Gender (In)Equality

15.05.2021

Group Members

Sadia Sayed

Shristi Roy

Ruheen Fatima Qureshi

Arfaa Shaikh

Mithibai College of Arts,
Chauhan Institute of Science &
Amrutben Jivanlal College of Commerce and Economics
NAAC Reaccredited 'A' Grade, CGPA: 3.57

Department of Statistics

Post-Graduation

CERTIFICATE

This is to certify that

Ms. Arfaa Imtiyaz Shaikh
Ms. Ruheen Fatima Qureshi
Ms. Sadia Mujeebuddin Sayed
Ms. Shristi Vikramaditya Roy

of M.Sc. Part-II have successfully completed the project entitled

"Gender (In) Equality"

during the academic year 2020-2021.

Asst. Prof. Chaitanya Alshi
(Co-ordinator)
M.Sc. Statistics External
Examiner

ACKNOWLEDGMENT

We would like to express our heartfelt thanks and gratitude to all, who out of concern had come forward in completing our project.

We express our sincere thanks to the Principal of Mithibai College for extending her support.

Our deepest sense of gratitude to the Department of Statistics for providing us with the required amenities and co-operation which helped us completing our project successfully.

Special thanks to Asst.Prof. Chaitanya Alshi for guiding and helping us whenever required. Also special thanks to HOD Mr. Amrit Rajwadkar, Prof. Sunila Datar, Asst Prof. Dhaval Shah, Asst Prof. Pratik Patil, Ms. Priyanka Dangar, Ms. Madhura Barve for their helpful inputs and guidance.

And finally, our special thanks to all those who provided us with their vital information, opinions, and precious time and support.

INDEX

Objective

Objective	Technique
• To study causes leading to gender disparities in all aspects of society, and find out reasons leading to gender inequality.	• Pareto Analysis
• To study gender inequality in the workplace , the gender pay gap, and equal opportunities for both genders. • To classify respondent based on demographics who support to mandate paternity leave across all sectors	• K- Means Clustering • Decision Tree
• To study causes leading to gender disparities in all aspects of society. • To study the position of men and women concerning social representation & participation.	• Random Forest • Association Rule
• To analyse people's opinions about the misuse of laws made to uplift Gender Inequality in India.	• K- Modes Clustering
• To study the impact of Covid-19 on gender inequality.	• Odds Ratio
• Effects of media on Gender Inequality. • To study factors promoting Gender Equality.	• Text Classification • Factor Analysis

Introduction

Rightly did Swami Vivekanand say, 'Just as a bird can not fly with one wing only, a Nation can not march forward if the women are left behind. Men and women are the two holes of a perfect whole. Strength is borne of their union; their separation results in weakness. Each has what the other does not have. Each completes the other and is completed by the other.

Man and woman both are equal and play a paramount role in the creation and development of their families in particular and the society in general. Indeed, the struggle for equality has been one of the major concerns of the women's movement all over the world. Gender inequality is, therefore, a form of inequality that is distinct from other forms of socioeconomic inequalities. Gender inequality in India is a crucial reality.

India has fallen 28 spots to rank **140th** among **156** countries on the World Economic Forum's Global Gender Gap index. In 2020, India had ranked 112th among 153 countries on the index. The literacy rate in the country is **74.04%**, **82.14%** for males, and **65.46%** for females.

The gender pay gap in India refers to the difference in earnings between women and men in the paid employment and labor market. According to a report by TeamLease, more than 72% of women feel gender discrimination is still prevalent in the workplace. At 53 percentage points, India has one of the worst gender gaps (difference between the sexes) in the world when it comes to labor force participation a recent World Economic Forum report that only 14.3% of science researchers in India are women data from the National Sample Survey Organisation shows that the percentage of women working in finance, insurance, real estate, and business services, which includes information technology services, is only 13.4% across rural and urban populations. The report also said that the estimated earned income of women in India was one-fifth of men's, which put India among the bottom 10 countries globally on that indicator.

The Indian Penal code, in its basic form, is the main criminal code of India, which lists all the cases and punishments that a person committing any crimes is liable to be charged with, and covers any Indian citizen or a person of Indian origin. The assumed mindset that all violence is male generated, does not only create a gender divide in the society but provides a shield to the crimes perpetrated by women. 'While new data collection on the socioeconomic impacts of Covid-19 must prioritise sex-and age disaggregated data to measure the gendered impacts of the pandemic on adults and children, existing data suggests that Covid-19 will deepen existing gender inequalities', from UNICEF website.

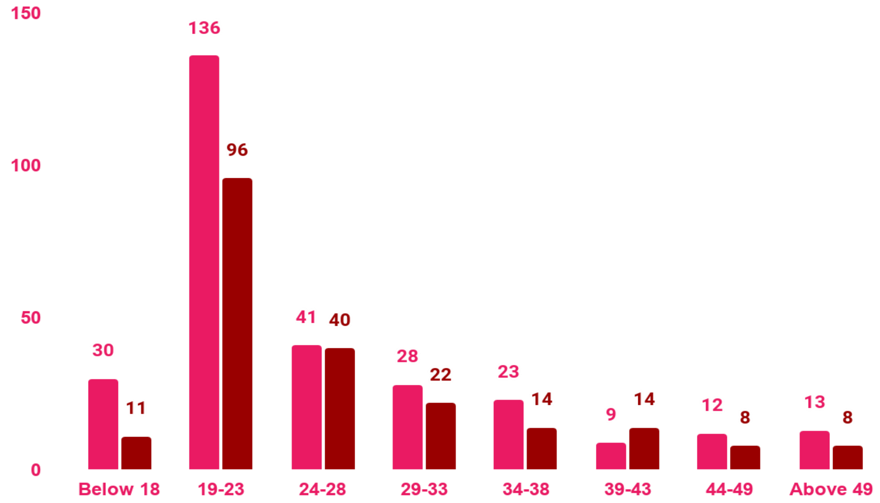
Women commit crimes for the same reasons that men do. Crime has no gender, and neither should our laws.

There is no provision on paternity leave in Indian labour law for private sector workers. The civil servants (Central Government) however are entitled to paternity leave. Being a country where our family is of first and foremost importance for us, an employer needs to keep in mind that having a child is a start to the chapter of family for almost all, hence, it is an utter necessity to provide a reasonable amount of maternity as well as paternity leave.

Gender equality, also known as sexual equality, is the state of equal ease of access to resources and opportunities regardless of gender, including economic participation and decision-making; and the state of valuing different behaviors, aspirations, and needs equally, regardless of gender. It has been quoted by UNICEF **"India will not fully develop unless both girls and boys are equally supported to reach their full potential"**. Educating Indian children from an early age about the importance of gender equality could be a meaningful start in that direction.

GRAPHICAL ANALYSIS

Gender In AGE



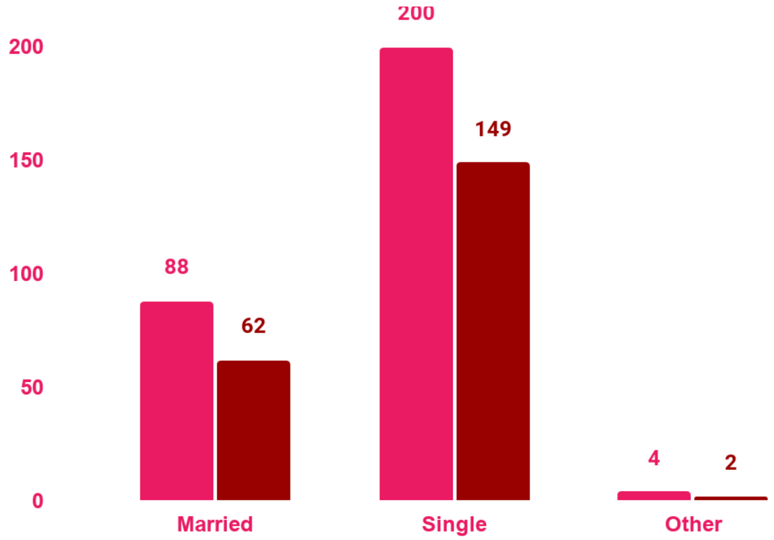
FEMALE

Majority of females are in the age group 19-23

MALE

Majority of Males are in the age group 19-23

Present Marital STATUS



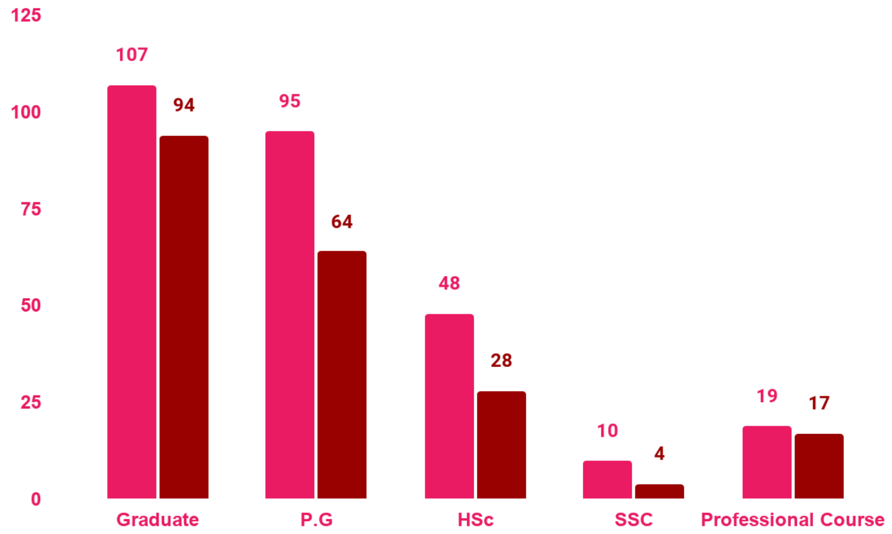
FEMALE

Majority of females are Single

MALE

Majority of males are Single

GENDER in EDUCATION



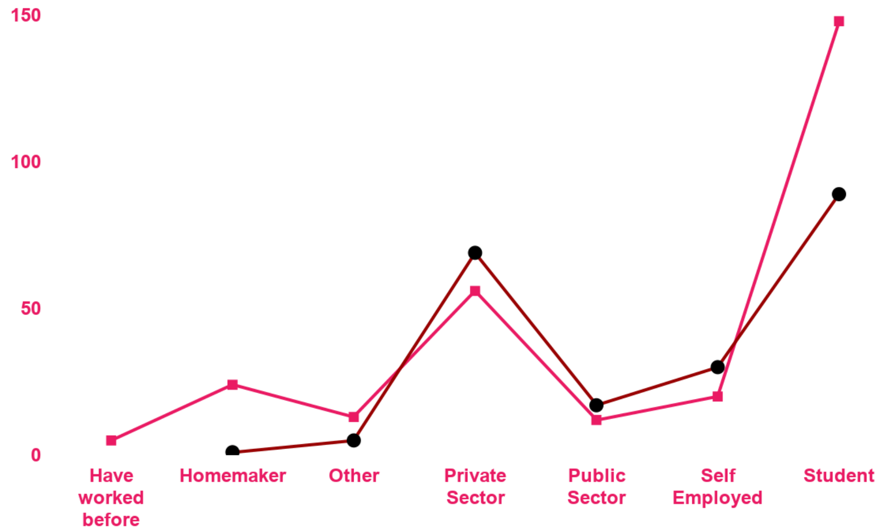
FEMALE

Females have a higher rate in education

MALE

Majority of the males are graduated

GENDER in OCCUPATION

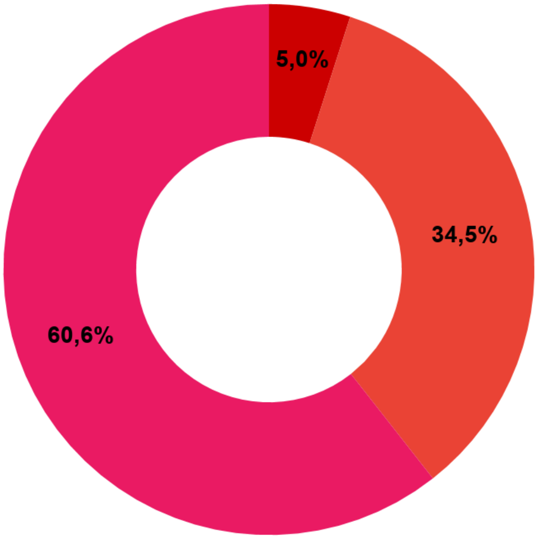


FEMALE

Though the proportion of females in education is more, we see less participation of women in the workforce.

MALE

FAMILY TYPE



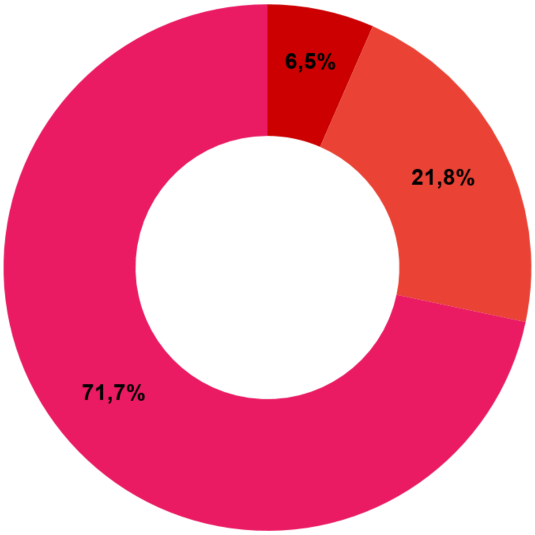
NUCLEAR FAMILY

Majority of the respondents lives in a Nuclear family

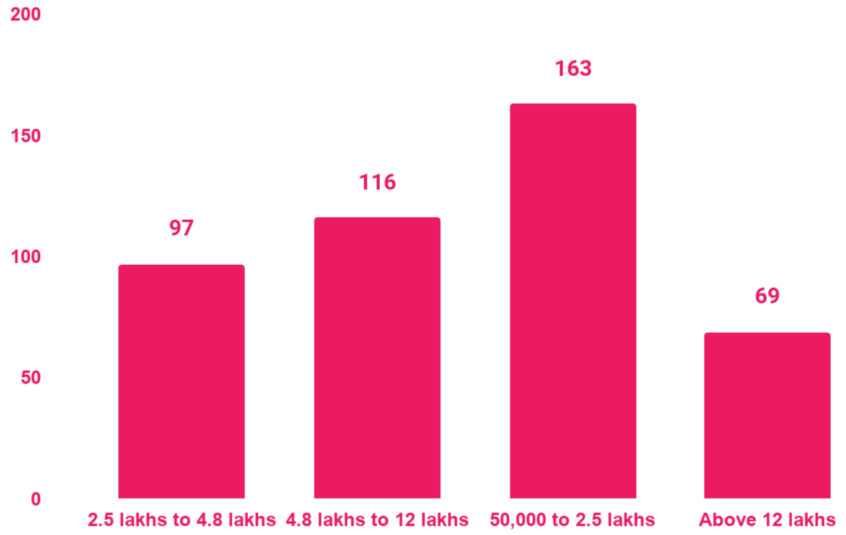
JOINT FAMILY

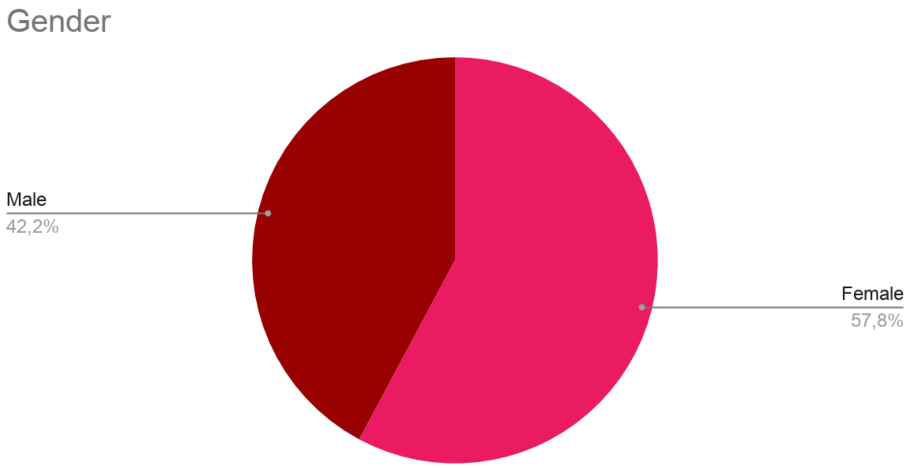
ALONE

AREA TYPE



FAMILY INCOME





Objective:

To find reasons for gender inequality.

