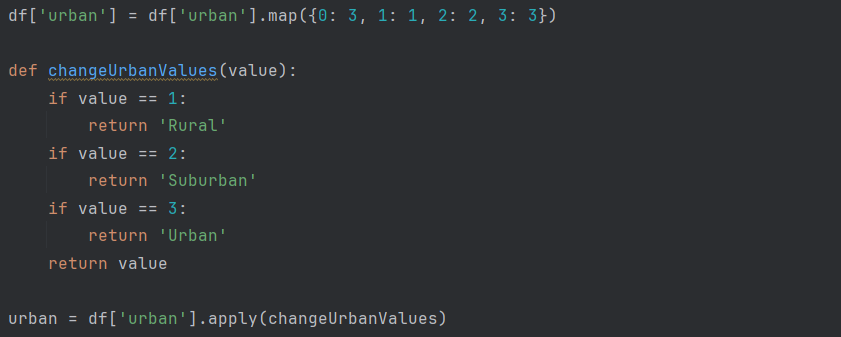
**Age Feature Engineering**:

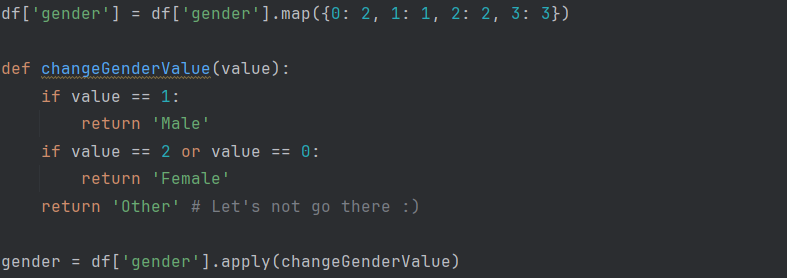
* Created a new 'age' column with descriptive age group labels (e.g., 'Under 10', 'Primary Children', 'Secondary Children', 'Adults', 'Elder Adults', 'Older People') using the makeAgeGroup function.
* Created a numerical 'age\_group' feature by mapping the age values to numerical codes (1-6) using the makeAgeGroupFeature function.
* Dropped the original 'age' column as the new 'age\_group' feature was created.

**Total Count Feature**:

* Created a 'total\_count' feature by summing all the columns for each row, representing the overall score or count.
* Analyzed the distribution of 'total\_count' and divided it into five categories: 'Normal', 'Mild', 'Moderate', 'Severe', and 'Extremely Severe' based on quantile values.
* Created a 'target' column by mapping the 'total\_count' values to the respective category labels using the buildTarget function.
* To address data imbalance, adjusted the category boundaries by moving 10 steps in the buildTargetMove10Steps function and updated the 'target' column.
  1. **Key Features:**
     + 1. **DASS (Depression Anxiety Stress Scales) Responses**:
* Q1 to Q42: Responses to the 42 DASS questions, which are likely the most important features for detecting depression levels.
  + - 1. **Demographic Features**:
* education: Education level (mapped to descriptive labels).
* urban: Area type lived in during childhood (Rural, Suburban, Urban).
* gender: Gender (mapped to descriptive labels).
* religion: Religion (mapped to descriptive labels).
* race: Race (mapped to descriptive labels).
* age\_group: Age group (numerical codes).

****

****

****





* + - 1. **Personality Traits**:

TIPI1 to TIPI10: Responses to the Ten Item Personality Inventory (TIPI) questions, which measure different personality traits (e.g., extroversion, agreeableness, conscientiousness).

TIPI1 Extraverted, enthusiastic. TIPI2 Critical, quarrelsome.

TIPI3 Dependable, self-disciplined. TIPI4 Anxious, easily upset.

TIPI5 Open to new experiences, complex. TIPI6 Reserved, quiet.

TIPI7 Sympathetic, warm. TIPI8 Disorganized, careless.

TIPI9 Calm, emotionally stable. TIPI10 Conventional, uncreative.

The TIPI items were rated "I see myself as:" \_\_\_\_\_ such that

1 = Disagree strongly 2 = Disagree moderately

3 = Disagree a little 4 = Neither agree nor disagree

5 = Agree a little 6 = Agree moderately

7 = Agree strongly



1. **Data Analysis & Methodology:**

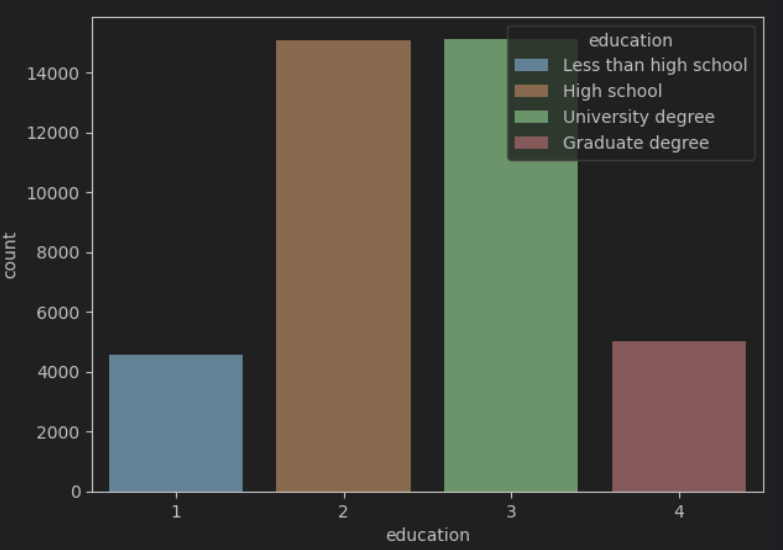
**Exploratory Data Analysis:**

It involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.

**Data Distributions:**

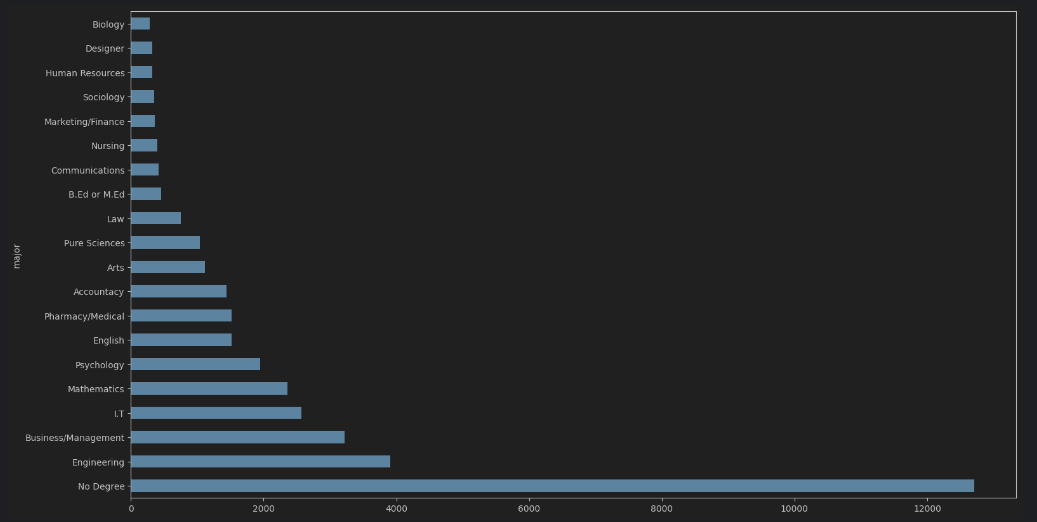
* 1. **Education:** The dataset seems to capture a diverse range of education levels, from less than high school to graduate degrees.

The higher representation of high school and university degree levels could be useful for analyzing potential differences in depression, anxiety, or stress levels across these common education groups.

****

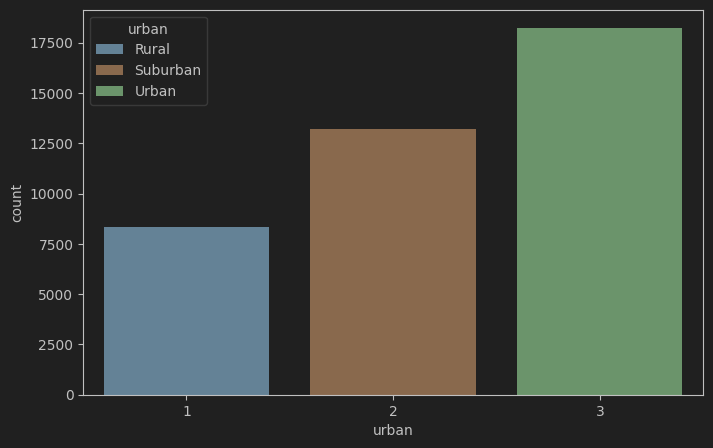
* 1. **Majors:** The dataset seems to capture a diverse range of majors, including STEM fields, social sciences, humanities, and professional degrees.

The "No Degree" category has the highest count, indicating a significant portion of the participants did not have a degree or major.



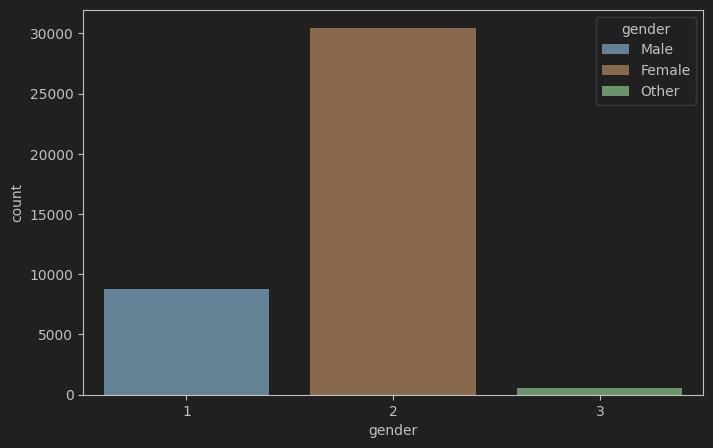
* 1. **Location:** The dataset captures participants from diverse geographic areas, including urban, suburban, and rural regions.

The higher representation of urban and suburban areas could be useful for analyzing potential differences in depression, anxiety, or stress levels among these population segments.

****

* 1. **Gender: The** dataset captures participants from all three gender categories: Male, Female, and Other.