PARETO ANALYSIS

Pareto Analysis is a Statistical Technique in decision making that is used for the selection of a limited number of tasks that produce a significant overall effect. It uses the Pareto Principle. It is also known as 80/20 rule. The idea is that by doing 20% of the work you can generate 80% of the benefit of doing the whole job. This is also known as 'vital few' and the 'trivial many' effect.

The Pareto Principle has many applications in quality control. It is the basis for the Pareto Diagram, one of the key tools used in total quality control and Six Sigma. A Pareto chart is used to graphically summarize and display the relative importance of the differences between groups of data. Pareto chart organizes and displays information to show the relative importance of the differences between groups of data. Pareto chart organizes and displays information to show the relative importance of various problems or causes of problems. It is essentially a special form of vertical bar chart that puts items in order (highest to lowest) relative to some measurable effect of interest such as frequency, cost or time.

The chart is based on the Pareto Principle, which states that when several factors affect a situation, a few factors will account for most of the impact. Pareto describes a phenomenon in which 80 percent of variation observed can be explained by a mere 20 percent of the causes of that variation. The Pareto curve makes it clear as to where effort must be concentrated to give maximum effect.

The Pareto Chart is a very simple but effective tool for prioritizing problem causes, which is why it is widely used for problem-solving in the manufacturing industry.

The Pareto Chart is a descending bar graph that shows the frequencies of occurrences or relative sizes of the various problems or causes of a particular problem.

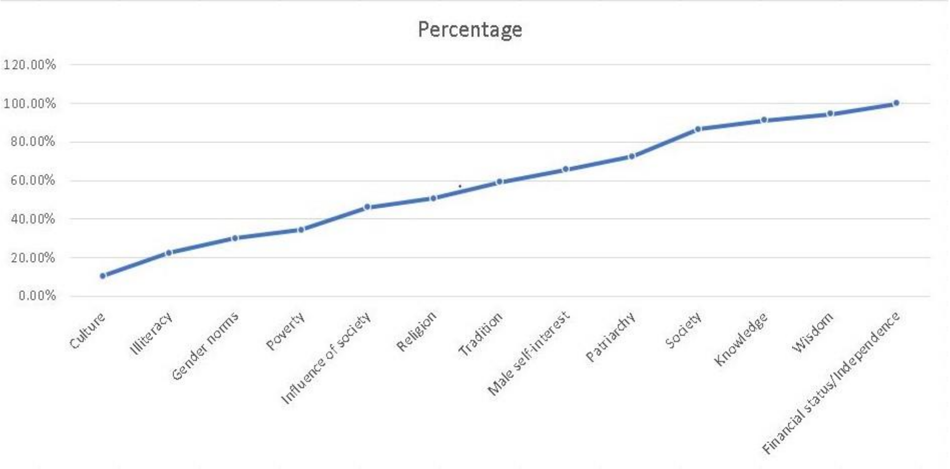
The problem categories or causes are shown on the x-axis of the bar graph. Aside from its main bar graph, the Pareto chart may also include a line graph that indicates the cumulative percentage of occurrences at each bar of the bar graph.

This line graph, referred to as the 'cumulative percentage line', is used to determine which of the bars belong to the 'vital few' and which ones are relegated to the 'trivial many'.

Analysis

Variables used:

- Culture
- Illiteracy
- Gender norms
- Poverty
- Influence Of Society
- Religion
- Society
- Tradition
- Patriarchy
- Male Self Interest
- Financial status/Independence
- Knowledge
- Wisdom



CONCLUSION:

From Pareto Analysis we conclude the reasons leading to gender inequality are:

- SOCIETY
- KNOWLEDGE
- WISDOM
- FINANCIAL STATUS / INDEPENDENCE

Objective:

To study gender inequality in the workplace , the gender pay gap, and equal opportunities for both genders.

K-Means Clustering

K-means clustering is a clustering method that subdivides a single cluster or a collection of data points into K different clusters or groups.

The algorithm analyzes the data to find organically similar data points and assigns each point to a cluster that consists of points with similar characteristics. Each cluster can then be used to label the data into different classes based on the characteristics of the data.K-Means clustering works by constantly trying to find a centroid with closely held data points. This means that each cluster will have a centroid and the data points in each cluster will be closer to its centroid compared to the other centroids.

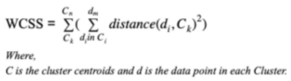
K-Means Algorithm

1. Selecting an appropriate value for K which is the number of clusters or centroids
2. Selecting random centroids for each cluster
3. Assigning each data point to its closest centroid
4. Adjusting the centroid for the newly formed cluster in step 4
5. Repeating step 4 and 5 till all the data points are perfectly organised within a cluster space

Choosing The Right Number Of Clusters

The number of clusters that we choose for a given dataset cannot be random. Each cluster is formed by calculating and comparing the distances of data points within a cluster to its centroid. An ideal way to figure out the right number of clusters would be to calculate the Within-Cluster-Sum-of-Squares (WCSS).

WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids.



$$\text{WCSS} = \sum_{C_k}^{C_n}\Big(\sum_{d_i \, in \, C_i}^{d_m} distance(d_i, C_k)^2\Big)$$

Where,

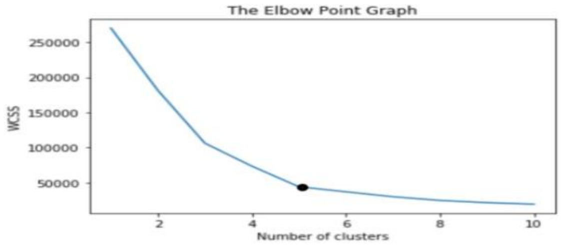
C is the cluster centroids and d is the data point in each Cluster.

The idea is to minimise the sum. Suppose there are n observation in a given dataset and we specify n number of clusters (k = n) then WCSS will become zero since data points themselves will act as centroids and the distance will be zero and ideally this forms a perfect cluster, however this doesn't make any sense as we have as many clusters as the observations. Thus there exists a threshold value for K which we can find using the Elbow point graph.

Elbow method

We can find the optimum value for K using an Elbow point graph. We randomly initialise the K-Means algorithm for a range of K values and will plot it against the WCSS for each K value.

The resulting graph would look something like what's shown below:



For the above-given graph, the optimum value for K would be 5. As we can see that with an increase in the number of clusters the WCSS value decreases. We select the value for K on the basis of the rate of decrease in WCSS. For example, from cluster 1 to 2 to 3 in the above graph we see a sudden and huge drop in WCSS. After 5 the drop is minimal and hence we chose 5 to be the optimal value for K.

The Random Initialisation Trap

One major drawback of K-Means clustering is the random initialisation of centroids. The formation of clusters is closely bound by the initial position of a centroid. The random positioning of the centroids can completely alter clusters and can result in a random formation.

The solution is K-means++. K-Means++ is an algorithm that is used to initialise the K-Means algorithm.

K Means++

The algorithm is as follows:

1. Choose one centroid uniformly at random from among the data points.

2. For each data point say x, compute D(x), which is the distance between x and the nearest centroid that has already been chosen.

3. Choose one new data point at random as a new centroid, using a weighted probability distribution where a point x is chosen with probability proportional to D(x)2.

4. Repeat Steps 2 and 3 until K centres have been chosen.

5. Proceed with standard k-means clustering.

Variables Used

Age

Received less support from senior leaders

Gender

Been passed over for most important assignments

Education

Felt isolated at the workplace

Occupation

Been denied a promotion

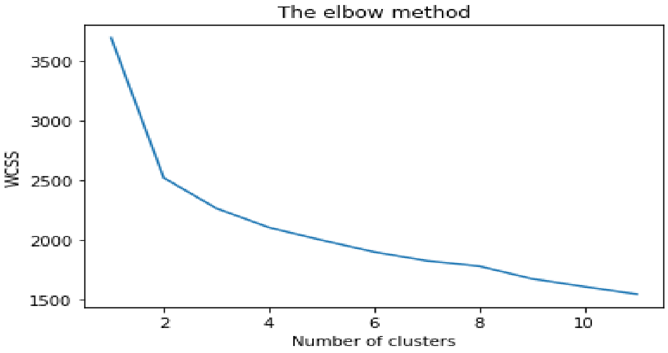
Earned less than a woman/man

Been turned down for a job

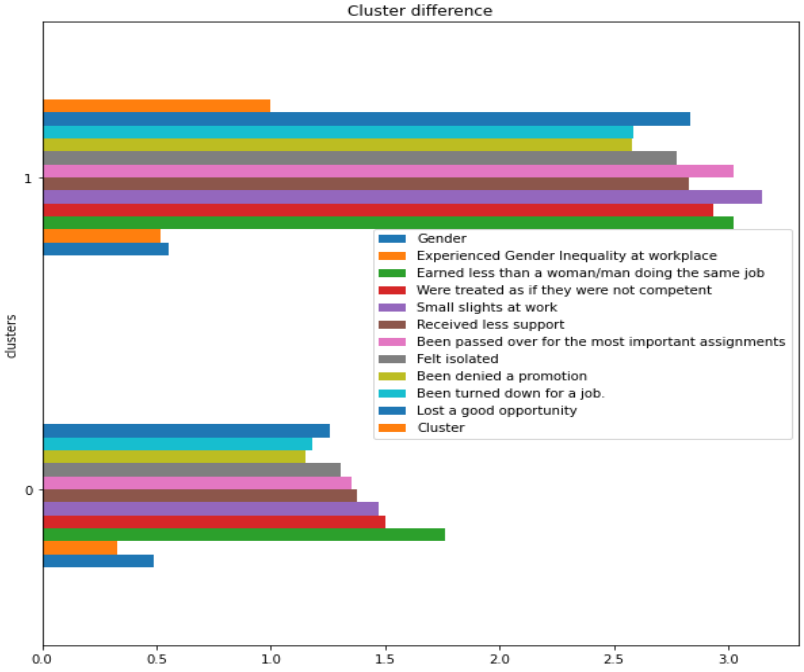
Were treated as if they were not competent

Lost a good opportunity

Experienced repeated small slights at work



As the elbow occurred at k = 2, we have made 2 cluster for our analysis.



Both clusters are significantly different from each other.

Based on the graphical representation of cluster centroids, it can be seen that,

Cluster_0 :

Not encountered instances of gender inequality at the workplace.

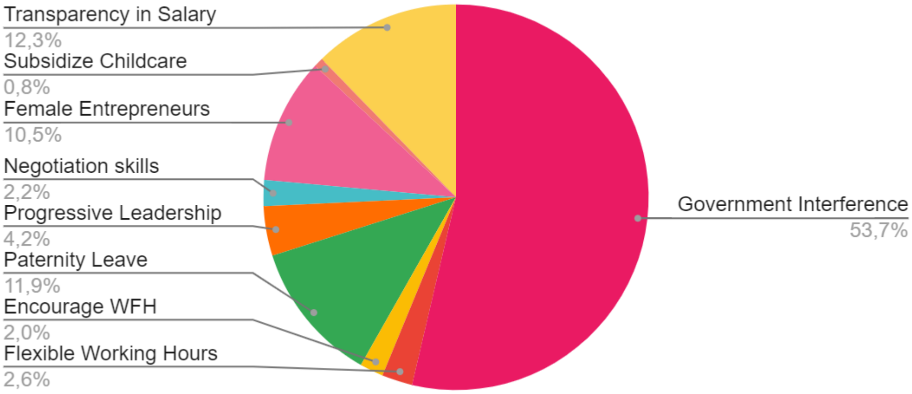
- The characteristics observed in this cluster show that the respondents have not experienced gender inequality at the workplace and they have agreed on being paid less than women/men doing the same job.

Cluster_1 :

Encountered instances of gender inequality at the workplace.

- In this cluster we observe that most of the respondents have agreed to the reasons which are considered to highlight gender inequality at the workplace. For eg: It is clearly seen that earning less than women/men doing the same job, not being competent enough, being passed over for most important assignments, been turned down for a job and losing good opportunities are the most highlighted factors.

Measures to overcome Gender Pay Gap:



- From the responses of our respondents we can clearly identify that Government interference can be a solid measure to overcome the Gender Pay Gap in India which is 19% percent at present between both genders.

Objective:

To understand the characteristics of respondents who support paternity leave being made mandatory across all sectors.

Commented [1]: Hi I'm just suggesting that we should bold this line so that it will look more lively

Commented [2]: ok

Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms (pre-defined target variable). Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems**. It uses tree-like structures and their possible combinations to solve a particular problem.

The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data (training data).

Types of Decision Trees

Types of decision trees are based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:** Decision Tree which has a categorical target variable then it is called a Categorical variable decision tree.

2. **Continuous Variable Decision Tree:** Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

Since our target variable is categorical, we are using a classification tree.

Important Terminology related to Decision Trees

1. **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.

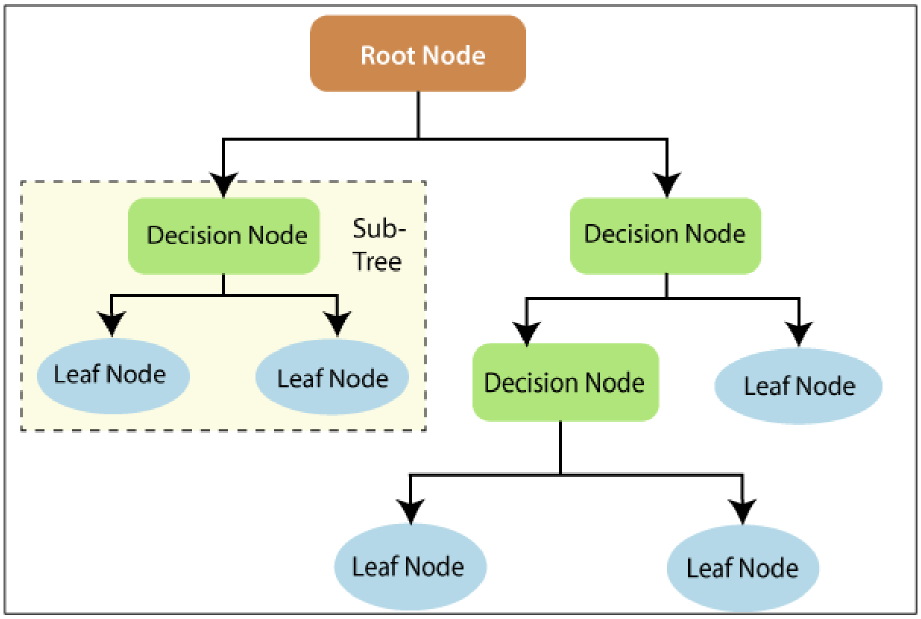
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.

4. **Leaf / Terminal Node:** Nodes that do not split are called Leaf or Terminal nodes.

5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

6. **Branch / Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.

7. **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.



We make some assumptions while implementing the Decision-Tree algorithm. These are listed below:

1. At the beginning, the whole training set is considered as the root.

2. Feature values need to be categorical. If the values are continuous then they are discretized prior to building the model.

3. Records are distributed recursively on the basis of attribute values.

4. Order to place attributes as root or internal node of the tree is done by using some statistical approach.

Decision Tree Splitting Methods:

1. **Gini Impurity:**

Gini Impurity is a method for splitting the nodes when the target variable is categorical. It is the most popular and the easiest way to split a decision tree. The Gini Impurity value is:

$$Gini\ Impurity = 1 - Gini$$

Lower the Gini Impurity, higher is the homogeneity of the node. **The Gini Impurity of a pure node is zero.**

Steps to split a decision tree using Gini Impurity:

1. For each split, individually calculate the Gini Impurity of each child node
2. Calculate the Gini Impurity of each split as the weighted average Gini Impurity of child nodes
3. Select the split with the lowest value of Gini Impurity
4. Until you achieve homogeneous nodes, repeat steps 1-3.

2. **<u>Information Gain:</u>**

Information Gain is used for splitting the nodes when the target variable is categorical. It works on the concept of the entropy and is given by:

$$Information\, Gain = 1 - Entropy$$

Entropy is used for calculating the purity of a node. **Lower the value of entropy, higher is the purity of the node.** The entropy of a homogeneous node is zero. Since we subtract entropy from 1, the Information Gain is higher for the purer nodes with a maximum value of 1. Now, let's take a look at the formula for calculating the entropy:

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Steps to split a decision tree using Information Gain:

1. For each split, individually calculate the entropy of each child node
2. Calculate the entropy of each split as the weighted average entropy of child nodes
3. Select the split with the lowest entropy or highest information gain.

4. Until you achieve homogeneous nodes, repeat steps 1-3.

3. Chi-Square:

It is an algorithm to find out the statistical significance between the differences between sub-nodes and parent nodes. We measure it by the sum of squares of standardized differences between observed and expected frequencies of the target variable.

1. It works with the categorical target variable “Success” or “Failure”.
2. It can perform two or more splits.
3. Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node.
4. Chi-Square of each node is calculated using the formula,
5. Chi-square = $((\text{Actual} — \text{Expected})^{2} / \text{Expected})^{1}/2$
6. It generates a tree called CHAID (Chi-square Automatic Interaction Detector)

Steps to Calculate Chi-square for a split:

1. Calculate Chi-square for an individual node by calculating the deviation for Success and Failure both
2. Calculated Chi-square of Split using Sum of all Chi-square of success and Failure of each node of the split
3. Select the split where Chi-Square is maximum.

4. Reduction in Variance:

Reduction in variance is an algorithm used for continuous target variables (regression problems).

1. Used for continuous variables
2. This algorithm uses the standard formula of variance to choose the best split.
3. The split with lower variance is selected as the criteria to split the population

Steps to calculate Variance:

1. Calculate variance for each node.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

2. Calculate variance for each split as a weighted average of each node variance.
3. The node with lower variance is selected as the criteria to split.

Advantages of Decision Tree:

- Easy to Understand
- Useful in Data exploration
- Less Data cleaning is required
- Data type is not a constraint
- Non-Parametric model

Analysis:

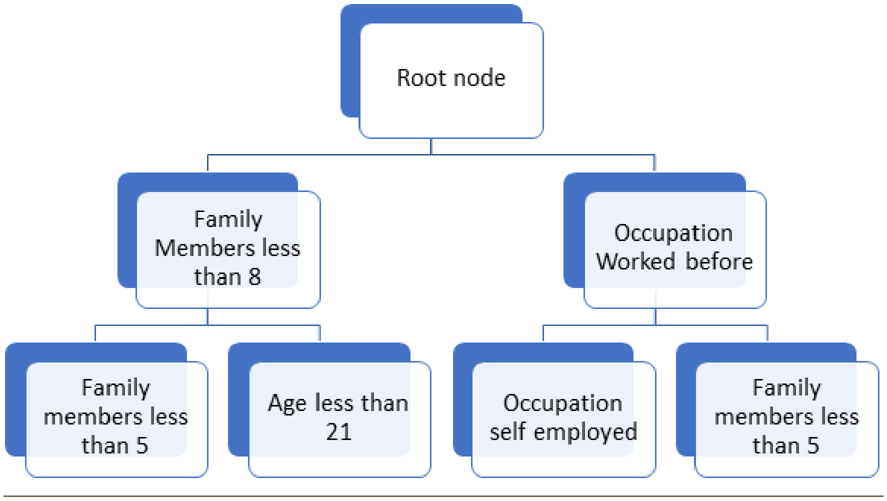
Training Dataset: The sample data used to fit model

Testing Dataset: The sample of data used to provide unbiased evaluation final fit model on Training dataset.

Model Information

Split Criterion	Entropy
Number of branches used	2
Maximum tree depth requested	3
Maximum tree depth achieved	3
Number of observations read	505

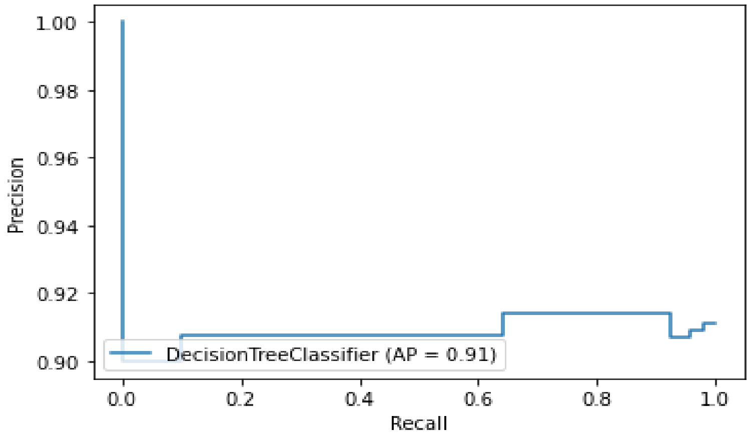
Number of observations used	505
Number of observations in training set	404
Number of observations in testing set	101


Root node
Family Members less than 8
Occupation Worked before
Family members less than 5
Age less than 21
Occupation self employed
Family members less than 5

Classification Report:

	Training Set	Testing Set
Precision	0.94	0.91
Recall	0.99	0.98
F-1 Score	0.97	0.94

Precision-Recall Curve:



Since precision-recall curves do not consider true negatives, they should only be used when specificity is of no concern for the classifier.

The AUC-PR curve for testing data shows a good model fit measure.

Interpretations:

- A person with family members less than 8 support paternal leave being made mandatory across all sectors.
- A person with family members less than 8 and age less than 21 does not support paternity leave being made mandatory across all sectors.
- A person with occupation who has worked before does not support paternity leave being made mandatory across all sectors.
- A person with occupation worked before and self employed supports paternity leave being made mandatory across all sectors.
- A person with occupation worked before and has family members less than 5 does not support paternity leave being made mandatory across all sectors.

Objective:

To study Gender disparities based on roles performed by both men and women.

Random Forest

Random forests or **random decision forests** is a method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct the decision trees habit of overfitting to their training set.

The entire random forest is built on top of weak learners (decision trees), giving you the analogy of using trees to make forest. The term “Random” indicates that each decision tree is built with a random subset of data.

Random forest is a flexible , easy to use machine learning algorithm that produces even without hyper-parameter tuning , a great result most of the time .It is also one of the most used algorithms, because of its simplicity and diversity.

Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Variables used:

- Personal details like Age, Gender, Marital Status, Occupation, Income etc
- Area you reside, Earning Members, Family Number
- Household Chores, Financial Responsibility, Looked after you
- Shop for groceries, Prepare food, Do cleaning and laundry

Feature Importance:

Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sklearn provides a great tool for this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results so the sum of all importance is equal to one.

By looking at the feature importance you can decide which features to possibly drop because they don’t contribute enough (or sometimes nothing at all) to the prediction process. This is important because a general rule in machine learning is that the more features you have the more likely your model will suffer from overfitting and vice versa.

Important Hyperparameters:

The hyperparameters in random forest are either used to increase the predictive power of the model or to make the model faster. Let's look at the hyperparameters of sklearns built-in random forest function.

1. **Increasing the predictive power**

   Firstly, there is the n_estimators hyperparameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.

Another important hyperparameter is max_features, which is the maximum number of features random forest considers to split a node. Sklearn provides several options, all described in the documentation.

The last important hyperparameter is min_sample_leaf. This determines the minimum number of leafs required to split an internal node.

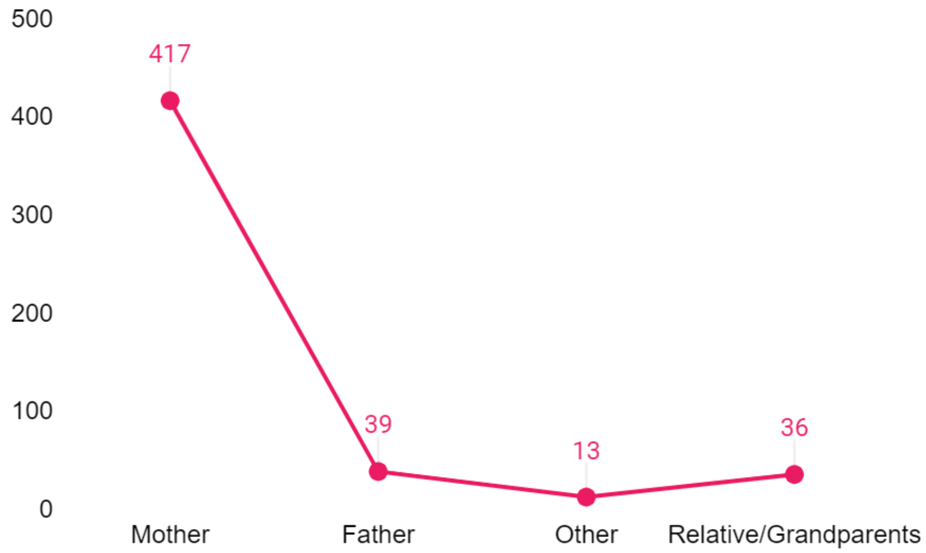
2. Increasing the model's speed

The n_jobs hyperparameter tells the engine how many processors it is allowed to use. If it has a value of one, it can only use one processor. A value of “-1” means that there is no limit.

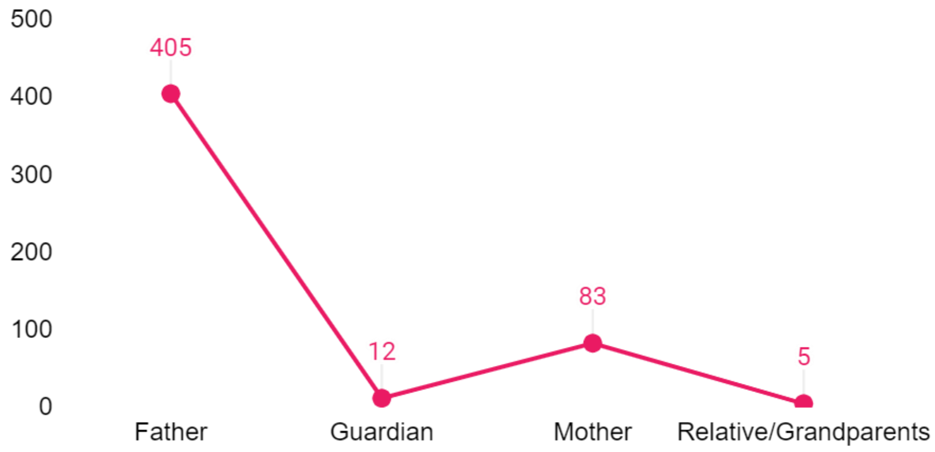
The random_state hyperparameter makes the model’s output replicable. The model will always produce the same results when it has a definite value of random_state and if it has been given the same hyperparameters and the same training data.

Interpretations

Exploratory Data Analysis



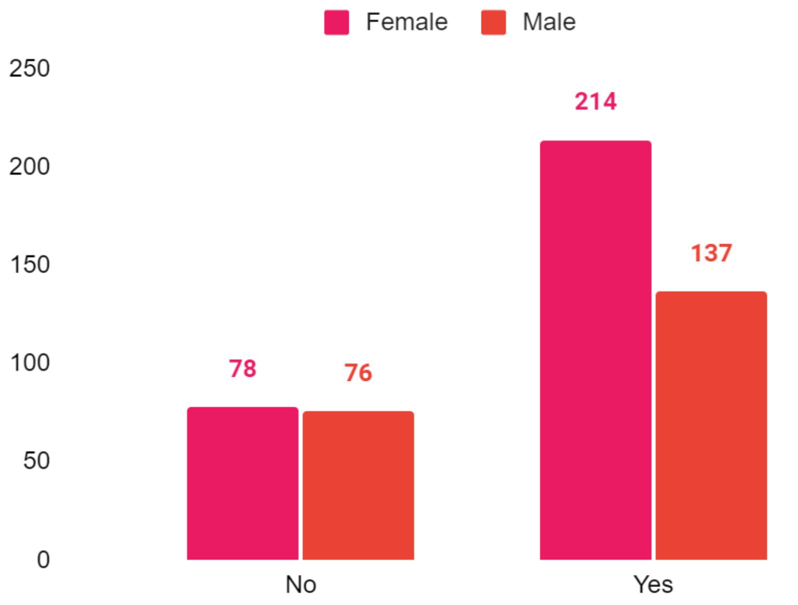
The Above is a representation of **who was responsible for upbringing of individuals while growing up**, it clearly shows that most of the time it is Mother who is supposed take this responsibility.



This is the graph of **who was responsible for financial needs of respondents** , again here it depicts the disparity by showing that its mostly the males taking up this financial responsibilities.

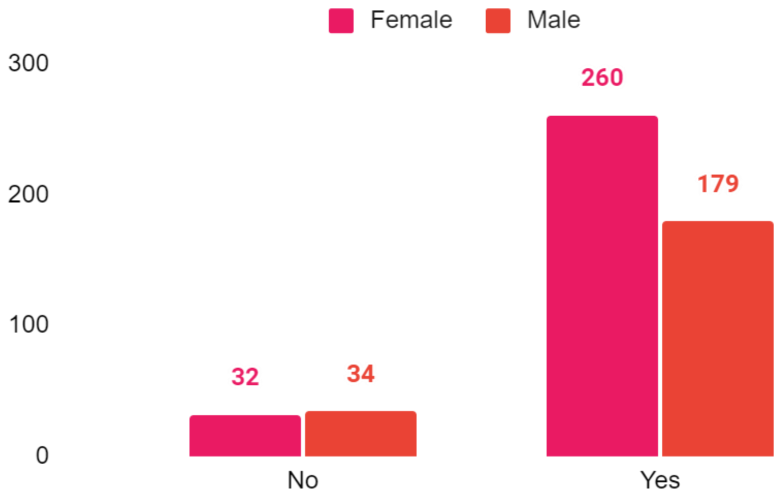
Disparities in day to day basic tasks

- ➔ Prepare Food



- ➔ Cleaning and Laundry

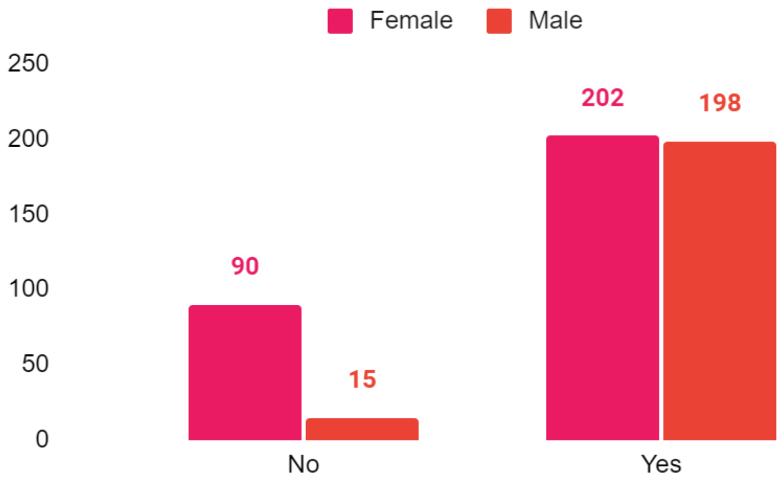
Here we can see the differe



Difference between males and females and whether they were taught this tasks while growing up. It shows that its women who are always taught more comparatively which indicates disparities between genders indirectly.

- Shop for groceries

This task shows equal distribution of both genders



Feature Importance :

The important features of our model with n_estimators=50 are;

Age	0.1632
Number of Family Members	0.0981
Occupation	0.0975
Qualification	0.0852

These variables contribute to 69% of accuracy of the Random Forest model, so by tackling and working on these variables we can improve and also overcome gender disparities in society.

OBJECTIVE:

To study the social representation and participation of men and women within families.

Association Rule

This algorithm, introduced by R Agarwal & R Srikant in 1994 has great significance in data mining. Name of the algorithm is Apriori because it uses prior knowledge of frequent item sets properties.

In data mining Association rules are “if - then” statements that help us to show probability of relationships between data items within large data sets in various types of databases.

Association rule mining has a number of applications, and are widely use to discover correlation in transactional data. This has applications in domains such as market basket analysis, sales, and marketing. Companies like Big Bazaar, Walmart, DMART, etc use association rules to analyze customer buying behavior.

A-priori Algorithm

DEFINITION: Apriori algorithm is a classical algorithm in data mining. It is used for mining the frequent item sets and relevant association rules

MEASURES:

SUPPORT: This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. In Table below, the support of {apple} is 4 out of 8, or 50%. Item sets can also contain multiple items. For instance, the support of {apple, beer, rice} is 2 out of 8, or 25%.

Support {apple} = 4/8

TID 1	Apple, beer, rice, meat
TID 2	Apple, beer, rice
TID 3	Apple, beer
TID 4	Apple, pear
TID 5	Milk, beer, rice, meat
TID 6	Milk, beer, rice

TID 7	Milk, beer
TID 8	Milk, pear

Table. Example of transaction

Support shows that Apple has 50% of overall transactions.

If you discover that sales of items beyond a certain proportion tend to have a significant impact on your profits, you might consider using that proportion as your *support threshold*. You may then identify item sets with support values above this threshold as significant item sets.

CONFIDENCE: This says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears. In Table 1, the confidence of {apple -> beer} is 3 out of 4, or 75%.