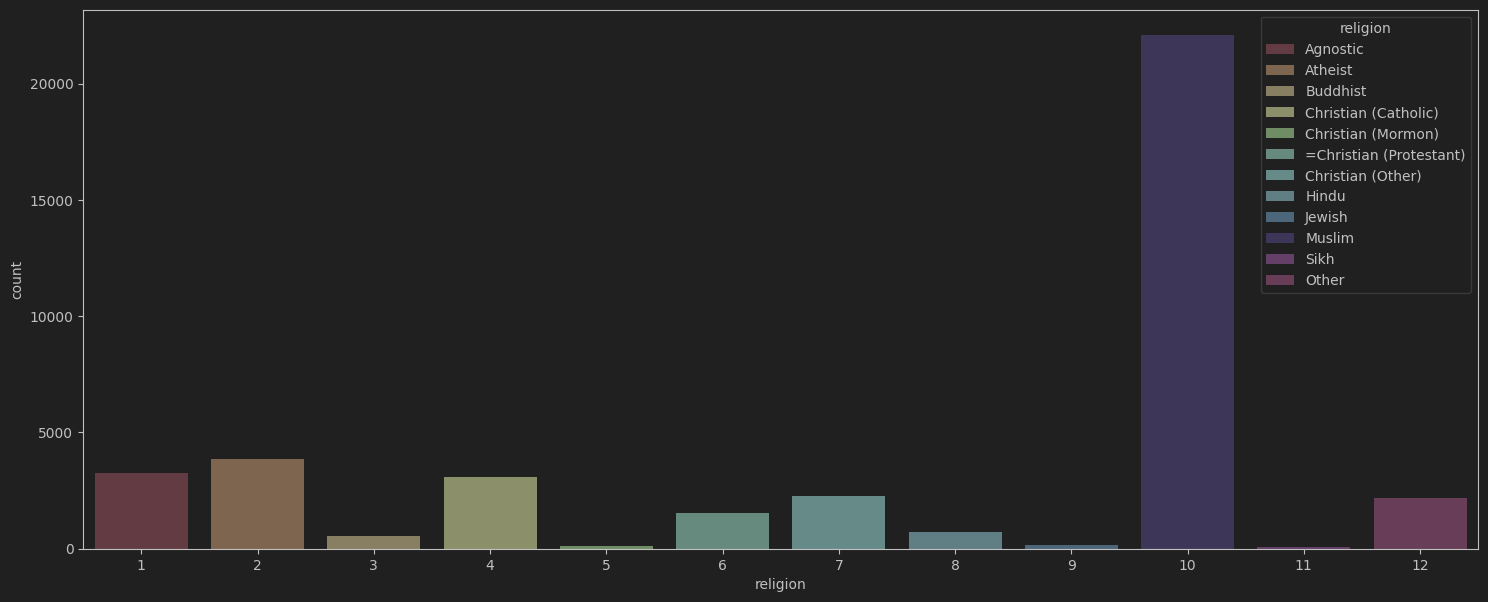
The higher representation of female participants could be useful for analyzing potential gender differences in depression, anxiety, or stress levels.



* 1. **Religion:** dataset captures participants from diverse religious backgrounds, including major world religions and non-religious affiliations.

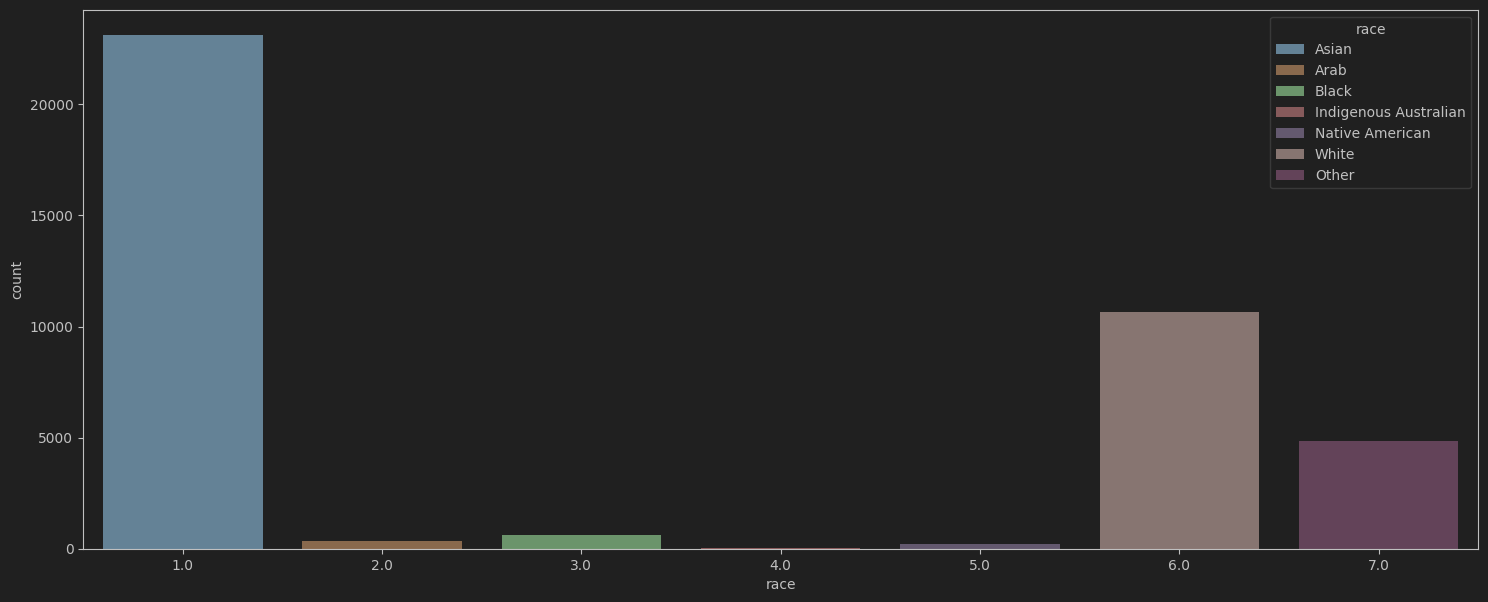
The Muslim category has the highest count, represented by the tallest bar.

Other religious affiliations like Buddhist, Christian (Catholic, Mormon, Protestant, Other), Hindu, Jewish, Sikh, and Other have relatively lower counts.

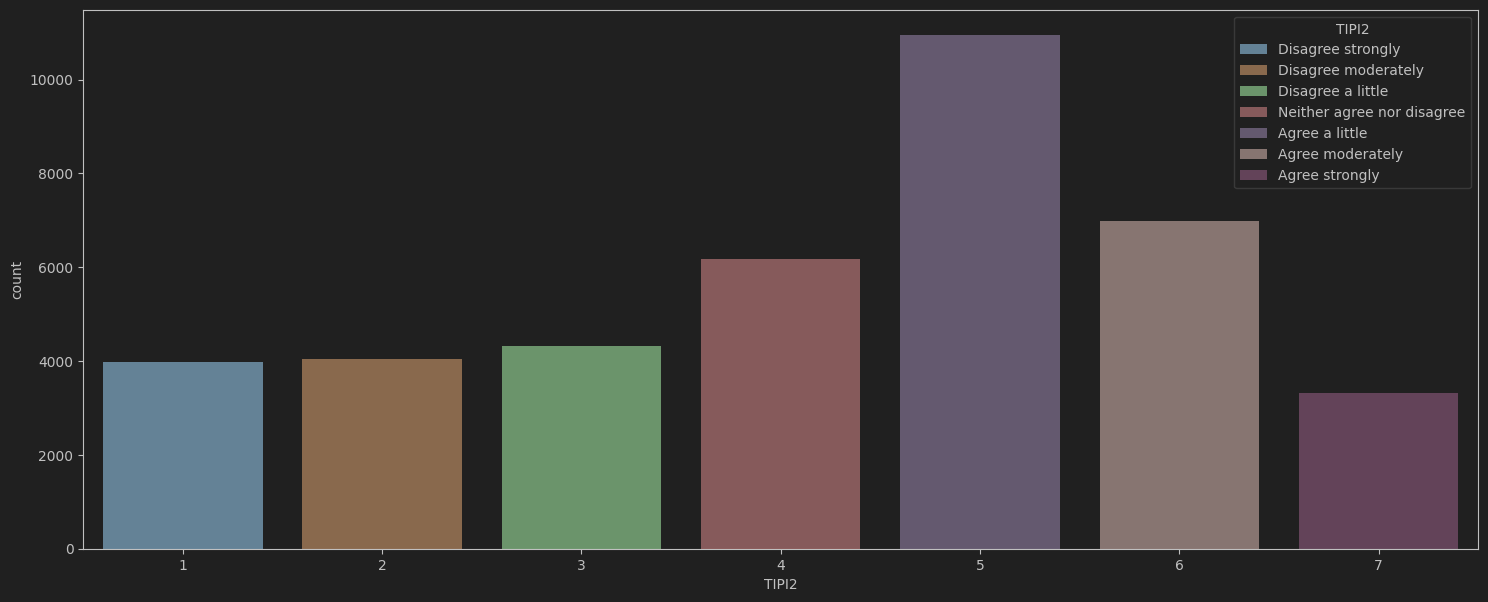
****

* 1. **Race: T**he study aims to represent a broader population, comparing the data's racial makeup to known demographics can reveal potential biases in the sample.

Majority are Asians .