Confidence {Apple -> Beer} = support {Apple,beer}/support{Apple}

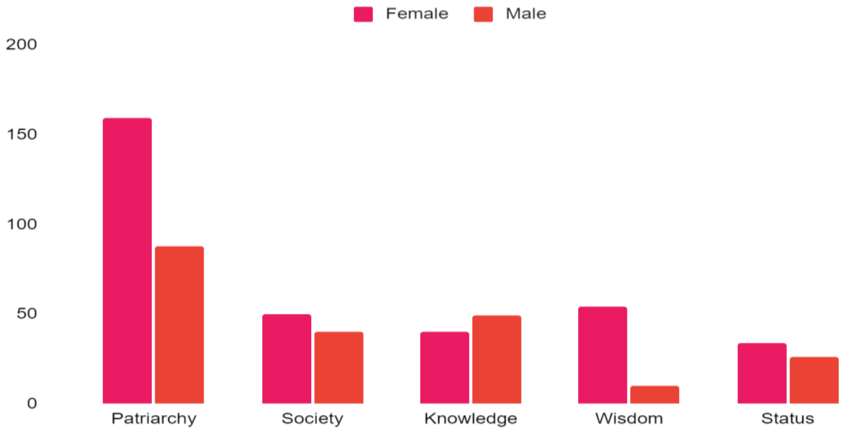
Confidence indicates that the 75% rule has been found to be true.

One drawback of the confidence measure is that it might misrepresent the importance of an association. This is because it only accounts for how popular apples are, but not beers. If beers are also very popular in general, there will be a higher chance that a transaction containing apples will also contain beers, thus inflating the confidence measure. To account for the base popularity of both constituent items, we use a third measure called lift.

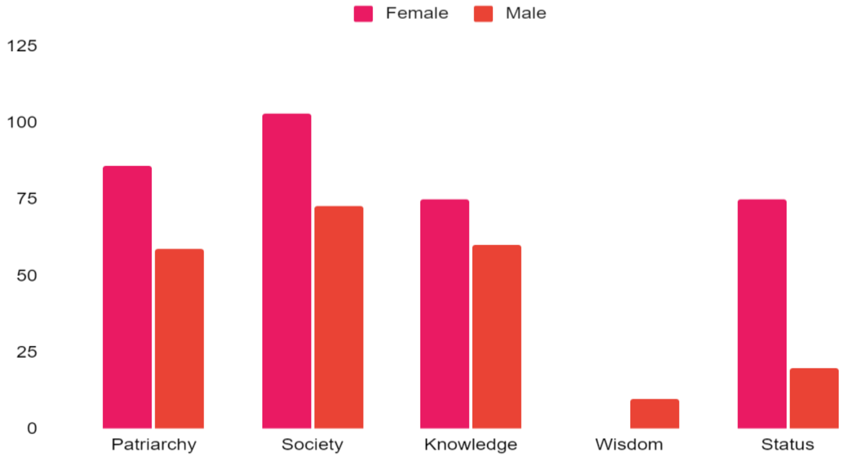
LIFT: This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. In Table 1, the lift of {apple -> beer} is 1, which implies no association between items. A lift value greater than 1 means that item Y is *likely* to be bought if item X is bought, while a value less than 1 means that item Y is *unlikely* to be bought if item X is bought.

Lift > 1 positive association, Lift =1 no association, Lift < 1 negative association.

Graphical representation of roles and position of Men within families determined by both genders,



Graphical representation of roles and position of Women determined by both genders,



- By using Association rule we are trying to find the main reasons for the determined roles and position of men and women within families.
- To find out relevant rules we used a-priori algorithm.
- In this case we set the minimum support as 0.03 and confidence as 0.2 and so we get the relevant rules generated as follows,

Association Rules

Rules	Support	Confidence	Lift
Female > Patriarchy	0.314852	0.544521	1.11329
Society > Female	0.099010	0.55556	1.09608
Society > Male	0.174257	0.44444	1.05370
Knowledge > Male	0.079208	0.550562	1.30533
Status > Male	0.051485	0.434333	1.03738
Female > Patriarchy	0.203960	0.352740	1.09959
Status > Female	0.051485	0.565217	1.09775
Male > Society	0.144554	0.342723	1.08852
Society > Male	0.144554	0.459119	1.08852
Knowledge > Male	0.118812	0.44444	1.05374

Results:

It has been observed that all the rules have been found to be significant.

The Lift value for all the rules is greater than "1", which indicates that the occurrence of the rule-body has a positive effect on the rule-head.

- If a person is a **female** then there is a high support of her selecting **Patriarchy** as one of the main **roles** of men with a high percentage of confidence.
- Whereas, the **male** respondents in our data chose **Financial status, Society, and knowledge** as the main positions of **men** within families and society.
- **Female** respondents in our data, after analyzing association rules, showed that **Patriarchy**, and **Financial status** as two of the main factors of the roles and participation of **women** within families and society.
- While, the **men** respondents believed that **Society, and Knowledge** are the two main factors that determine **women's** roles and position within families and society.

Objective:

To analyse people's opinions about the misuse of laws made to uplift Gender Equality in India.

K-Modes Clustering

Cluster Analysis is the unsupervised learning technique that finds the interesting patterns in the data objects without knowing class labels. Most of the real-world dataset consists of categorical data. Therefore, the k-modes clustering algorithm is the most widely used to group the categorical data.

K-Modes Clustering is a modified version of the standard k-means clustering process optimized to cluster categorical data. It does so by using the simple matching dissimilarity measure also referred to as the Hamming distance instead of the Euclidean distance to calculate the distance between two objects. Furthermore, it uses modes instead of means to represent the cluster centroids. While the k-means algorithm is a very popular choice when clustering numerical data, it performs poorly when applied to categorical data. The reason is that to cluster categorical data, the categorical values first have to be transformed into numerical values which distorts the clustering due to the usage of the Euclidean distance that leads the k-means algorithm to consider two distant values as close, simply based on the proximity of their numerical representations. A solution that works even with high dimensional categorical data is k-modes clustering.

Hamming Distance

As the data used in k modes are categorical, we can't calculate the Euclidean distance. So, what we are left with is dissimilarity measure also known as hamming distance.

Dissimilarity measure works on a simple principle;

D(xi, yj) = 0 if xi = yj   i= 1,2,.....k and j= 1,2.....n (sample size)

= 1 if xi is not equal to yj

where, xi is the value of attribute corresponding to ith cluster and yj is the value of the same attribute corresponding to jth observation

Steps for K-Modes Clustering

1. When clustering a categorical data set into k clusters using k-modes is to transform the categorical values into numerical values or dummy binary variables.

2. Randomly select k different objects out of the data set as the initial cluster centroids.

3. Thirdly, we have to calculate the distance between each object and centroid and

assign each object to the cluster containing its closest centroid.

4. Lastly, we select the new mode of each cluster centroid and compare it with the previous one. If the two modes differ, we repeat the last two steps. Thus, the modes

5. get updated with each iteration of the k-modes clustering process.

Analysis

Variables used:

- Age
- Gender
- Occupation
- Education
- Marital Status
- Statements related to laws

Note: All variables are of categorical type

Procedure:

1. We considered a sample of size 500 and based on it using CAO initialization, the initial value of k in k-modes was decided arbitrarily.
2. For k=2, unique centroids were obtained and corresponding modes were calculated depending upon all the attributes .
3. The modal value was then compared with the initial cluster centroids
4. Those modal values turned out to be the same as that of cluster centroids.
5. Based on this, k=2 was fixed and so three clusters were formed as:

Cluster 0: Aware and believe in misuse of laws

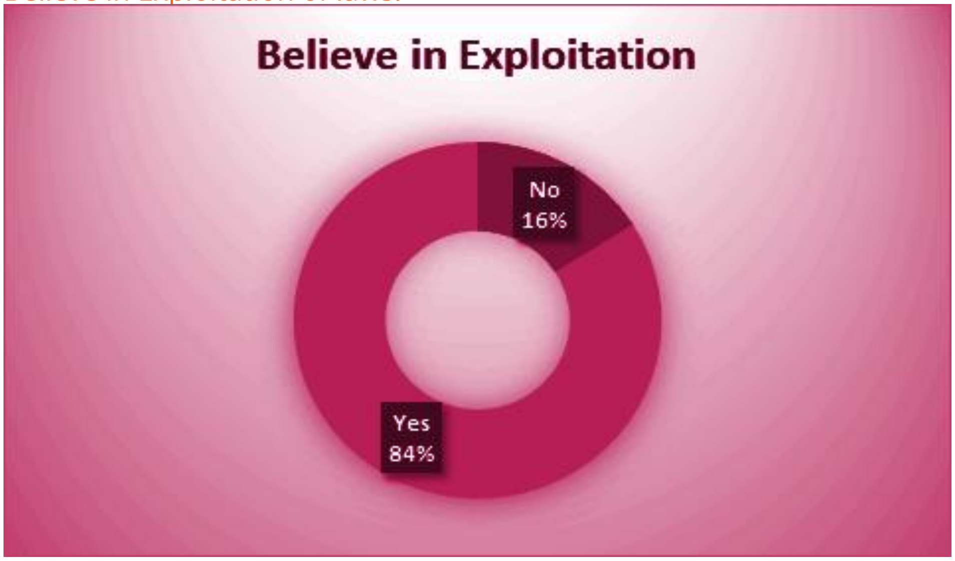
Cluster 1: Not aware and do not believe in misuse of laws

Cluster Centroid:

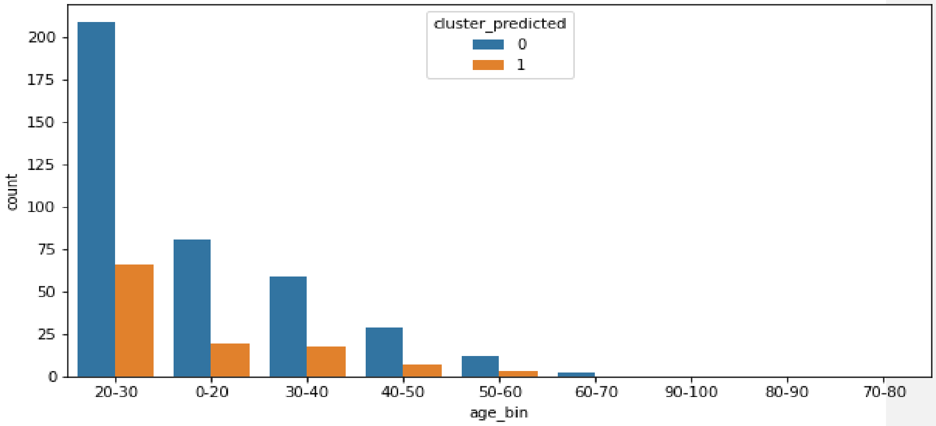
Age Group	**Gender**	**Qualification**	**Marital Status**	**Family Type**	**Clusters**
20-30	Female	P.G	Single	Nuclear Family	Aware and Believe

20-30	Male	P.G	Single	Joint Family	Not aware and do not believe

Believe in Exploitation of laws:



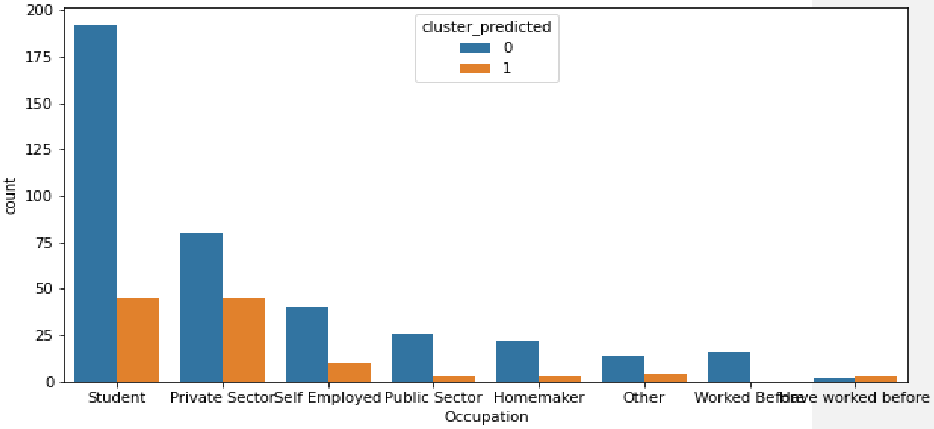
➢ Age



Interpretation:

According to our research people from Age Group 20-30 are more likely to be aware and also believe in misuse of laws as compared to people from higher Age Group.

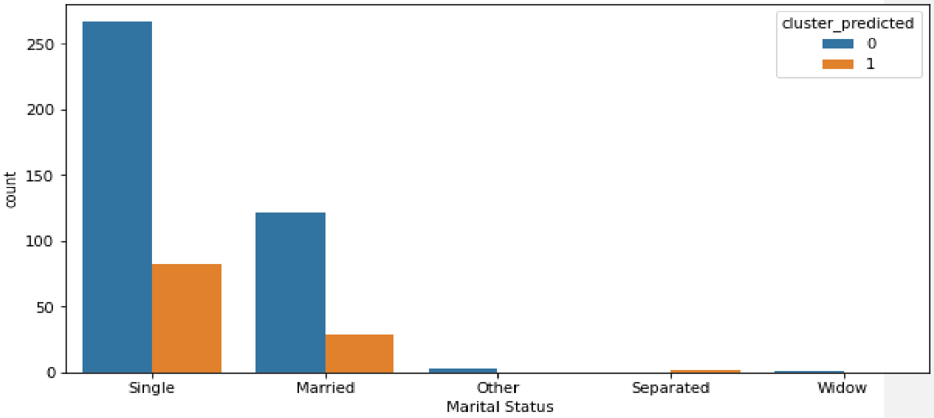
➢ Occupation



Interpretation:

The proportion of awareness is very high in students and working professionals as compared to the population who are homemakers and this can be the reason why there is biasness and inequality in our law system.

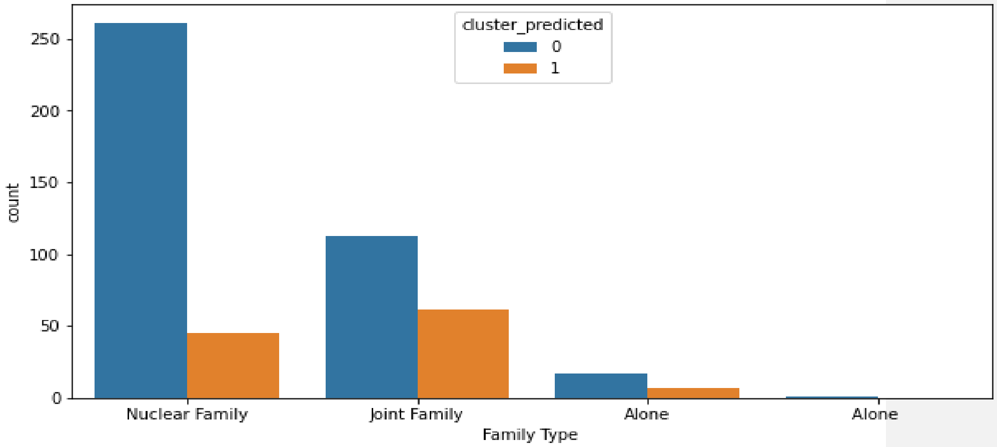
➢Marital Status



Interpretation:

We can observe a fair amount of distribution of awareness in this class.

➢ Family Type



Interpretation:

People belonging to Nuclear Family are more likely to be aware as compared to those who are from joint family.

Objective:

To study the impact of Covid-19 on Gender Inequality.

Odds Ratio

An odds ratio (OR) is a measure of association between a certain property A and a second property B in a population. Specifically, it tells you how the presence or absence of property A has an effect on the presence or absence of property B.

An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. Odds ratios are most commonly used in case-control studies, however they can also be used in cross-sectional and cohort study designs as well (with some modifications and/or assumptions). It can be calculated for a 2x2 contingency table, i.e. for variables having two categories.

How to calculate Odds Ratio?

		Dependent Variables	
		Y=1	Y=0
Independent Variable	X=1	a	b
	X=0	c	d

Odds Ratio = (a/c)/(b/d)

Interpretation of Odds Ratio:

If OR > 1, there is a positive association between outcome of dependent and independent variable.

If OR = 1, there is no association between outcome of dependent and independent variable.

If OR < 1, there is negative association between outcome of dependent and independent variable.

We calculated the odds using the Pivot Tables in excel for the following variables which were an activity with respect to gender. All the variables used are binary in nature. All the variables which denote activities were in accordance to the time spent on each of them during the pandemic whether for the particular activity the time spent has increased or decreased for it.