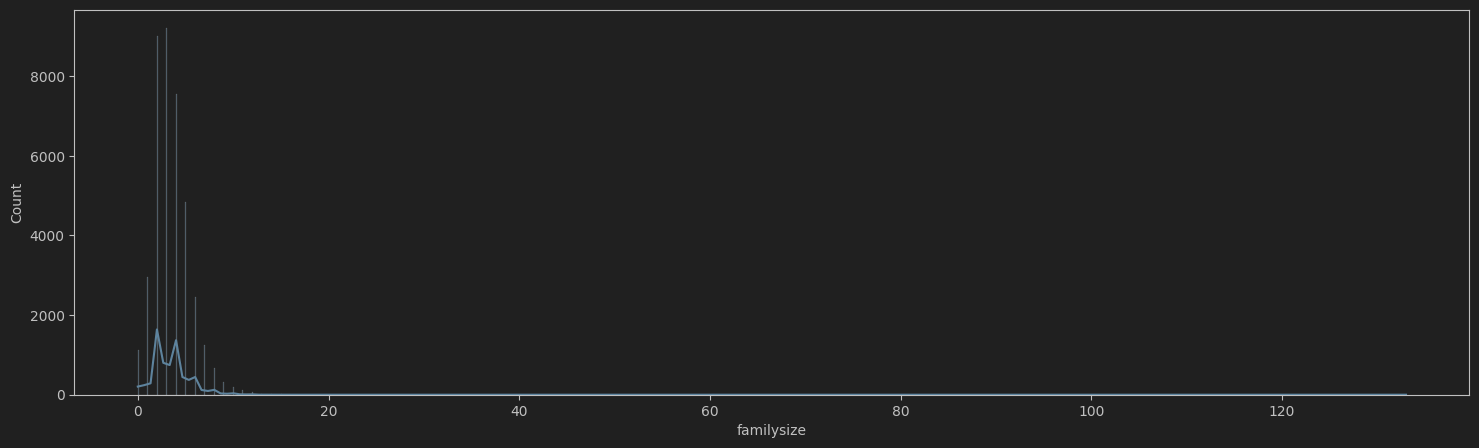
* 1. **TIPI:** TIPI2 labels represent a personality dimension like agreeableness (e.g., disagree strongly, disagree moderately, etc.), the chart shows how many people fall into each category. This can help identify the most common personality types within the data.



**There are graphs for other TIPI’s as well in python notebook!**

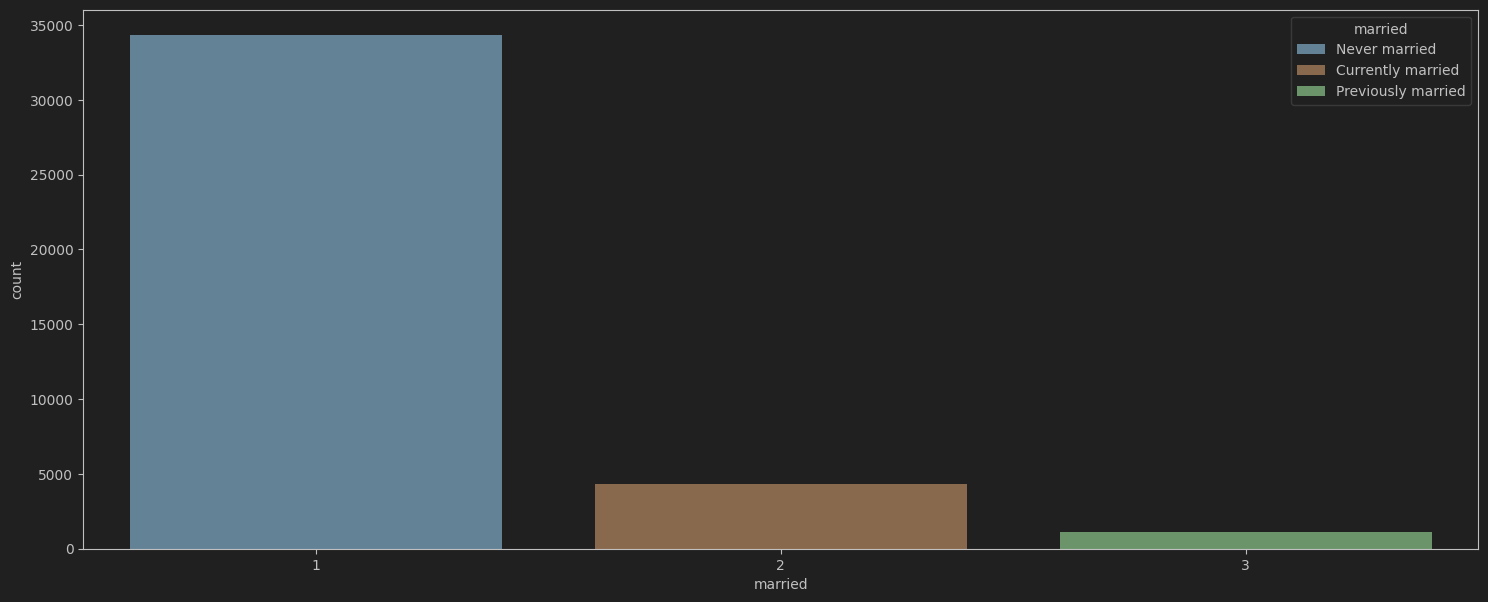
* 1. **Family size: The** histogram provides a good starting point to understand family size distribution within the data. There are so many outliers in family data set.

A symmetrical, bell-shaped curve might suggest a normal distribution of family sizes, where most families fall around an average size with fewer families having very small or very large sizes.



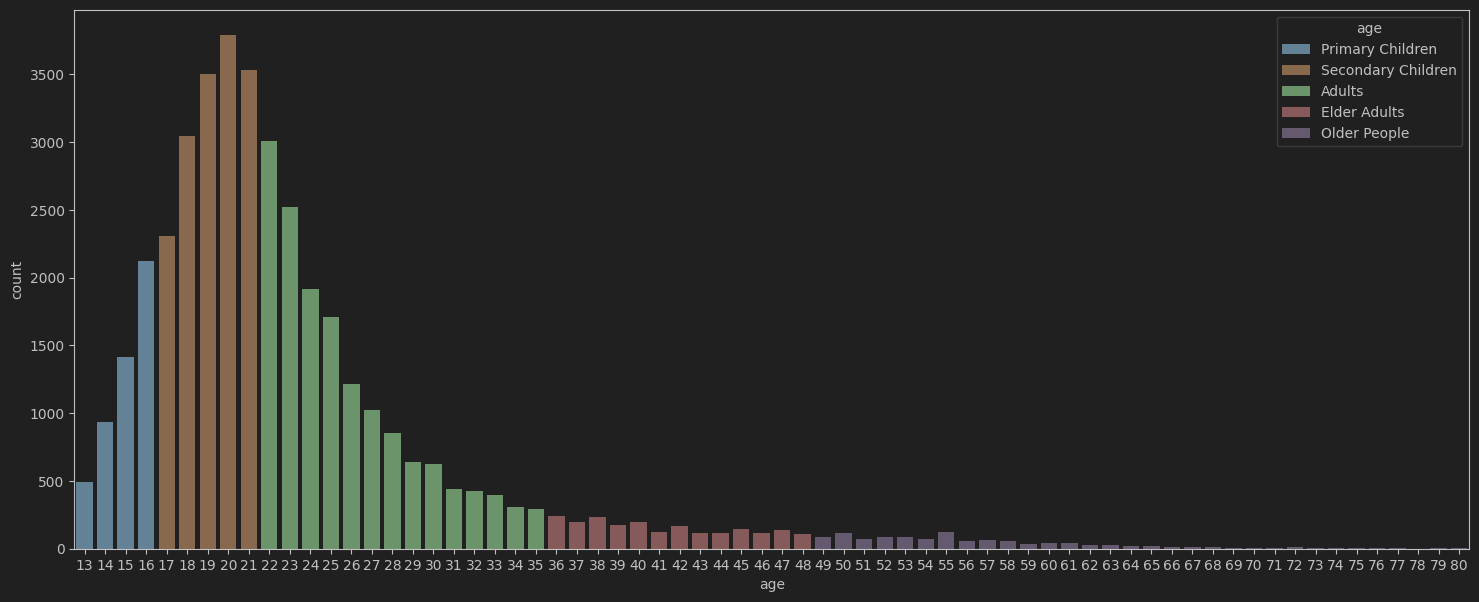
* 1. **Marital Status:** The study aims to represent a specific population, consider how this distribution compares to known demographics. A higher proportion of "never married" individuals might suggest a younger demographic or a population with a specific lifestyle focus.

Most of People who have depression are un-married

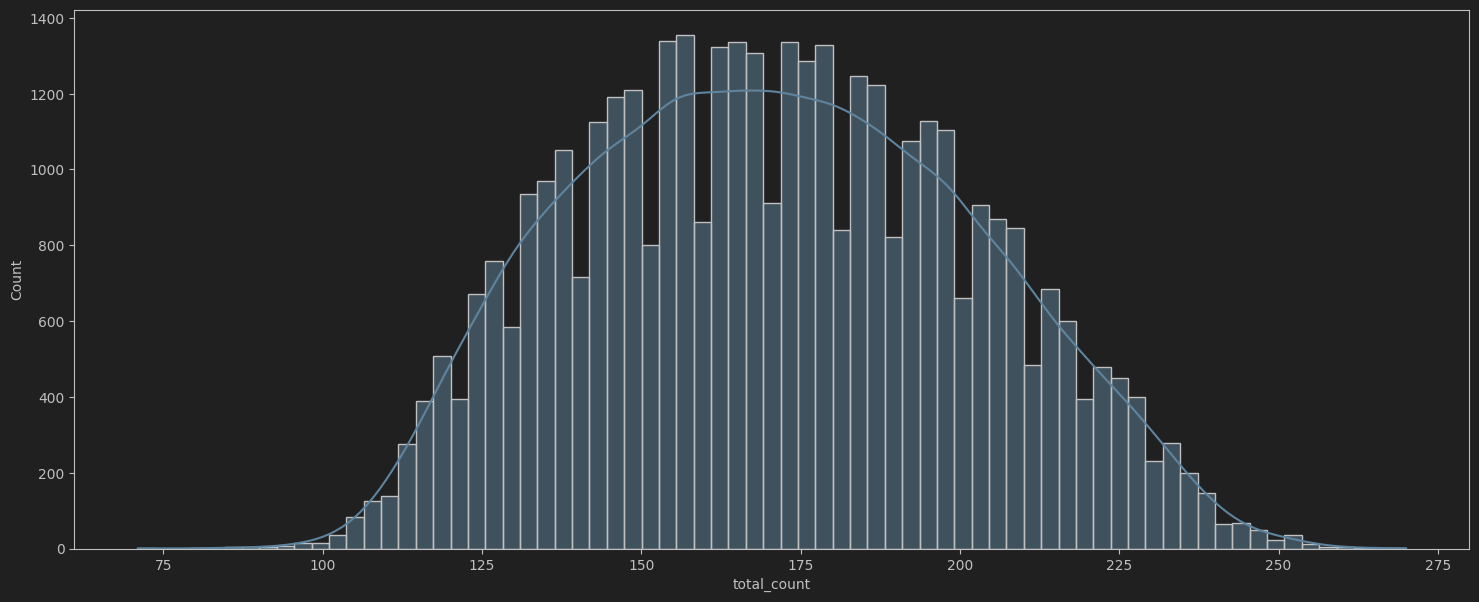


* 1. **Age:** The plot depicts the age distribution of the data, likely representing counts of people in different age groups.