Variables Used:

- Gender (Male & Female)
- Gender Inequality during the Pandemic
- Duration for Work
- Duration for Studies
- Duration for Cooking
- Duration for Cleaning
- Duration for Gaming
- Duration for Reading
- Duration for Hobbies
- Duration for Self-Relaxation
- Duration for Childcare
- Duration for Sleep

Interpretation of Odds Ratio:

- ❖ Gender Inequality is 1.73 times more likely to be observed during the pandemic.
- ❖ For males, working hours are 1.11 times more likely to increase than females during the pandemic.
- ❖ For females, study hours are 0.64 times less likely to increase than males during the pandemic.
- ❖ For females, the cooking hours are 1.56 times more likely to increase than males during the pandemic.
- ❖ For females, the cleaning hours are 1.98 times more likely to increase than males during the pandemic.
- ❖ For males, the gaming hours are 1.68 times more likely to increase than females during the pandemic.
- ❖ For females, the reading hours are 0.96 times less likely to increase than males during the pandemic.
- ❖ For females, the time for hobbies is 0.92 times less likely to increase than males during the pandemic.
- ❖ For females, the self-relaxation hours are 0.66 times less likely to increase than males during the pandemic.

- For females, the time spent on childcare is 1.14 times more likely to increase than males during the pandemic.
- For females, the sleeping hours are 0.68 times less likely to increase than males during the pandemic.

Objective:

Effects of media on Gender Inequality.

Text Classification

NAÏVE BAYES TEXT CLASSIFIER:

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

Our training dataset should consist of a comment and corresponding labels as follows: Comments Label Can't believe blindly. Be careful while sharing data. positive Ok neutral Not secured at all negative

COMMENTS	LABELS
Media plays a big role in our society toward gender inequality	positive
Yes	neutral
No, it is all about mindset and family education	negative

ASSIGNING LABELS TO SENTENCES:

Since, our data just consists of comments, we need to assign labels to each comment to train our model. This was done by computing average sentiment scores for each sentence.

The Python output for average sentiment score is,

sentiment score is less than 0 then label assigned to the sentence is “negative”, if the sentiment score is 0 then label assigned to the sentence is “neutral” and if the sentiment score is greater than 0 then label assigned to the sentence is “positive”.

Data Pre-processing

As part of the preprocessing phase following is done

1. All the words in corpora (dataset) are converted to lowercase
2. Numbers are removed
3. Punctuation is removed
4. English stop words are removed
5. White space is removed

Note: English stop words are words like be, an, about, there, etc. Stop words are the words which do not add much meaning to the sentence.

Training Naive Bayes Model

We have divided the data in 80:20 ratio, of which 80% is train data and 20% is test data.

Understanding How Probability Predicts the Label for Test

Example

We consider the following example “chances misusing” which is the review from test data to show how probability predicts labels

Finding probability of review belonging to a class is done as follows

P(review belonging to class c) = $total\ number\ of\ training\ examples\ belonging\ to\ class\ c/total\ number\ of\ training\ examples$

and [P(test word j in class c)] = P(chances)*P(misusing)

Probability of review belonging to whichever particular class is higher, the

review will belong to that particular class (say positive, negative, neutral)

Laplace Correction

Suppose the word ‘chances’ has never occurred in training set then its probability will be 0 using the formula

P(test word j in class c) = $counts\ of\ word\ "j"\ in\ class\ c/counts\ of\ words\ in\ class\ c$

and which will eventually lead the probability of a review belonging to class c to be 0. In such a case we use Laplace correction which gives the formula for probability of test word j in class c as follows:

P(test word j in class c) = $counts\ of\ word\ "j"\ in\ class\ c+1/counts\ of\ words\ in\ class\ c+\ |V|+1$

where $|V|$ is vocabulary length and here vocabulary refers to all the unique words in the entire training data irrespective of class (positive, negative and neutral)

Why add 1 to the denominator?

We are considering that there are no more unknown test words. The numerator will always be “1” even when a word that never occurred in the training dataset occurs in the test set. In other words, we are assuming that an unknown test word occurred once (i.e. it’s count is 1) and so this also needs to be adjusted in the denominator. This is like adding the unknown word to the vocabulary of your training dataset, implying that the total count will be $|V| + 1$.

Python output:

COMMENTS	PRE-PROCESSED COMMENTS

Media plays a big role in our society toward gender inequality	media plays big role society towards gender inequality
No, it is all about mindset and family education	no mindset family education

Splitting the Data

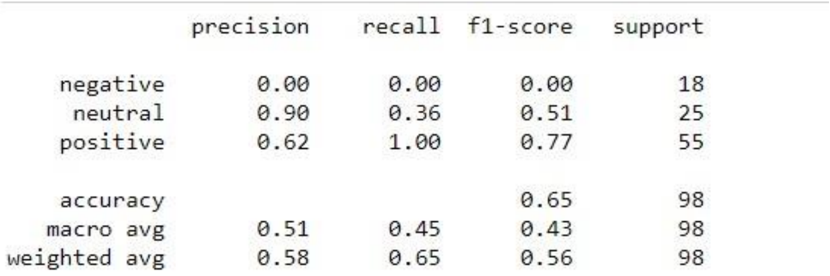
First, you separate the columns into dependent and independent variables(or features and label). Then you split those variables into a train and test set.

We have divided the data in 80:20 ratio, of which 80% is train data and 20% is test data.

Final Python Output:

Classification Report:

For Test

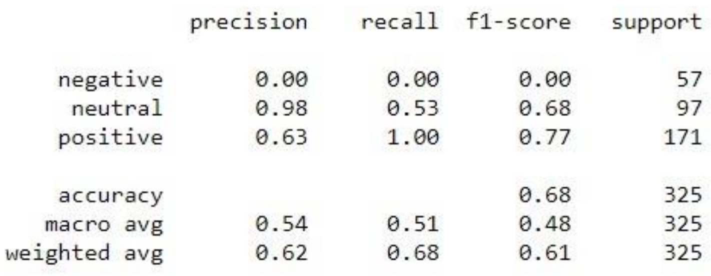


```
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00        18
     neutral       0.90      0.36      0.51        25
    positive       0.62      1.00      0.77        55

    accuracy                           0.65        98
   macro avg       0.51      0.45      0.43        98
weighted avg       0.58      0.65      0.56        98
```

For Train



```
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00        57
     neutral       0.98      0.53      0.68        97
    positive       0.63      1.00      0.77       171

    accuracy                           0.68       325
   macro avg       0.54      0.51      0.48       325
weighted avg       0.62      0.68      0.61       325
```

Mean Cross Validation Score: 0.62

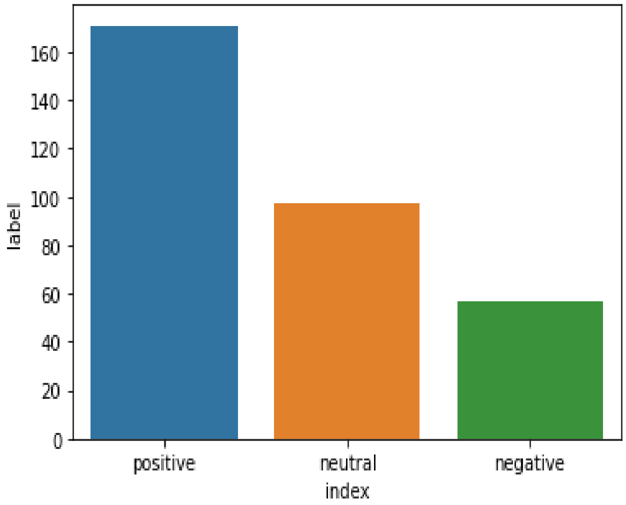
Model Accuracy: 0.68

The above is a text classifier model which helps to predict the probability of each word from a particular review and in turn knowing the probability of that particular review belonging to each of the three classes positive, negative and neutral. The review will belong to that particular class for which the probability is higher. This can be used to know the positive and negative review of people on the influence of the media on Gender Inequality and Gender Equality.

WORDCLOUD:

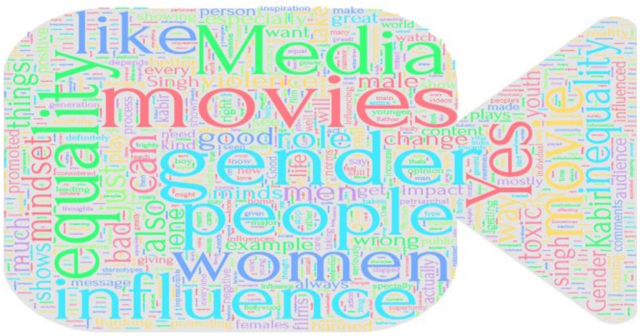
A word cloud is a graphical representation of frequently used words in a collection of text files. The height of each word in this picture is an indication of frequency of occurrence of the word in the entire text. It is a good way to convey the general idea of the text.

To accomplish this task, first the overall responses were collected, cleaned, preprocessed and visualized to understand what views were being expressed by the people on ‘Influence of Media on Gender Inequality'.



This graph shows the count of different opinions expressed by people on “Influence of Media on Gender Inequality”.

Majority of them had a positive opinion that media does influence gender inequality in a negative way.



From the masked cloud we can understand the most expressed words are: **MEDIA, MOVIES, GENDER, INFLUENCE, INEQUALITY, MINDSET, YOUTH and WOMEN.**

1. The media tends to demean men in caring or domestic roles, or those who oppose violence. Such portrayals can influence perceptions in terms of what society may expect from men and women, but also what they may expect from themselves. They promote an unbalanced vision of the roles of women and men in society.

2. young age, children are influenced by the gendered stereotypes that media present to them.

3. Women are frequently portrayed in stereotypical and hyper-sexualised roles in advertising and the film industry, which has long-term social consequences.

4. We strongly believe in the transformative role media can play in achieving gender equality in societies. By creating gender-sensitive and gender-transformative content and breaking gender stereotypes.

Objective:

To study and identify factors promoting Gender Equality in Society.

Factor Analysis

Factor analysis is one of the unsupervised machine learning algorithms which is used for dimensionality reduction. This algorithm creates factors from the observed variables to represent the common variance i.e. variance due to correlation among the observed variables.

It is a way to find hidden patterns, show how those patterns overlap and show what characteristics are seen in multiple patterns. It is also used to create a set of variables for similar items in the set (these sets of variables are called dimensions). It can be a very useful tool for complex sets of data involving psychological studies, socioeconomic status and other involved concepts. A "factor" is a set of observed variables that have similar response patterns; They are associated with a hidden variable (called a confounding variable) that is not directly measured. Factors are listed according to factor loadings, or how much variation in the data they can explain.

Types of Factoring:

There are different types of methods to extract factors from data;

1. **Principal Component Analysis:** This is the most common method used by researchers. PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.
2. **Common factor analysis**: The second most preferred method by researchers, it extracts the common variance and puts them into factors. This method does not include the unique variance of all variables.
3. **Maximum likelihood method:** This method also works on correlation metric but it uses maximum likelihood method to factor.
4. Other methods of factor analysis: Alfa factoring outweighs least squares. Weight square is another regression-based method which is used for factoring.

In our project, we have used Principal Component Analysis to obtain factors.

Key terms and Concepts:

Factor Loading: Factor loading is the correlation coefficient for the variable and factor. Factor loading shows the variance explained by the variable on that particular factor. In the

SEM approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable.

Eigenvalues: Eigenvalues are also called characteristic roots. Eigenvalues show variance explained by that particular factor out of the total variance. From the commonality column, we can know how much variance is explained by the first factor out of the total variance. For example, if our first factor explains 68% variance out of the total, this means that 32% variance will be explained by the other factor.

Criteria for determining the number of factors: According to the Kaiser Criterion, Eigenvalues is a good criterion for determining a factor. If Eigenvalues is greater than one, we should consider that a factor and if Eigenvalues is less than one, then we should not consider that a factor. According to the variance extraction rule, it should be more than 0.7. If variance is less than 0.7, then we should not consider that a factor.

Rotation method: Rotation method makes it more reliable to understand the output. Eigenvalues do not affect the rotation method, but the rotation method affects the Eigenvalues or percentage of variance extracted. There are several rotation methods available: (1) Varimax rotation method, (2) Quartimax rotation method, (3) Direct Oblimin rotation method, and (4) Promax rotation method.

In our project, we have used Varimax Rotation to obtain factors.

Variables used:

X1: Women's role to take care of home

X2: Men's role to take care of financial needs

X3: Is Gender Pay Gap justified

X4: Does Gender Inequality have negative effects on society

X5: Is awareness needed to promote Gender Equality

X6: Is Gender Equality an important aspect in development of society

X7: Everyone benefits from Gender Equality

X8: Prevents violence

X9: Independence

X10: Good for economy

X11: Human Right

X12: Equality makes community safer and healthier

X13: Women entrepreneur/ empowerment

X14: Equal opportunities

X15: Trustworthy and progressive society

X16: Gender equality meaning men and women are equal has come a long way

X17: Efforts to achieve gender equality benefits mostly well to do people

X18: Equal contribution of both men and women in planning and decision making processes

Before proceeding with factor analysis on the variables we need to check whether factor analysis is appropriate for our data and are the variables correlated with each other which are the basic assumption for factor analysis.

KMO and Bartlett's test of sphericity

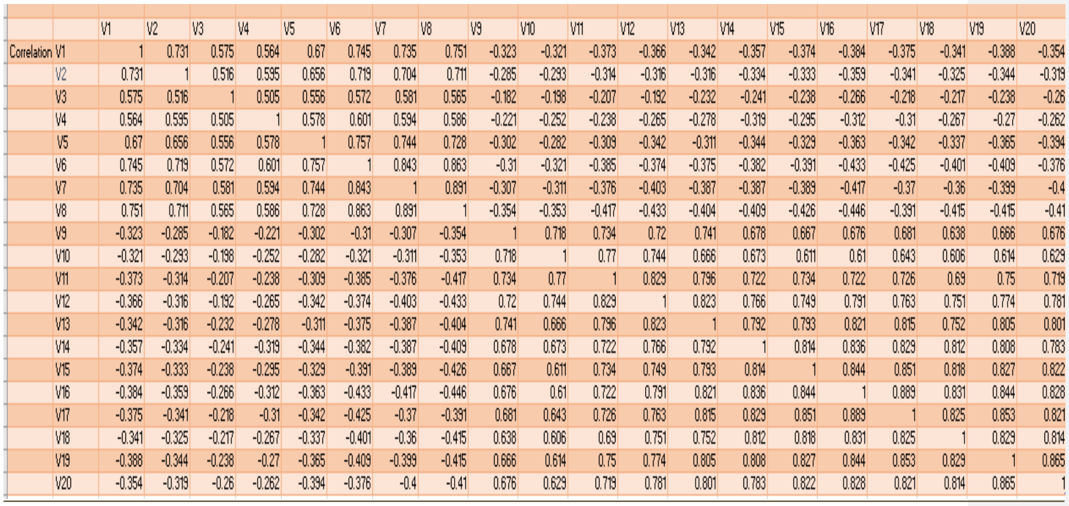
The Kaiser-Meyer-Olkin Measure of Sampling Adequacy is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors. High values (close to 1 .0) generally indicate that factor analysis may be useful with your data. If the value is less than 0.50, the results of the factor analysis probably won't be very useful.

Bartlett's test of sphericity tests the hypothesis that your correlation matrix is an identity matrix, which would indicate that your variables are: unrelated and therefore unsuitable for structure detection. Small values (less than 0.05) of the significance level indicate that factor analysis may be useful with your data.

Bartlett's test of Sphericity tests whether the correlation matrix is an identity matrix, which would indicate that the factor model is inappropriate.

Ho: Correlation matrix is an identity matrix.

H1: Correlation matrix is not an identity matrix.



		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
Correlation	V1	1	0.731	0.575	0.564	0.67	0.745	0.735	0.751	-0.323	-0.321	-0.373	-0.366	-0.342	-0.357	-0.374	-0.384	-0.375	-0.341	-0.388	-0.354
	V2	0.731	1	0.516	0.595	0.656	0.719	0.704	0.711	-0.285	-0.293	-0.314	-0.316	-0.316	-0.334	-0.333	-0.359	-0.341	-0.325	-0.344	-0.319
	V3	0.575	0.516	1	0.505	0.556	0.572	0.581	0.565	-0.182	-0.198	-0.207	-0.192	-0.232	-0.241	-0.238	-0.266	-0.218	-0.217	-0.238	-0.26
	V4	0.564	0.595	0.505	1	0.578	0.601	0.594	0.586	-0.221	-0.252	-0.238	-0.265	-0.278	-0.319	-0.295	-0.312	-0.31	-0.267	-0.27	-0.262
	V5	0.67	0.656	0.556	0.578	1	0.757	0.744	0.728	-0.302	-0.282	-0.309	-0.342	-0.311	-0.344	-0.329	-0.363	-0.342	-0.337	-0.365	-0.394
	V6	0.745	0.719	0.572	0.601	0.757	1	0.843	0.863	-0.31	-0.321	-0.385	-0.374	-0.375	-0.382	-0.391	-0.433	-0.425	-0.401	-0.409	-0.376
	V7	0.735	0.704	0.581	0.594	0.744	0.843	1	0.891	-0.307	-0.311	-0.376	-0.403	-0.387	-0.387	-0.389	-0.417	-0.37	-0.36	-0.399	-0.4
	V8	0.751	0.711	0.565	0.586	0.728	0.863	0.891	1	-0.354	-0.353	-0.417	-0.433	-0.404	-0.409	-0.426	-0.446	-0.391	-0.415	-0.415	-0.41
	V9	-0.323	-0.285	-0.182	-0.221	-0.302	-0.31	-0.307	-0.354	1	0.718	0.734	0.72	0.741	0.678	0.667	0.676	0.681	0.638	0.666	0.676
	V10	-0.321	-0.293	-0.198	-0.252	-0.282	-0.321	-0.311	-0.353	0.718	1	0.77	0.744	0.666	0.673	0.611	0.61	0.643	0.606	0.614	0.629
	V11	-0.373	-0.314	-0.207	-0.238	-0.309	-0.385	-0.376	-0.417	0.734	0.77	1	0.829	0.796	0.722	0.734	0.722	0.726	0.69	0.75	0.719
	V12	-0.366	-0.316	-0.192	-0.265	-0.342	-0.374	-0.403	-0.433	0.72	0.744	0.829	1	0.823	0.766	0.749	0.791	0.763	0.751	0.774	0.781
	V13	-0.342	-0.316	-0.232	-0.278	-0.311	-0.375	-0.387	-0.404	0.741	0.666	0.796	0.823	1	0.792	0.793	0.821	0.815	0.752	0.805	0.801
	V14	-0.357	-0.334	-0.241	-0.319	-0.344	-0.382	-0.387	-0.409	0.678	0.673	0.722	0.766	0.792	1	0.814	0.836	0.829	0.812	0.808	0.783
	V15	-0.374	-0.333	-0.238	-0.295	-0.329	-0.391	-0.389	-0.426	0.667	0.611	0.734	0.749	0.793	0.814	1	0.844	0.851	0.818	0.827	0.822
	V16	-0.384	-0.359	-0.266	-0.312	-0.363	-0.433	-0.417	-0.446	0.676	0.61	0.722	0.791	0.821	0.836	0.844	1	0.889	0.831	0.844	0.828
	V17	-0.375	-0.341	-0.218	-0.31	-0.342	-0.425	-0.37	-0.391	0.681	0.643	0.726	0.763	0.815	0.829	0.851	0.889	1	0.825	0.853	0.821
	V18	-0.341	-0.325	-0.217	-0.267	-0.337	-0.401	-0.36	-0.415	0.638	0.606	0.69	0.751	0.752	0.812	0.818	0.831	0.825	1	0.829	0.814
	V19	-0.388	-0.344	-0.238	-0.27	-0.365	-0.409	-0.399	-0.415	0.666	0.614	0.75	0.774	0.805	0.808	0.827	0.844	0.853	0.829	1	0.865
	V20	-0.354	-0.319	-0.26	-0.262	-0.394	-0.376	-0.4	-0.41	0.676	0.629	0.719	0.781	0.801	0.783	0.822	0.828	0.821	0.814	0.865	1

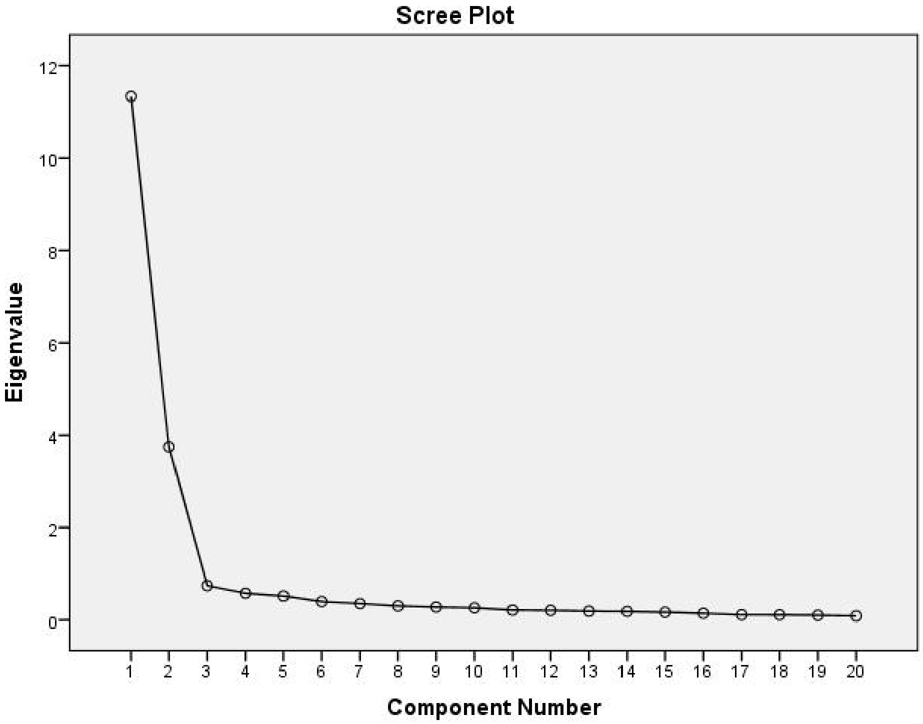
Kaiser Meyer Olkin Measure of Sampling Adequacy		0.959
Bartlett's Test of Sphericity	Approx Chi-Square	11419.748
	Degree of Freedom	190
	Sig.	.000

From the above table the KMO value > 0.5 which shows factor Analysis is useful for our data.

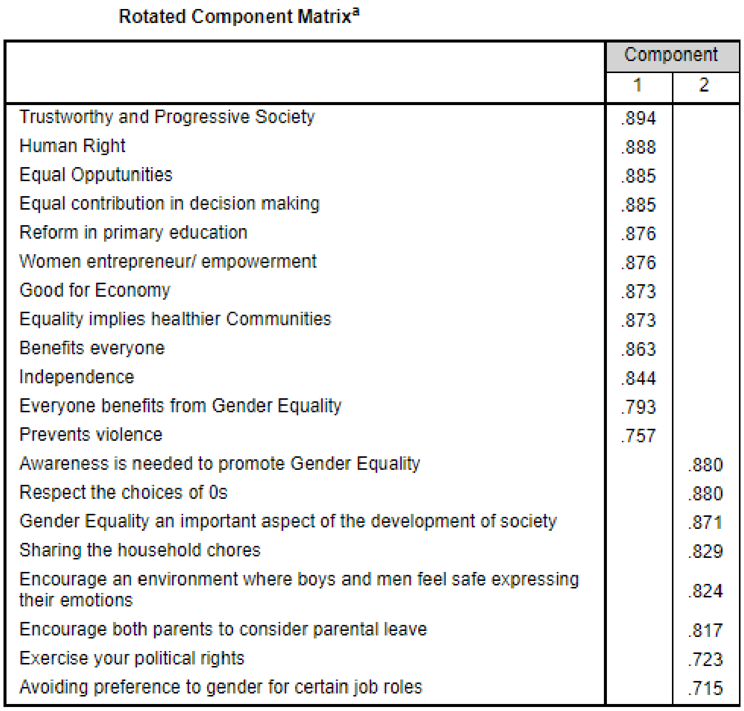
In this case, sig value 0 suggests that we should reject null hypothesis i.e. correlation matrix is not identity matrix i.e. good correlation exists between input attributes.

Scree Plot

A scree plot displays Eigenvalues associated with the component of factors in descending order versus the number of component or factors.



This scree plot shows that 2 factors explains most of the variability because the line starts to be parallel to the x axis after factor 2. The remaining factors explain a very small proportion of the variability and are likely not important.



Rotated Component Matrix[a]

	Component	
	1	2
Trustworthy and Progressive Society	.894	
Human Right	.888	
Equal Opputunities	.885	
Equal contribution in decision making	.885	
Reform in primary education	.876	
Women entrepreneur/ empowerment	.876	
Good for Economy	.873	
Equality implies healthier Communities	.873	
Benefits everyone	.863	
Independence	.844	
Everyone benefits from Gender Equality	.793	
Prevents violence	.757	
Awareness is needed to promote Gender Equality		.880
Respect the choices of 0s		.880
Gender Equality an important aspect of the development of society		.871
Sharing the household chores		.829
Encourage an environment where boys and men feel safe expressing their emotions		.824
Encourage both parents to consider parental leave		.817
Exercise your political rights		.723
Avoiding preference to gender for certain job roles		.715

Here we see that the loadings of each variable in Component 1 and 2 , we will only extract those variables which have higher loadings in Component 1 and 2.

	Variable	Factor 1	Factor2
V1	Trustworthy and Progressive Society	.894	
V2	Human Rights	.888	
V3	Equal contribution in decision making	.885	
V4	Women empowerment/entrepreneur	.876	

V5	Good for Economy	.873	
V6	Benefits everyone	.873	
V7	Independence	.844	
V9	Awareness is needed		.880
V10	Respect choice of others		.880
V11	Sharing the household		.829
V12	Encourage environment where everyone can express their emotions		.824
V13	Encourage both parents to consider parental leave		.817
V15	Avoiding preference for certain role		.723

<u>Total variance explained by this 2 factors are ;</u>

Factors	% of Variance	Cumulative %
Factor 1	46.046	46.066
Factor 2	29.349	75.395

Based on factor loadings obtained from the matrix we can classify the variables into 2 factors as follows

➔ **SOCIETY**

- **Trustworthy and Progresive Society**
- **Human Rights**
- **Equal contribution in decision making**
- **Women empowerment**
- **Good for economy**
- **Benefits everyone**
- **Independence**

➔ **At Domestic Level**

- **Awareness is needed**
- **Respect choice of others**
- **Sharing the household**
- **Encourage environment where everyone can express their emotions**
- **Avoiding preference for certain role**

Conclusion

I. Using Pareto Analysis we got to know that main reasons for Gender Inequality are Illiteracy, Culture, Influence of Society and Gender Norms. By curbing this reason we can definitely hope for a better tomorrow.

II. Using K-means clustering we studied gender inequality at the workplace and how people can be distributed in different clusters.

III. Using Decision trees we got to know who exactly in our research support paternity leave being made mandatory in all sectors.

IV. Using Random Forest and A-Priori algorithm we studied disparities in both genders and what determines Men's and Women's role in society and their participation.

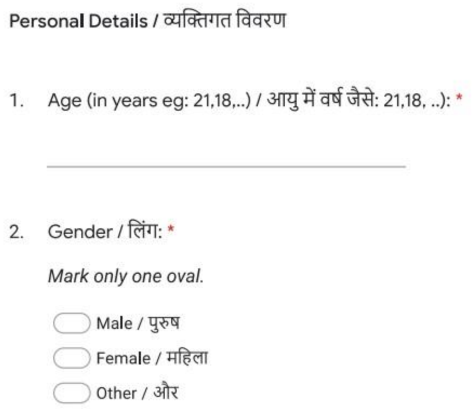
V. Using K-Modes clustering we got to know who among our respondents are aware about misuse of laws , mainly about laws which are biased against men and which leads to inequality.

VI. Odds Ratio helped us to study how Covid-19 is impacting the inequality between both genders. Who are more likely to overwork , experience lack of sleep, spend more time cooking.

VII. Using Sentiment and Text Classifier we analyzed opinions of people about how media influences inequality.

VIII. Using Factor Analysis we also got 2 factors viz Society and Domestic Levels which can help us work on how to promote equality.

Questionnaire



Personal Details / व्यक्तिगत विवरण

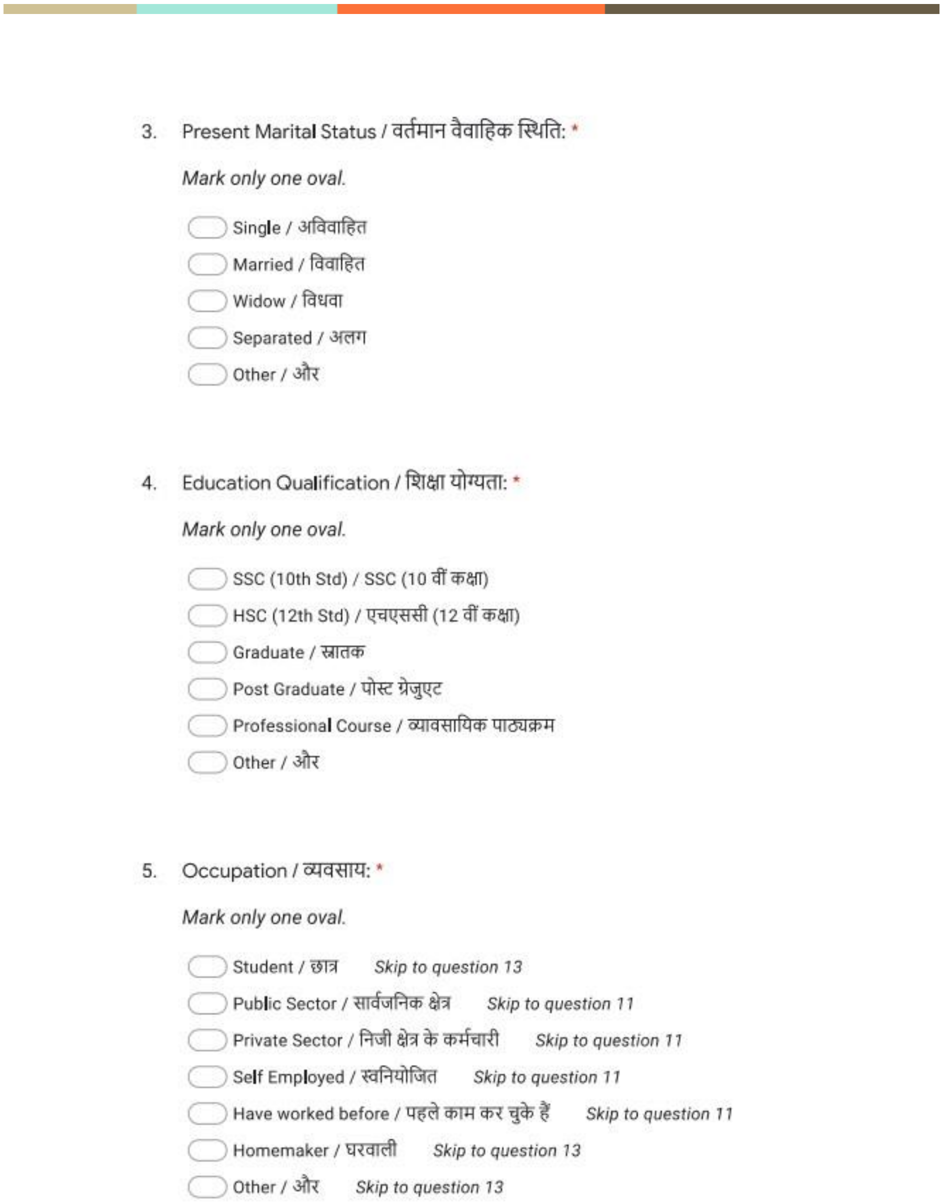
1. Age (in years eg: 21,18,..) / आयु में वर्ष जैसे: 21,18, ..): *

   ____

2. Gender / लिंग: *

   Mark only one oval.

   ○ Male / पुरुष
   ○ Female / महिला
   ○ Other / और



3. Present Marital Status / वर्तमान वैवाहिक स्थिति: *

Mark only one oval.

○ Single / अविवाहित

○ Married / विवाहित

○ Widow / विधवा

○ Separated / अलग

○ Other / और

4. Education Qualification / शिक्षा योग्यता: *

Mark only one oval.

○ SSC (10th Std) / SSC (10 वीं कक्षा)

○ HSC (12th Std) / एचएससी (12 वीं कक्षा)

○ Graduate / स्नातक

○ Post Graduate / पोस्ट ग्रेजुएट

○ Professional Course / व्यावसायिक पाठ्यक्रम

○ Other / और

5. Occupation / व्यवसाय: *

Mark only one oval.