Most of the people who participated are adults and Secondary.

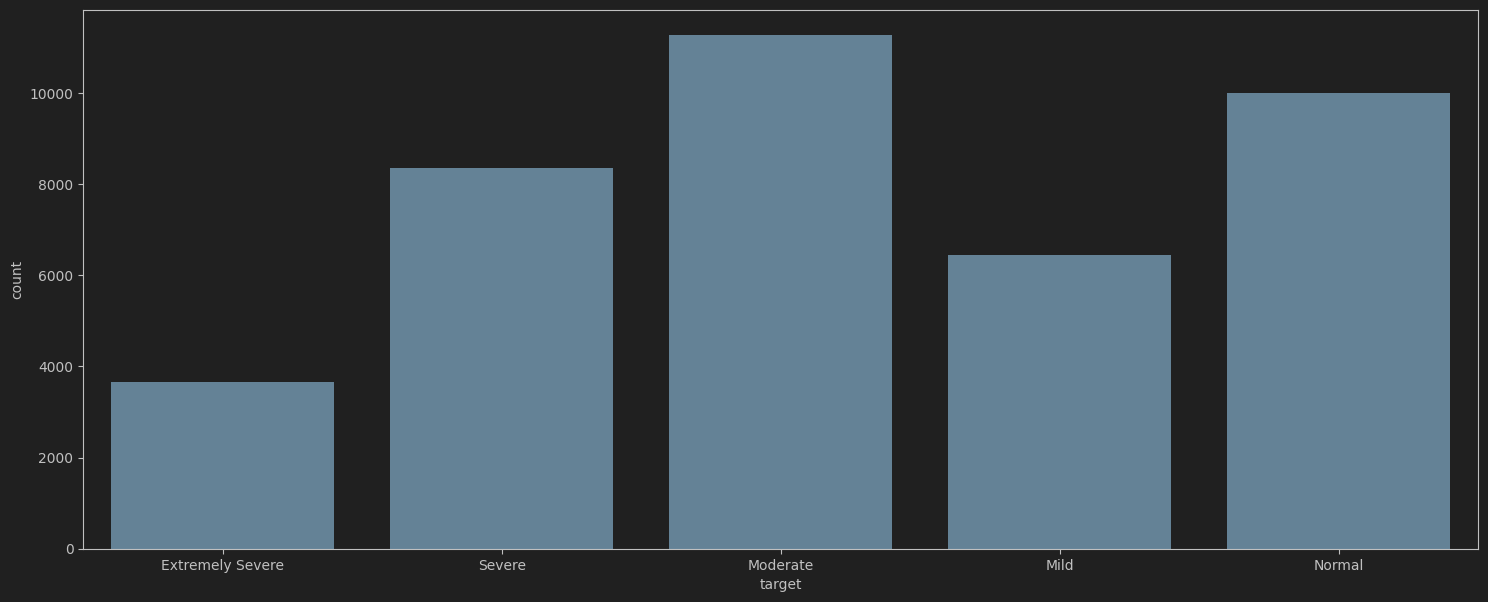


* 1. **Total Count**: To check whether data is balanced or not, this feature is created.



As total count is normally distributed, so divided results in five categories:

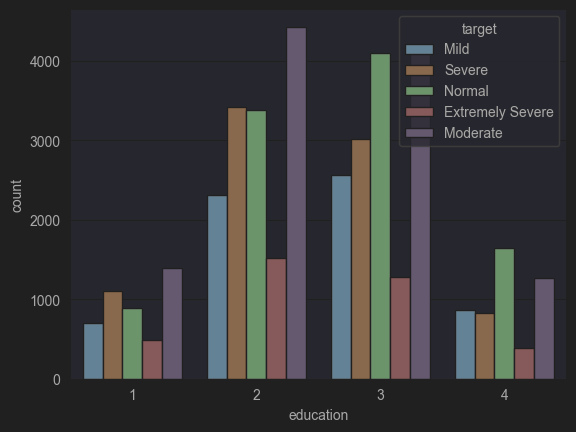
Normal Mild Moderate Severe Extremely Severe

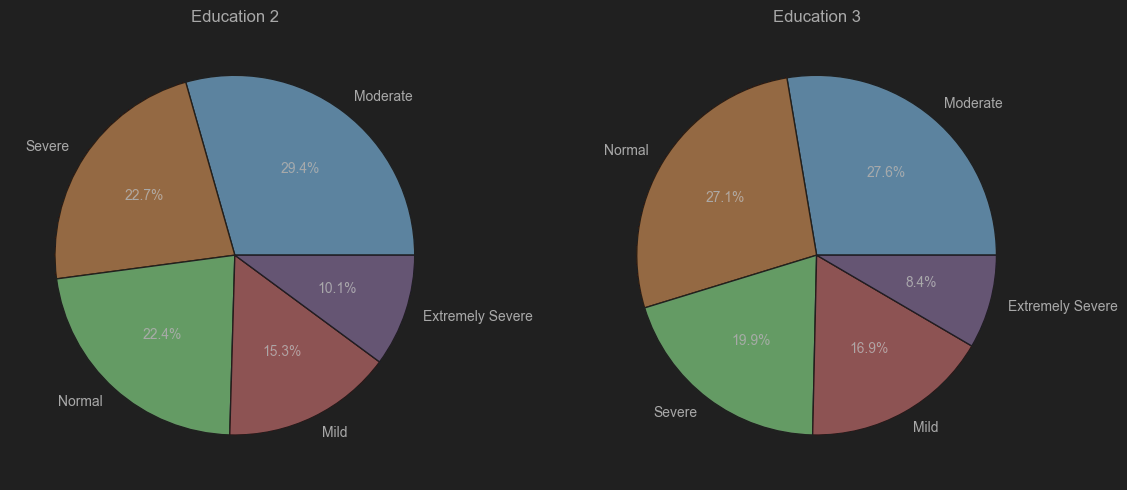


**Analysis of Results with different features:**

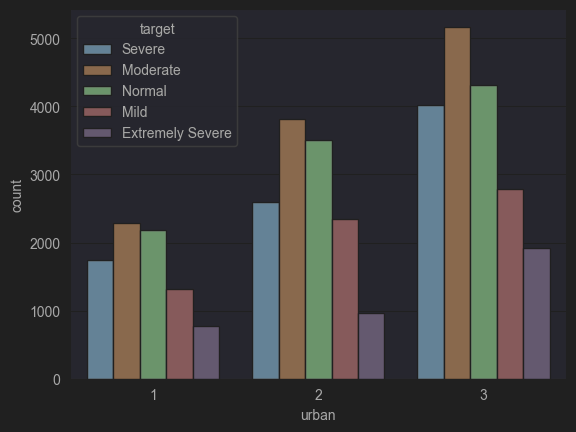
* 1. **Education:**

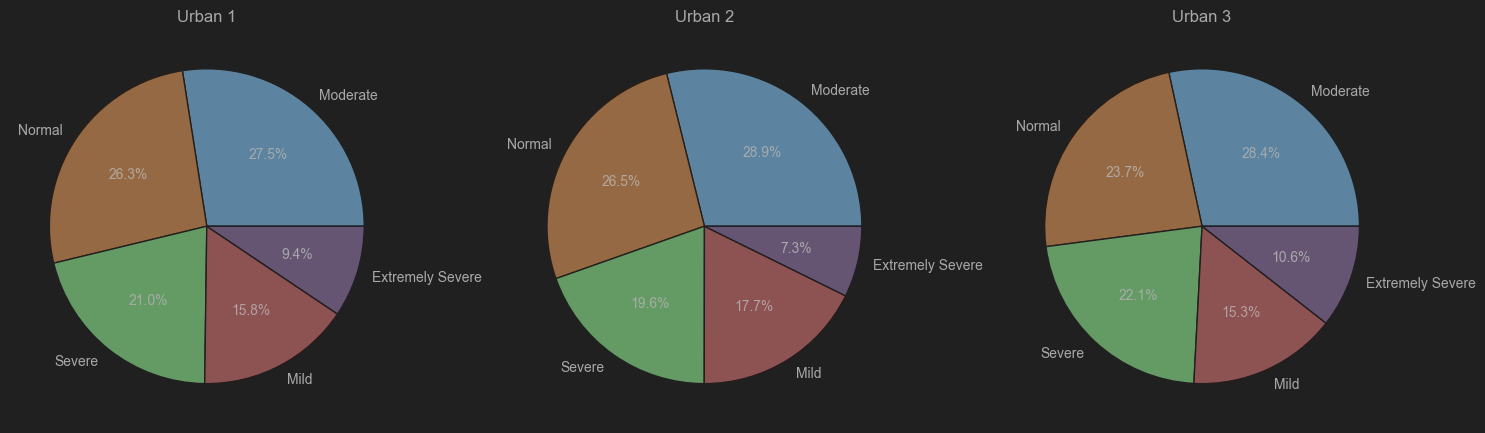
The graph shows that severe and extremely severe depression is prevalent across all education levels, though higher education may be associated with a shift toward more extreme cases rather than less severe forms. Lower education correlates with higher rates of severe depression specifically. Overall, it highlights the persistent challenge of addressing depression effectively, regardless of educational background.

****

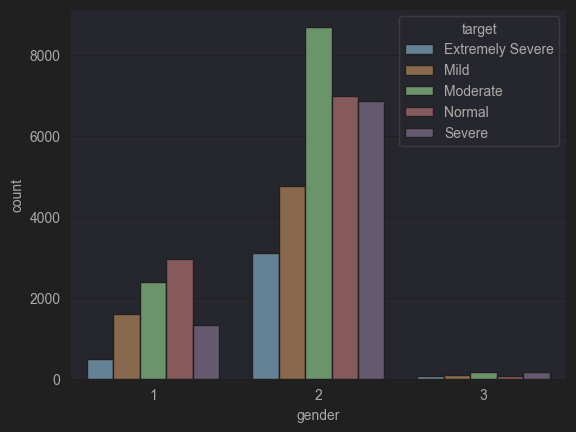
****

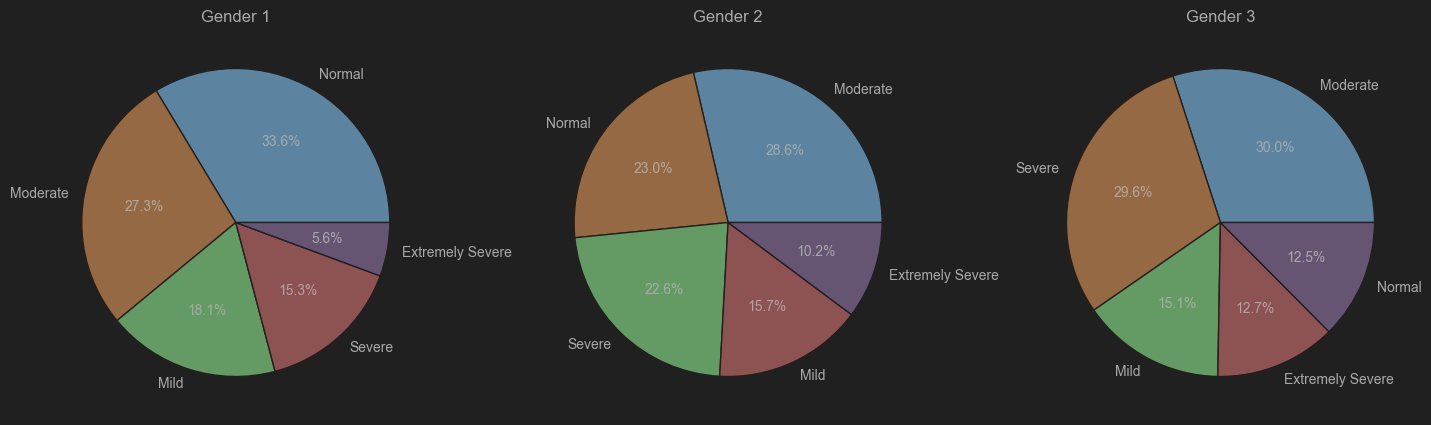
* 1. **Location:** The graph shows varying levels of severity for some condition across different urban locations. Location 2 has the highest overall counts, indicating a potentially higher prevalence or concentration of the condition in that area compared to locations 1 and 3. Location 1 exhibits lower severity levels, while location 3 has a more balanced distribution across severity categories

****

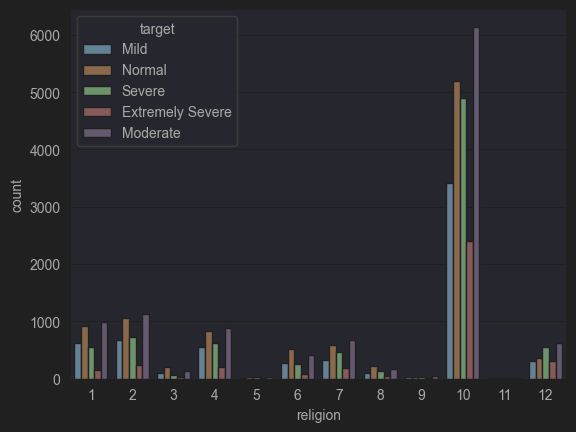
****

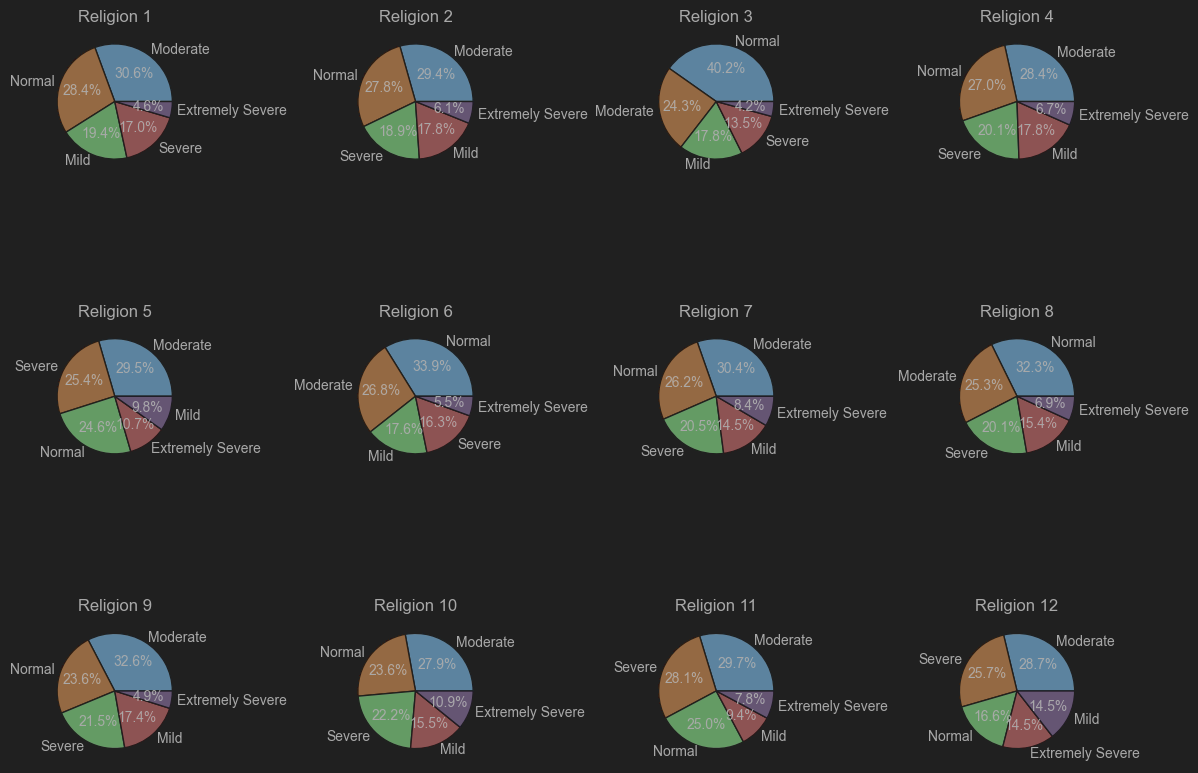
* 1. **Gender:** The graph shows significant differences in severity levels of a condition across different gender groups. Gender group 2(female) exhibits substantially higher counts across most severity categories, indicating a potentially higher prevalence of the condition compared to groups 1 and 3. Group 1(Malws has relatively lower counts across all severity levels. The "target" ideal state has low counts for all gender groups, highlighting a persistent challenge in achieving that target regardless of gender.

****

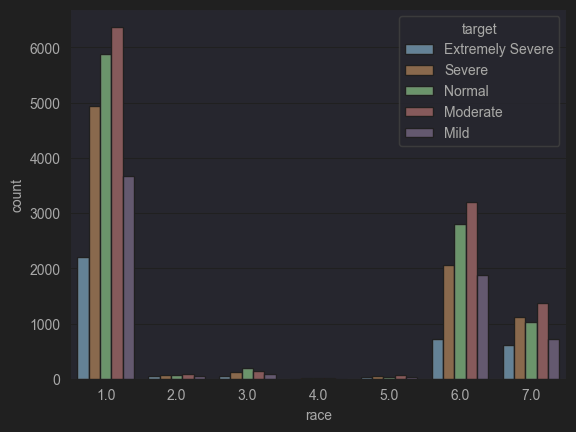
****

* 1. **Religion:** While the "moderate" category stands out with a significantly higher count for one religion, the counts across other severity levels vary across the different religious groups. Christians and Muslims have more depression.

****

****

* 1. **Race**: Native Americans and White people and Black People have more depression



* 1. **TIPI**: TIPI1 to TIPI10: Responses to the Ten Item Personality Inventory (TIPI) questions, which measure different personality traits (e.g., extroversion, agreeableness, conscientiousness).

TIPI1 Extraverted, enthusiastic. TIPI2 Critical, quarrelsome.

TIPI3 Dependable, self-disciplined. TIPI4 Anxious, easily upset.

TIPI5 Open to new experiences, complex. TIPI6 Reserved, quiet.

TIPI7 Sympathetic, warm. TIPI8 Disorganized, careless.

TIPI9 Calm, emotionally stable. TIPI10 Conventional, uncreative.

The TIPI items were rated "I see myself as:" \_\_\_\_\_ such that

1 = Disagree strongly 2 = Disagree moderately

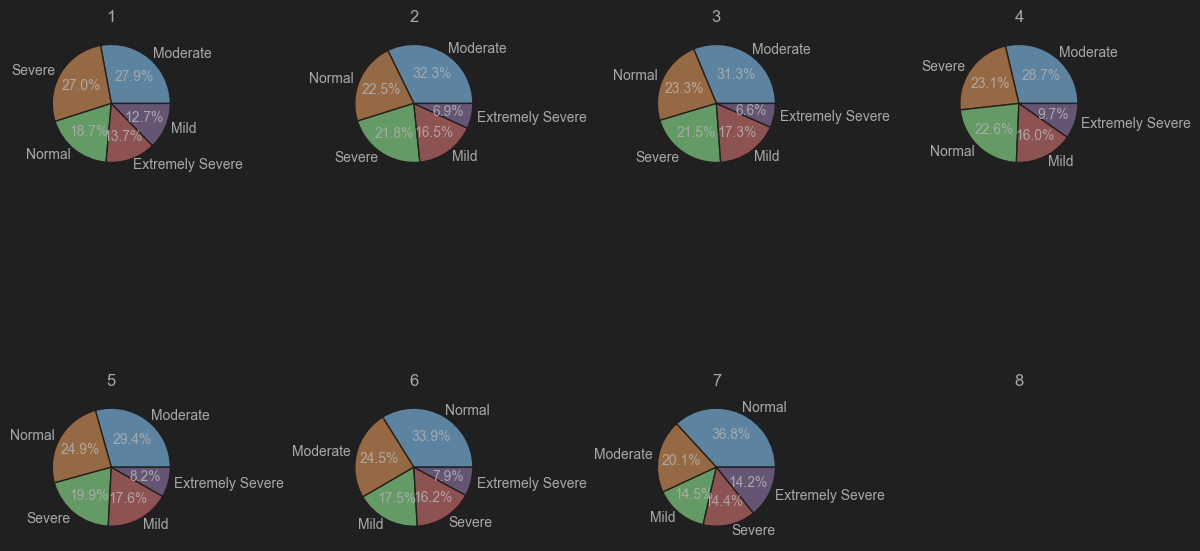
3 = Disagree a little 4 = Neither agree nor disagree