○ Student / छात्र *Skip to question 13*

○ Public Sector / सार्वजनिक क्षेत्र *Skip to question 11*

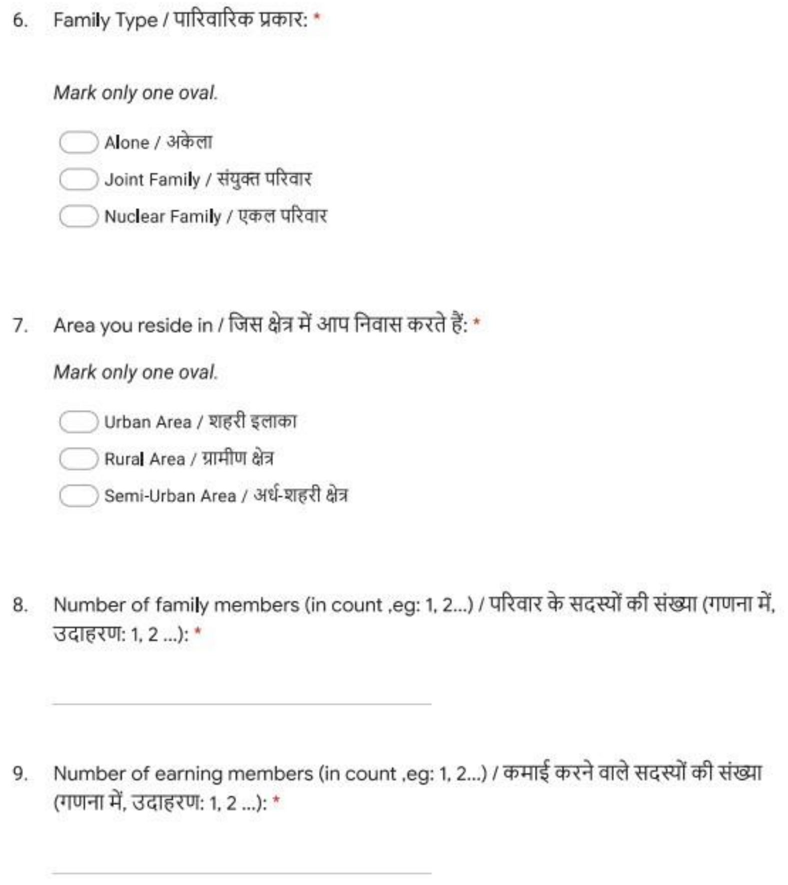
○ Private Sector / निजी क्षेत्र के कर्मचारी *Skip to question 11*

○ Self Employed / स्वनियोजित *Skip to question 11*

○ Have worked before / पहले काम कर चुके हैं *Skip to question 11*

○ Homemaker / घरवाली *Skip to question 13*

○ Other / और *Skip to question 13*



6. Family Type / पारिवारिक प्रकार: *

Mark only one oval.

○ Alone / अकेला

○ Joint Family / संयुक्त परिवार

○ Nuclear Family / एकल परिवार

7. Area you reside in / जिस क्षेत्र में आप निवास करते हैं: *

Mark only one oval.

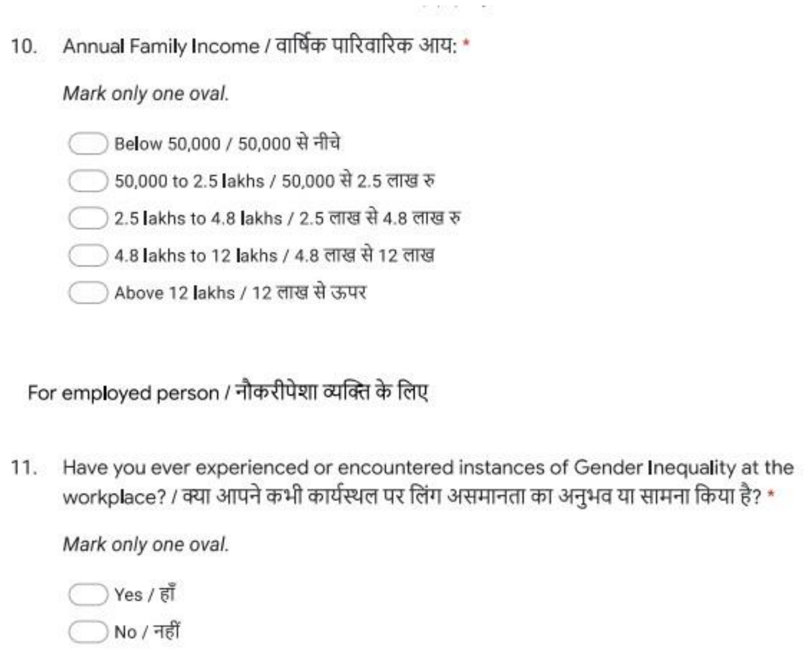
○ Urban Area / शहरी इलाका

○ Rural Area / ग्रामीण क्षेत्र

○ Semi-Urban Area / अर्ध-शहरी क्षेत्र

8. Number of family members (in count ,eg: 1, 2...) / परिवार के सदस्यों की संख्या (गणना में, उदाहरण: 1, 2 ...): *

9. Number of earning members (in count ,eg: 1, 2...) / कमाई करने वाले सदस्यों की संख्या (गणना में, उदाहरण: 1, 2 ...): *



10. Annual Family Income / वार्षिक पारिवारिक आय: *

Mark only one oval.

○ Below 50,000 / 50,000 से नीचे

○ 50,000 to 2.5 lakhs / 50,000 से 2.5 लाख रु

○ 2.5 lakhs to 4.8 lakhs / 2.5 लाख से 4.8 लाख रु

○ 4.8 lakhs to 12 lakhs / 4.8 लाख से 12 लाख

○ Above 12 lakhs / 12 लाख से ऊपर

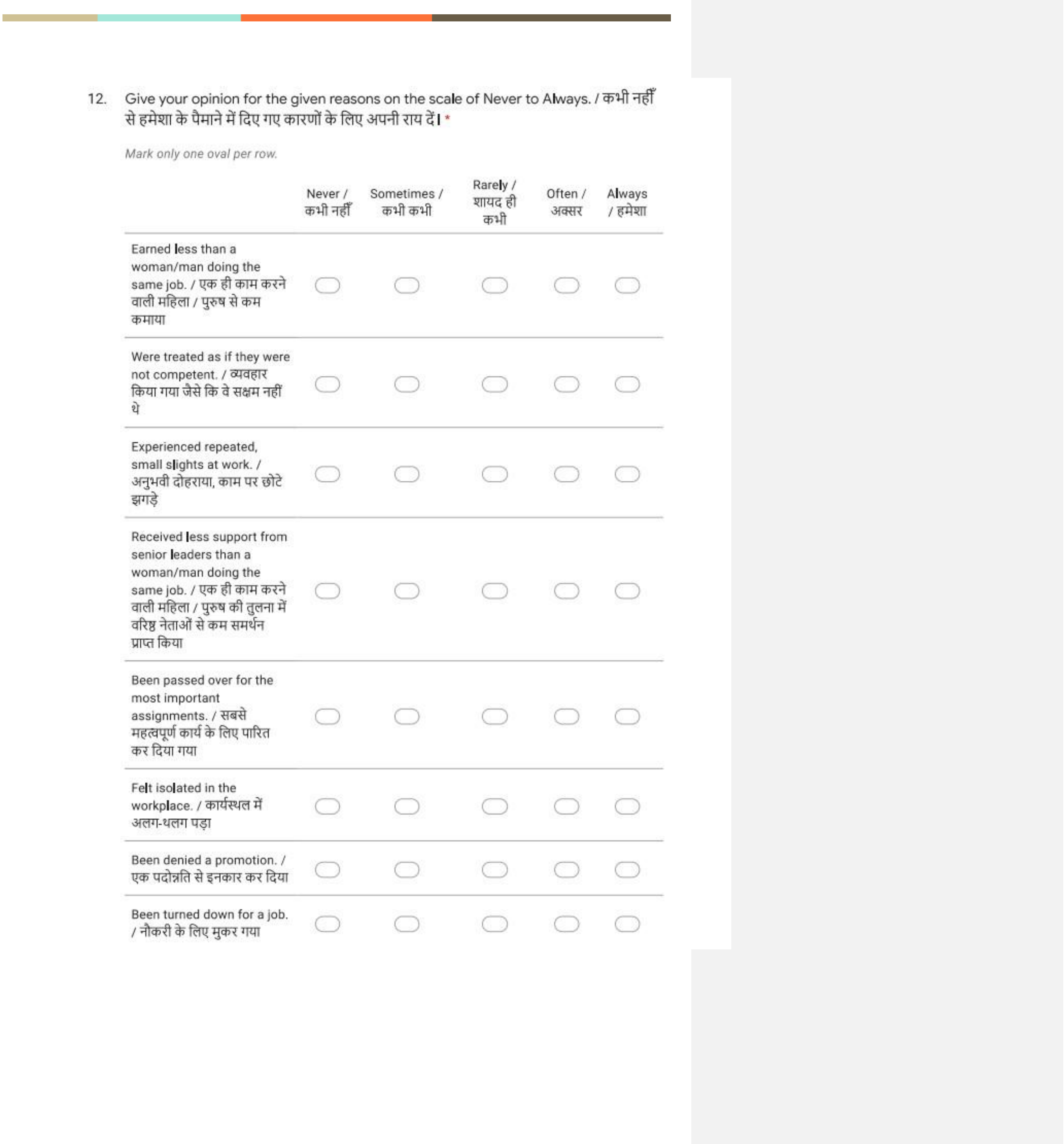
For employed person / नौकरीपेशा व्यक्ति के लिए

11. Have you ever experienced or encountered instances of Gender Inequality at the workplace? / क्या आपने कभी कार्यस्थल पर लिंग असमानता का अनुभव या सामना किया है? *

Mark only one oval.

○ Yes / हाँ

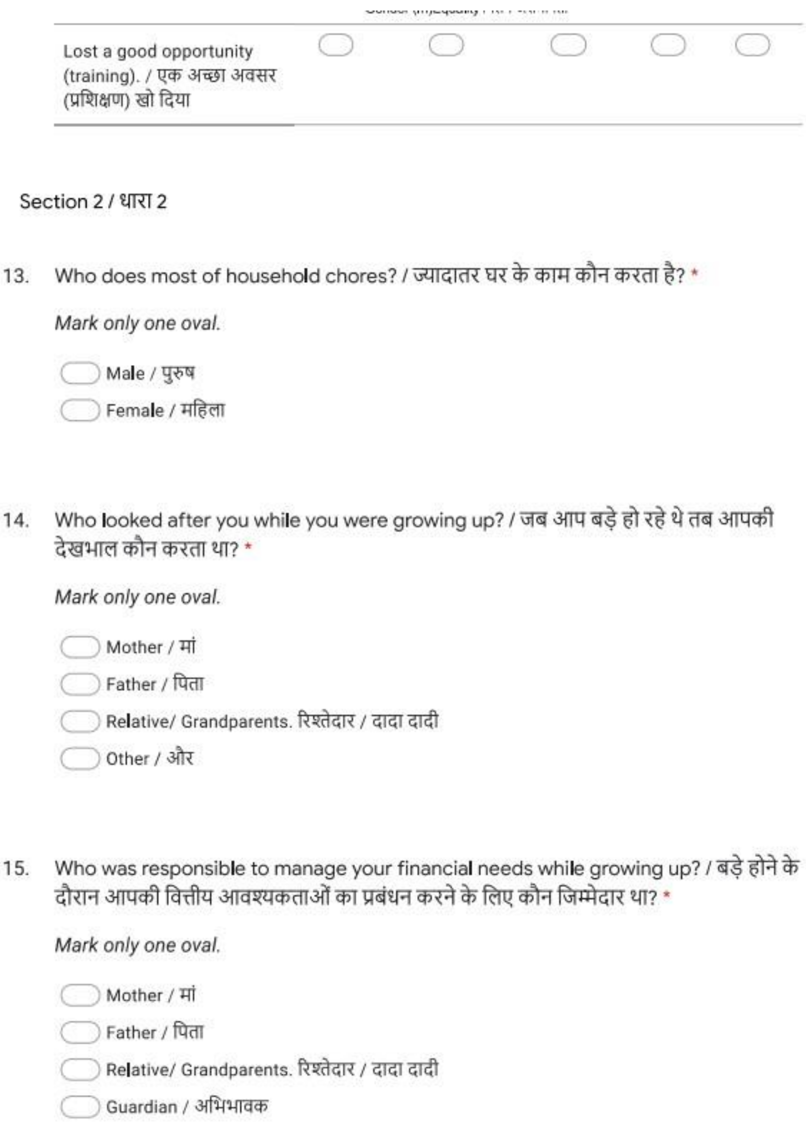
○ No / नहीं



12. Give your opinion for the given reasons on the scale of Never to Always. / कभी नहीँ से हमेशा के पैमाने में दिए गए कारणों के लिए अपनी राय दें। *

Mark only one oval per row.

	Never / कभी नहीँ	Sometimes / कभी कभी	Rarely / शायद ही कभी	Often / अक्सर	Always / हमेशा
Earned less than a woman/man doing the same job. / एक ही काम करने वाली महिला / पुरुष से कम कमाया	○	○	○	○	○
Were treated as if they were not competent. / व्यवहार किया गया जैसे कि वे सक्षम नहीं थे	○	○	○	○	○
Experienced repeated, small slights at work. / अनुभवी दोहराया, काम पर छोटे झगड़े	○	○	○	○	○
Received less support from senior leaders than a woman/man doing the same job. / एक ही काम करने वाली महिला / पुरुष की तुलना में वरिष्ठ नेताओं से कम समर्थन प्राप्त किया	○	○	○	○	○
Been passed over for the most important assignments. / सबसे महत्वपूर्ण कार्य के लिए पारित कर दिया गया	○	○	○	○	○
Felt isolated in the workplace. / कार्यस्थल में अलग-थलग पड़ा	○	○	○	○	○
Been denied a promotion. / एक पदोन्नति से इनकार कर दिया	○	○	○	○	○
Been turned down for a job. / नौकरी के लिए मुकर गया	○	○	○	○	○



Lost a good opportunity (training). / एक अच्छा अवसर (प्रशिक्षण) खो दिया	○	○	○	○	○

Section 2 / धारा 2

13. Who does most of household chores? / ज्यादातर घर के काम कौन करता है? *

Mark only one oval.

○ Male / पुरुष

○ Female / महिला

14. Who looked after you while you were growing up? / जब आप बड़े हो रहे थे तब आपकी देखभाल कौन करता था? *

Mark only one oval.

○ Mother / मां

○ Father / पिता

○ Relative/ Grandparents. रिश्तेदार / दादा दादी

○ Other / और

15. Who was responsible to manage your financial needs while growing up? / बड़े होने के दौरान आपकी वित्तीय आवश्यकताओं का प्रबंधन करने के लिए कौन जिम्मेदार था? *

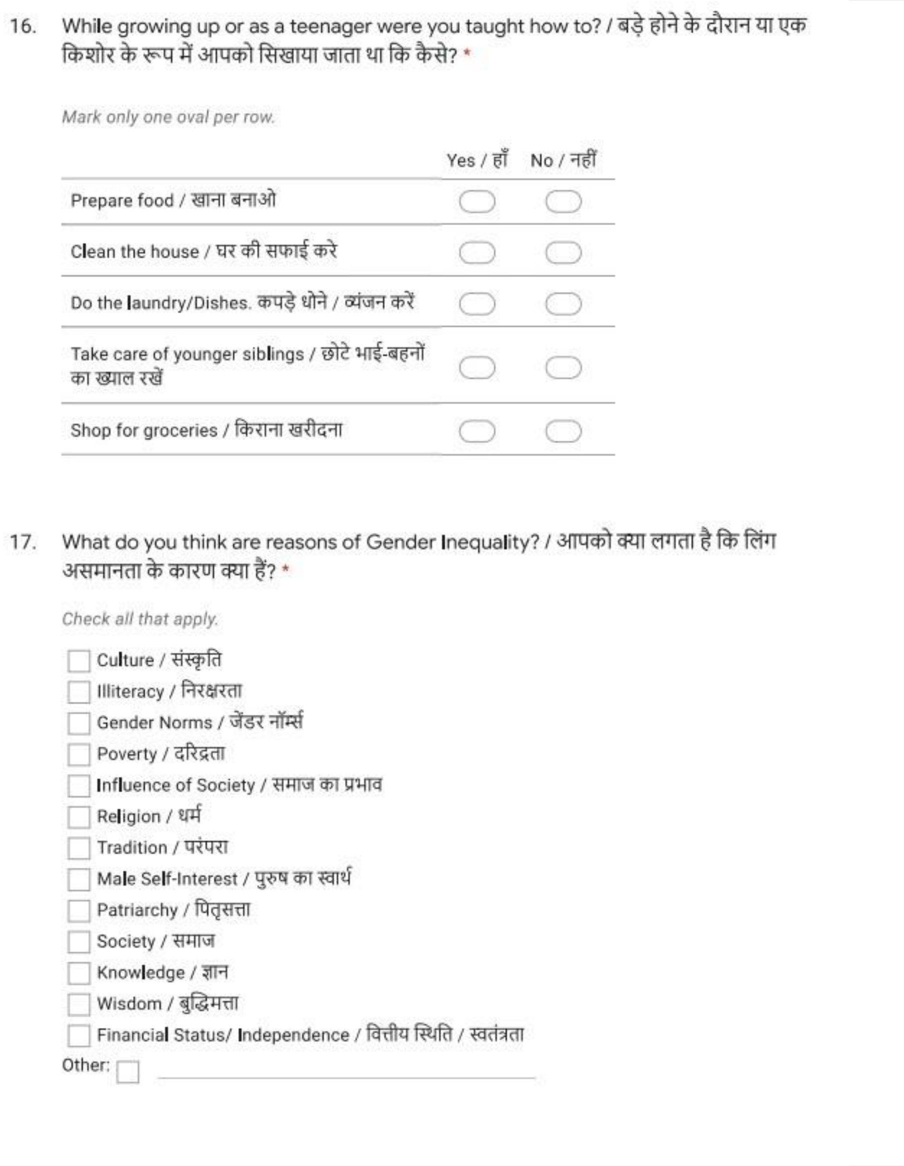
Mark only one oval.