5 = Agree a little 6 = Agree moderately

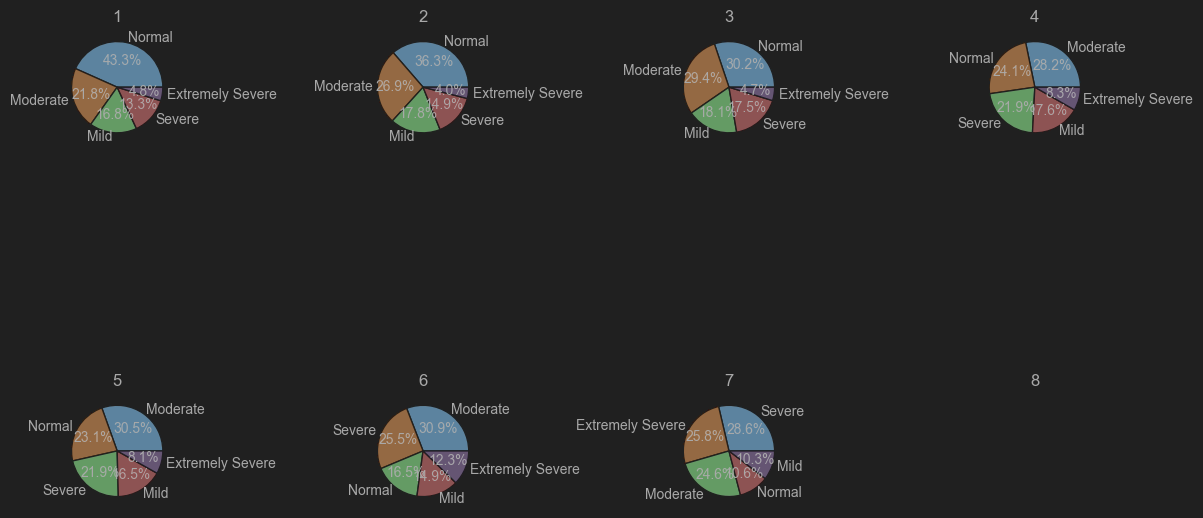
7 = Agree strongly

Exploratory graphs for above key features are as follows:

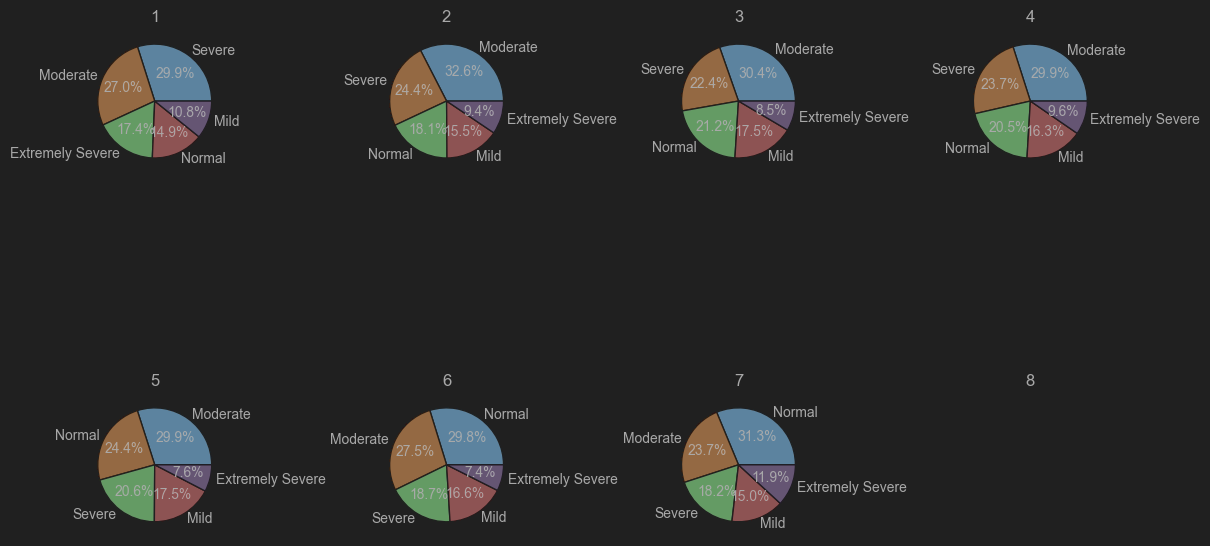
* 1. TIPI1: Extraverted, enthusiastic. People who disagree strongly or agree strongly on being extraverted and enthusiastic have more depression.



* 1. TIPI2: Critical, quarrelsome. People who moderately or strongly agree on being extraverted and enthusiastic have more depression.

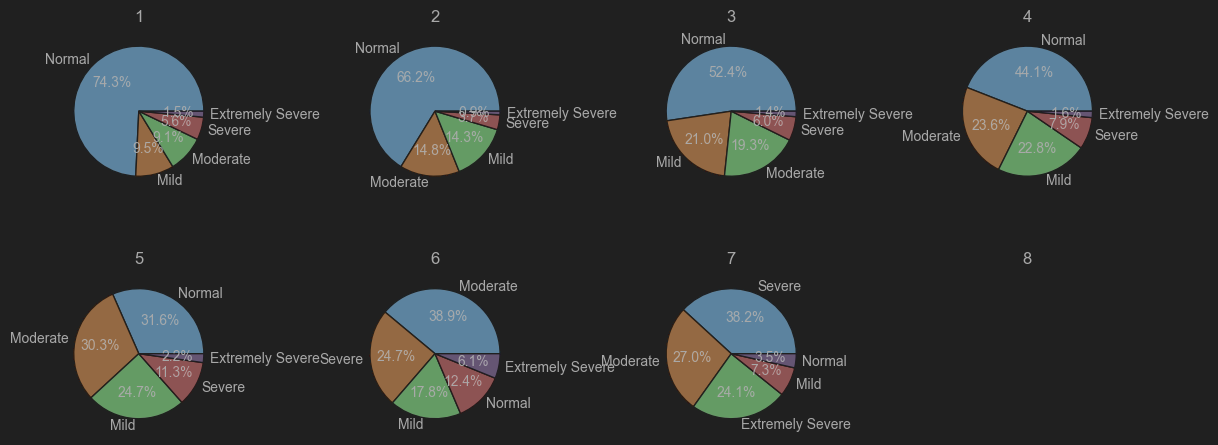


* 1. TIPI3: dependable, self-disciplined. People who disagree on being dependable and self-dependable and self-disciplined have more depression



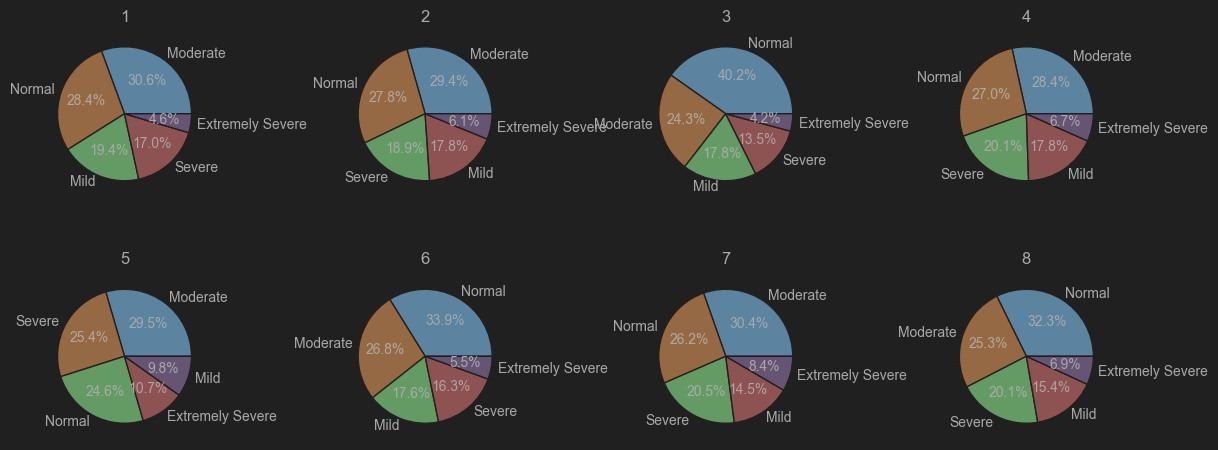
* 1. TIPI 4: Anxious, easily upset

People who moderately or strongly agree on being anxious and easily upset have more depression



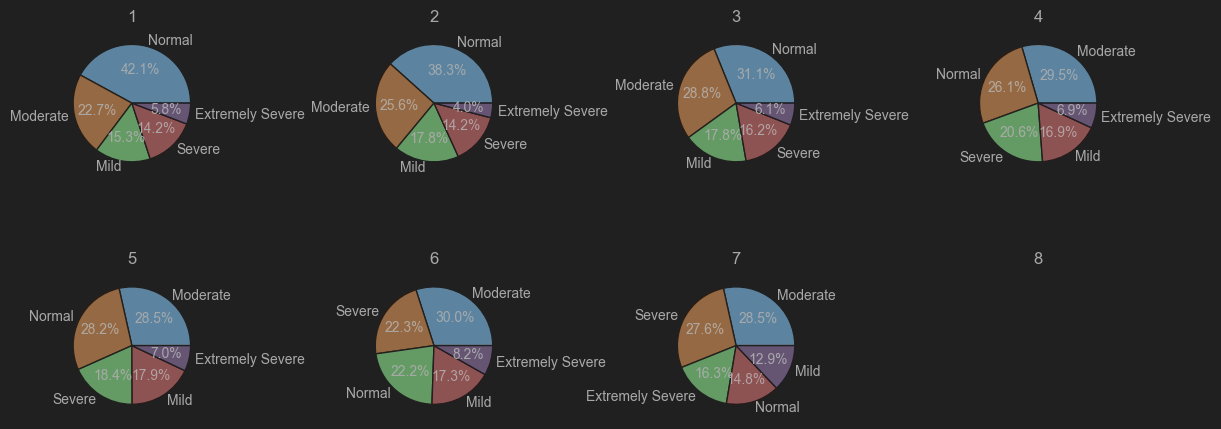
* 1. TIPI5: Open to new experiences, complex

People who disagree strongly or moderately on being open to new experiences have more depression.



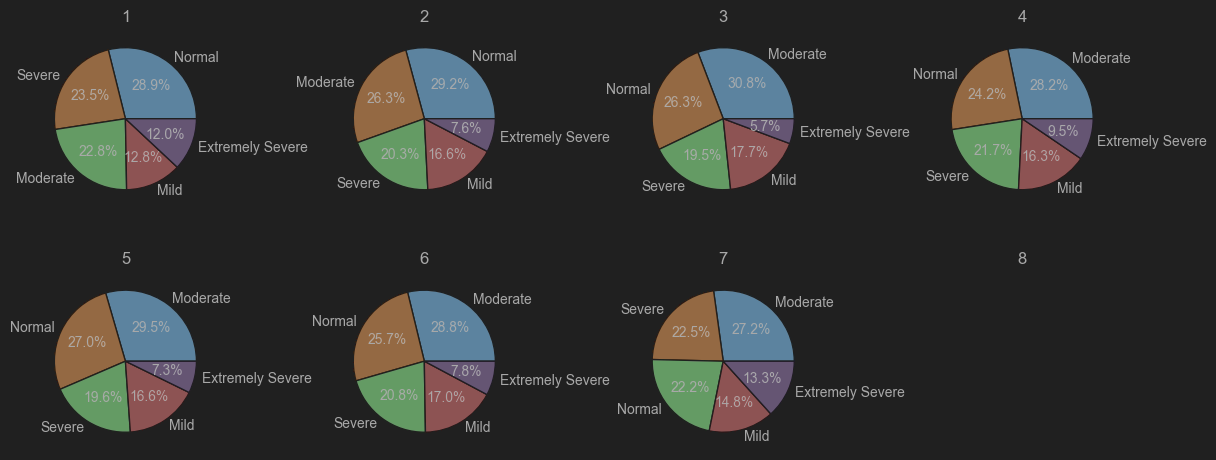
* 1. TIPI6: Reserved, quiet

People who agree strongly or moderately on being reserved have more depression.



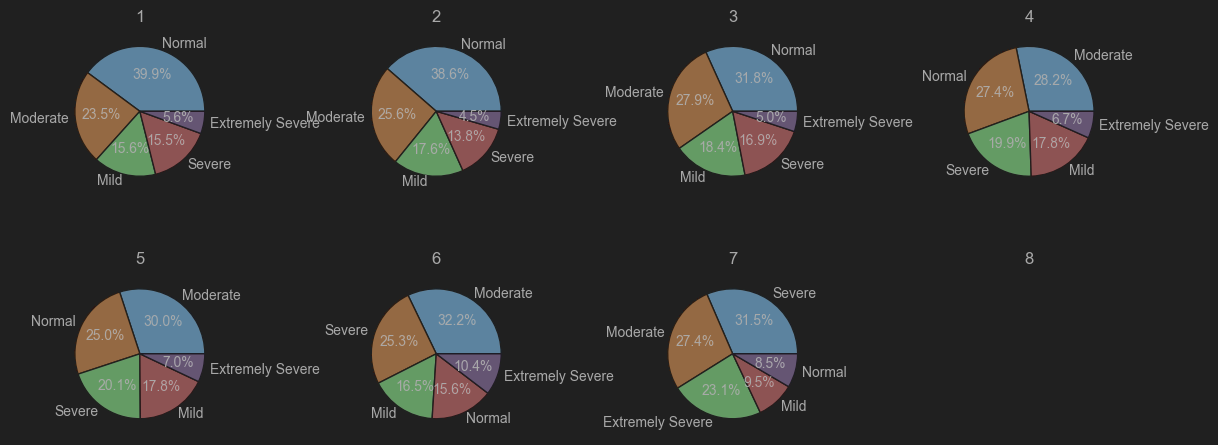
* 1. TIPI7: Sympathetic, warm

People who disagree strongly on being sympathetic and warm have more depression.



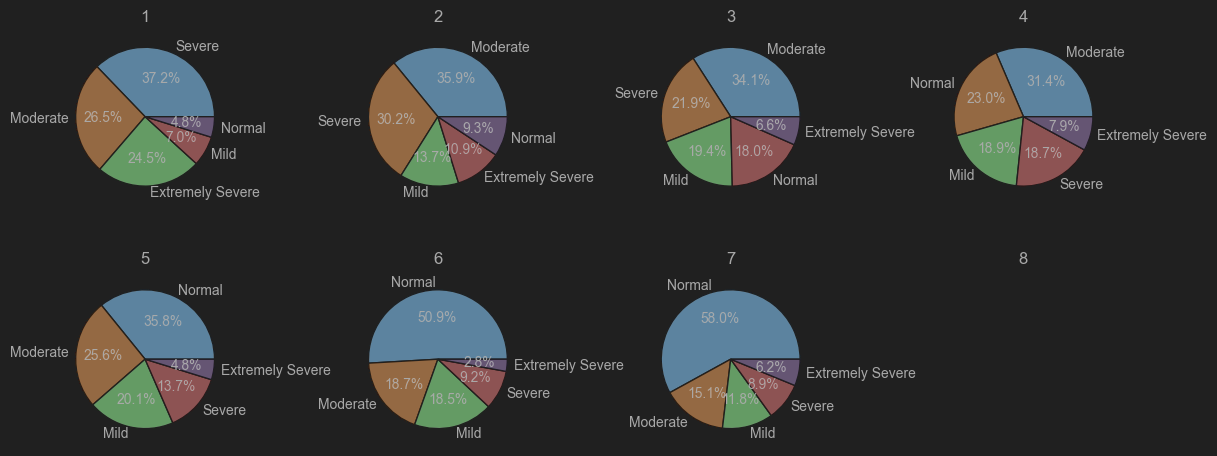
* 1. TIPI8: Disorganized, careless

People who agree strongly or moderately on being disorganized and careless have more depression.



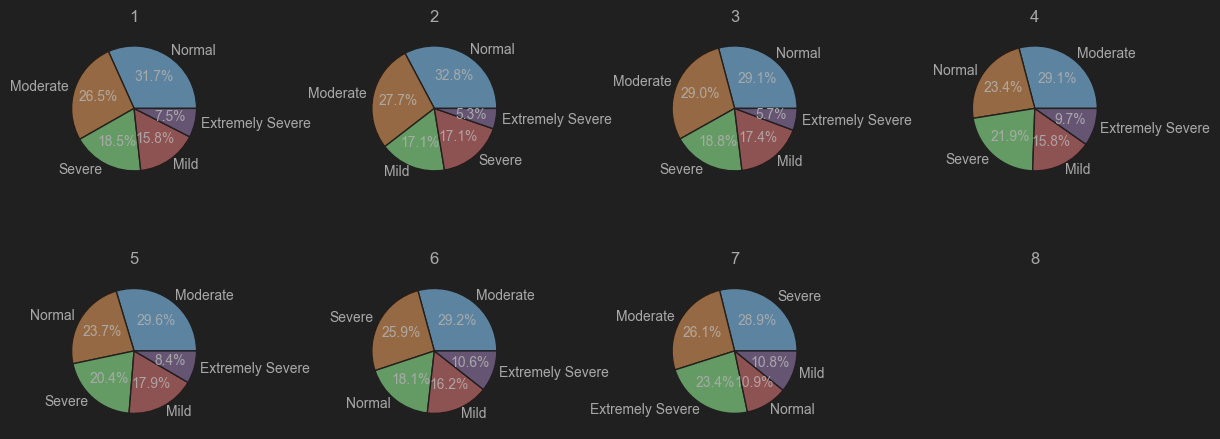
* 1. TIPI9: Calm, emotionally stable

People who disagree moderately or strongly on being calm and emotionally stable have more depression.

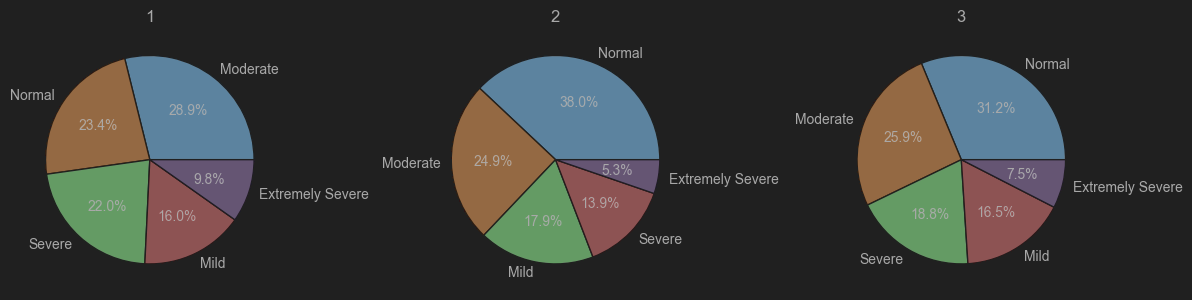


* 1. TIPI10: Conventional, uncreative

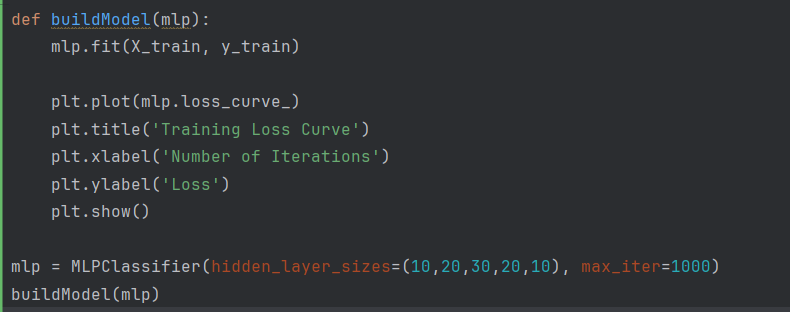
People who disagree moderately or strongly on being uncreative have more depression.