○ Mother / मां

○ Father / पिता

○ Relative/ Grandparents. रिश्तेदार / दादा दादी

○ Guardian / अभिभावक



16. While growing up or as a teenager were you taught how to? / बड़े होने के दौरान या एक किशोर के रूप में आपको सिखाया जाता था कि कैसे? *

Mark only one oval per row.

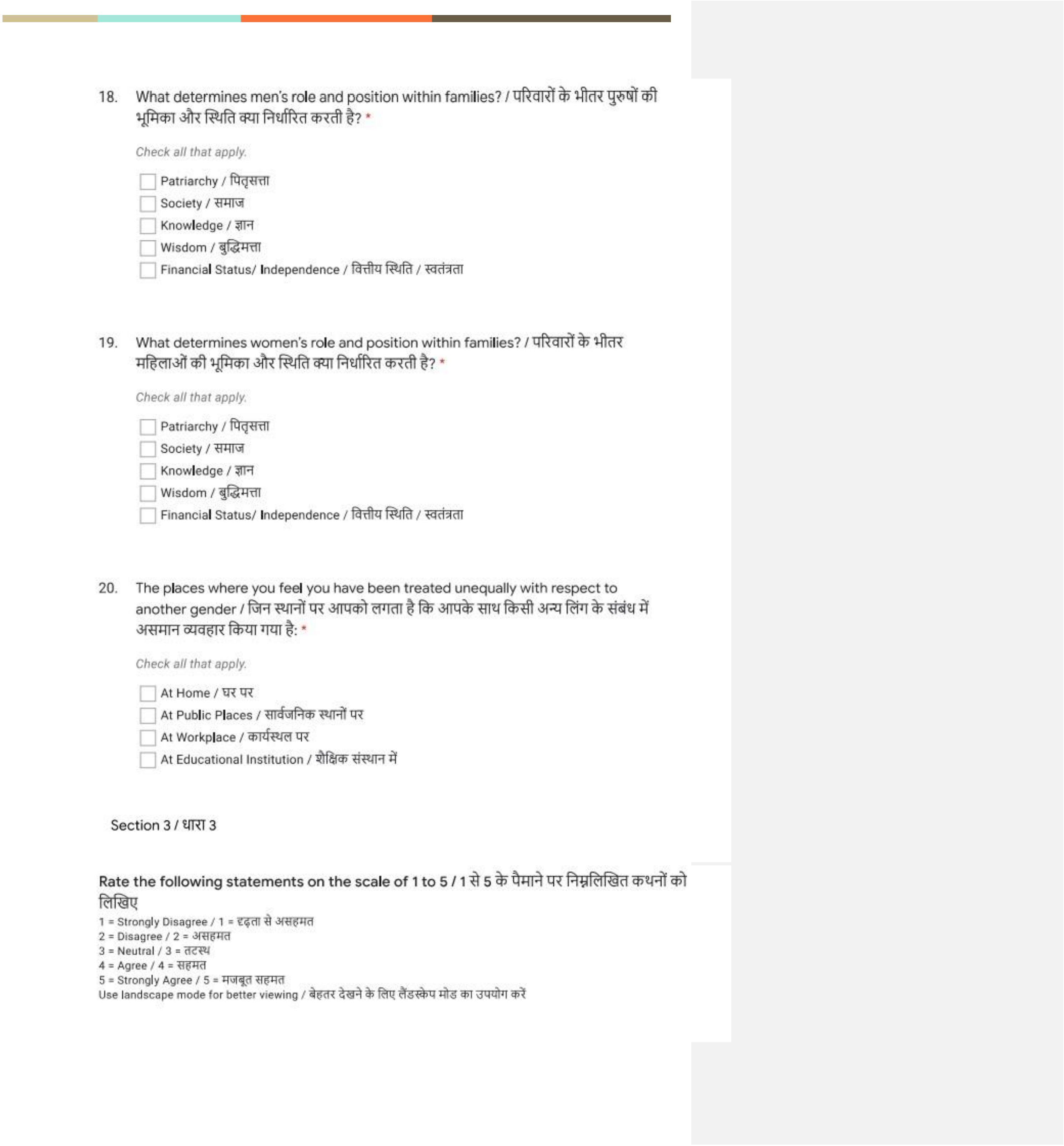
	Yes / हाँ	No / नहीं
Prepare food / खाना बनाओ	○	○
Clean the house / घर की सफाई करे	○	○
Do the laundry/Dishes. कपड़े धोने / व्यंजन करें	○	○
Take care of younger siblings / छोटे भाई-बहनों का ख्याल रखें	○	○
Shop for groceries / किराना खरीदना	○	○

17. What do you think are reasons of Gender Inequality? / आपको क्या लगता है कि लिंग असमानता के कारण क्या हैं? *

Check all that apply.

□ Culture / संस्कृति
□ Illiteracy / निरक्षरता
□ Gender Norms / जेंडर नॉर्म्स
□ Poverty / दरिद्रता
□ Influence of Society / समाज का प्रभाव
□ Religion / धर्म
□ Tradition / परंपरा
□ Male Self-Interest / पुरुष का स्वार्थ
□ Patriarchy / पितृसत्ता
□ Society / समाज
□ Knowledge / ज्ञान
□ Wisdom / बुद्धिमत्ता
□ Financial Status/ Independence / वित्तीय स्थिति / स्वतंत्रता

Other: □ ______________________



18. What determines men's role and position within families? / परिवारों के भीतर पुरुषों की भूमिका और स्थिति क्या निर्धारित करती है? *

Check all that apply.

□ Patriarchy / पितृसत्ता
□ Society / समाज
□ Knowledge / ज्ञान
□ Wisdom / बुद्धिमत्ता
□ Financial Status/ Independence / वित्तीय स्थिति / स्वतंत्रता

19. What determines women's role and position within families? / परिवारों के भीतर महिलाओं की भूमिका और स्थिति क्या निर्धारित करती है? *

Check all that apply.

□ Patriarchy / पितृसत्ता
□ Society / समाज
□ Knowledge / ज्ञान
□ Wisdom / बुद्धिमत्ता
□ Financial Status/ Independence / वित्तीय स्थिति / स्वतंत्रता

20. The places where you feel you have been treated unequally with respect to another gender / जिन स्थानों पर आपको लगता है कि आपके साथ किसी अन्य लिंग के संबंध में असमान व्यवहार किया गया है: *

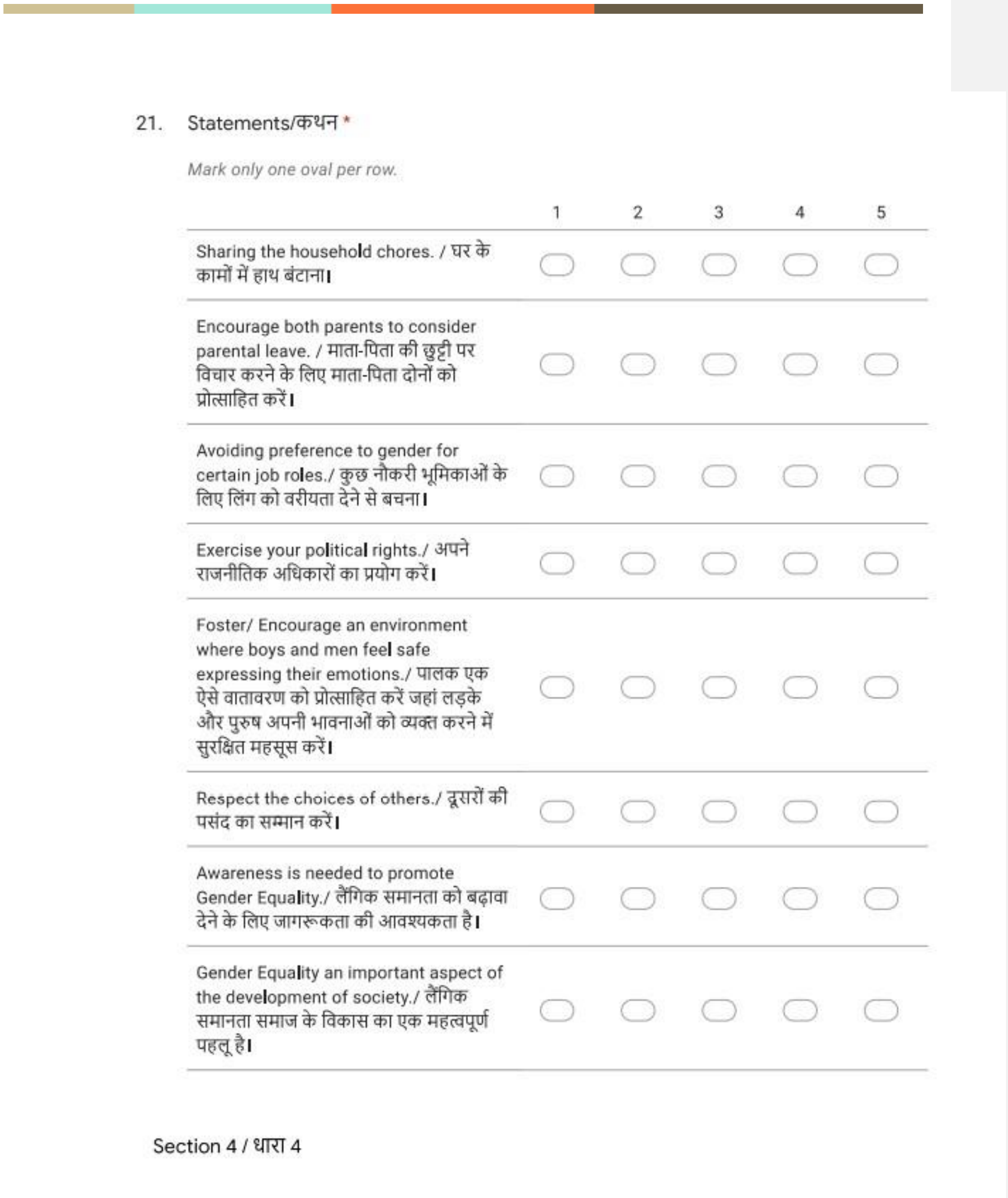
Check all that apply.

□ At Home / घर पर
□ At Public Places / सार्वजनिक स्थानों पर
□ At Workplace / कार्यस्थल पर
□ At Educational Institution / शैक्षिक संस्थान में

Section 3 / धारा 3

Rate the following statements on the scale of 1 to 5 / 1 से 5 के पैमाने पर निम्नलिखित कथनों को लिखिए

1 = Strongly Disagree / 1 = दृढ़ता से असहमत
2 = Disagree / 2 = असहमत
3 = Neutral / 3 = तटस्थ
4 = Agree / 4 = सहमत
5 = Strongly Agree / 5 = मजबूत सहमत
Use landscape mode for better viewing / बेहतर देखने के लिए लैंडस्केप मोड का उपयोग करें

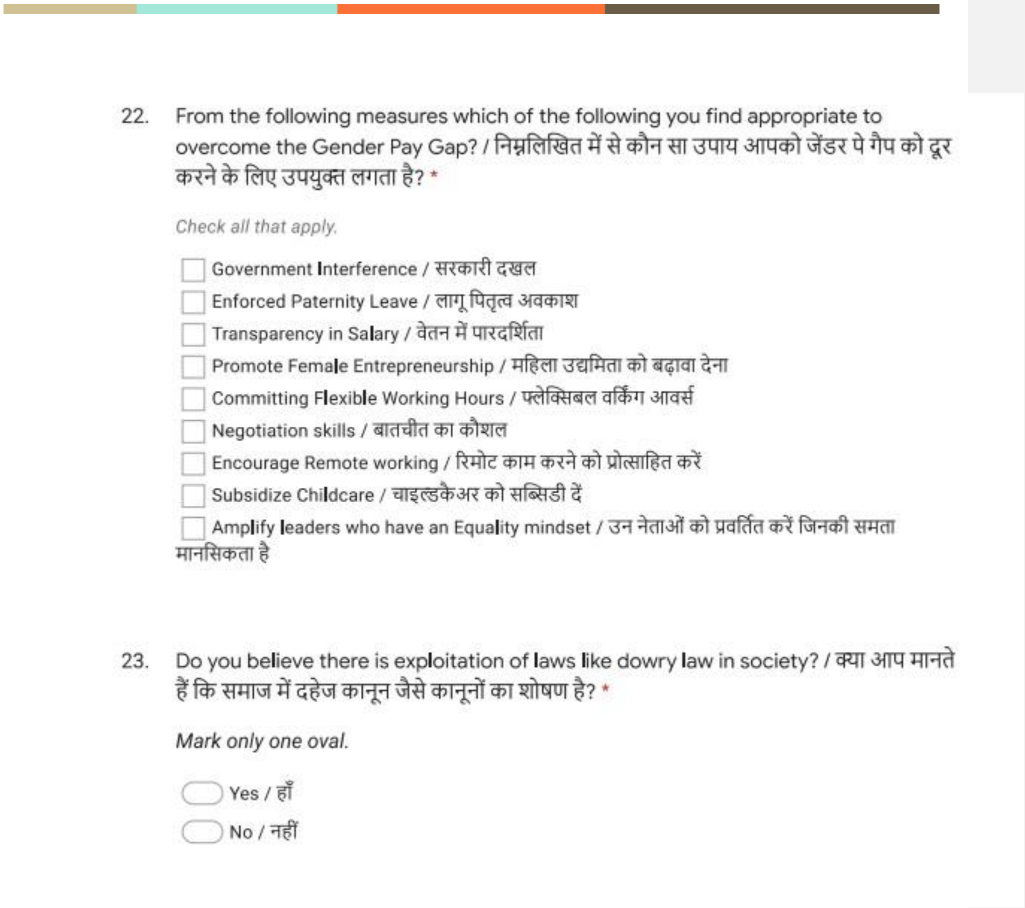


21. Statements/कथन *

Mark only one oval per row.

	1	2	3	4	5
Sharing the household chores. / घर के कामों में हाथ बंटाना।	○	○	○	○	○
Encourage both parents to consider parental leave. / माता-पिता की छुट्टी पर विचार करने के लिए माता-पिता दोनों को प्रोत्साहित करें।	○	○	○	○	○
Avoiding preference to gender for certain job roles./ कुछ नौकरी भूमिकाओं के लिए लिंग को वरीयता देने से बचना।	○	○	○	○	○
Exercise your political rights./ अपने राजनीतिक अधिकारों का प्रयोग करें।	○	○	○	○	○
Foster/ Encourage an environment where boys and men feel safe expressing their emotions./ पालक एक ऐसे वातावरण को प्रोत्साहित करें जहां लड़के और पुरुष अपनी भावनाओं को व्यक्त करने में सुरक्षित महसूस करें।	○	○	○	○	○
Respect the choices of others./ दूसरों की पसंद का सम्मान करें।	○	○	○	○	○
Awareness is needed to promote Gender Equality./ लैंगिक समानता को बढ़ावा देने के लिए जागरूकता की आवश्यकता है।	○	○	○	○	○
Gender Equality an important aspect of the development of society./ लैंगिक समानता समाज के विकास का एक महत्वपूर्ण पहलू है।	○	○	○	○	○

Section 4 / धारा 4



22. From the following measures which of the following you find appropriate to overcome the Gender Pay Gap? / निम्नलिखित में से कौन सा उपाय आपको जेंडर पे गैप को दूर करने के लिए उपयुक्त लगता है? *