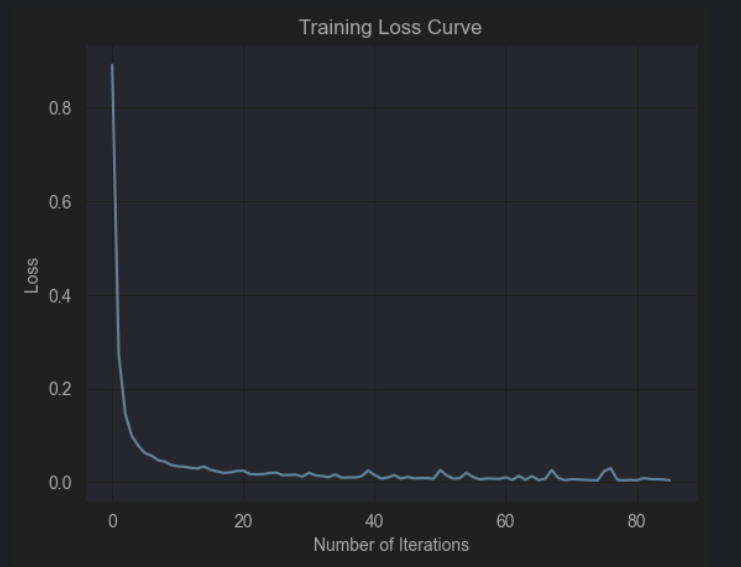


* 1. Marital Status: Single and divorced people have more depression.

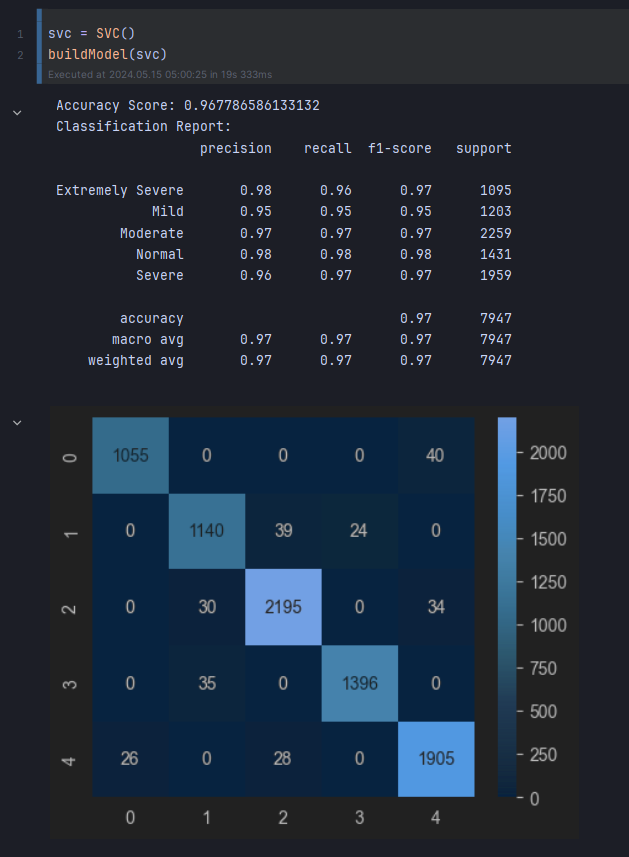


**Build Model - MLP:-**

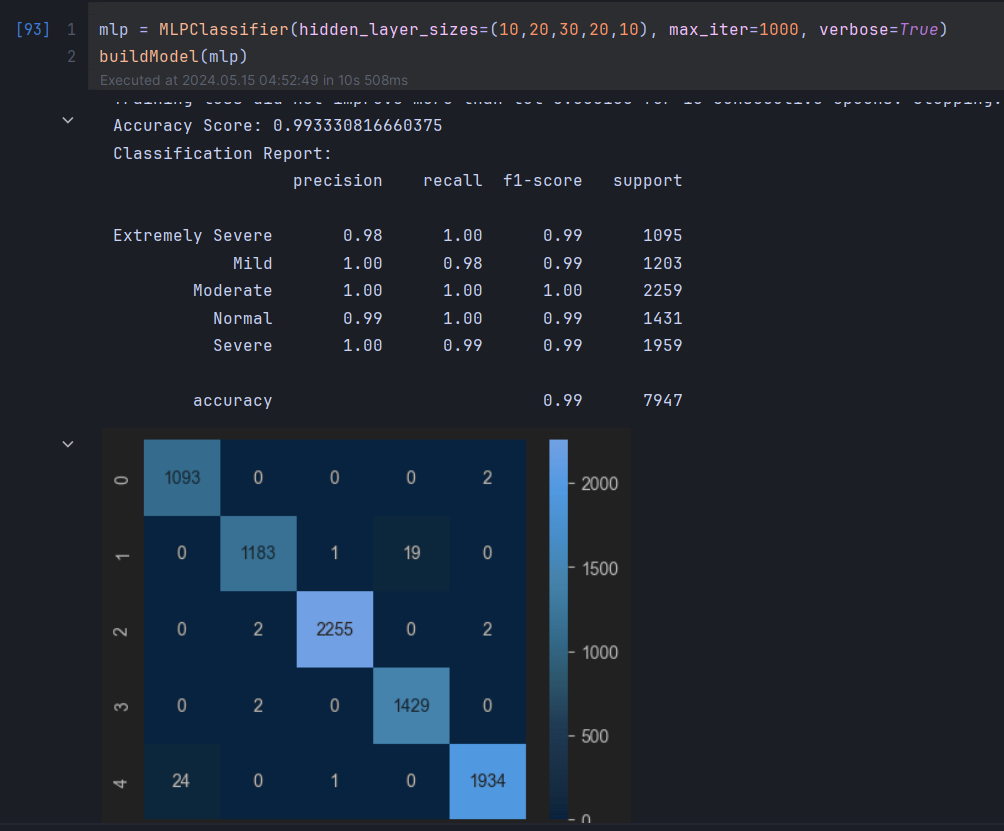




1. **Results:**
   1. Using SVC



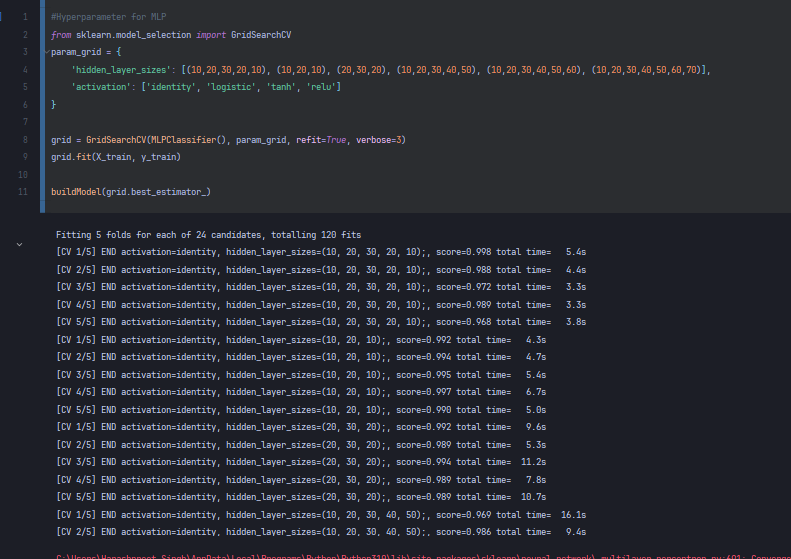
* 1. Using MLP



* 1. Using Logistic Regression

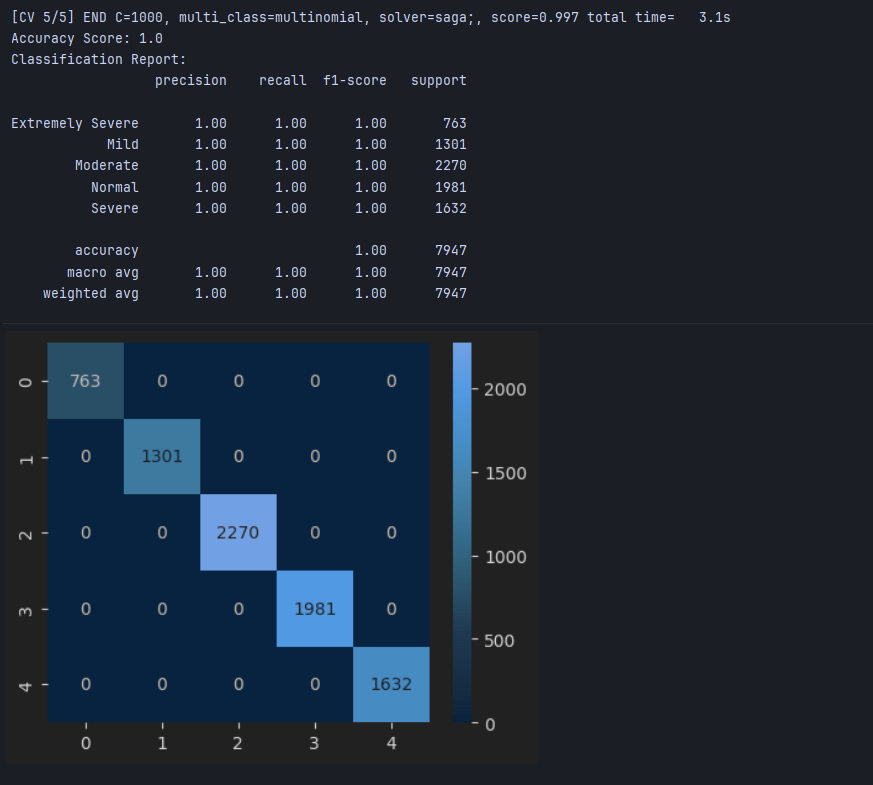


**Hyper Parameter Tuning on MLP :-**

****

****

**Hyper Parameter Tuning on Logistic Regression :-**

****

**INNER ECHO**

**Based on Machine Learning:**

So here are the glimpses of the software:

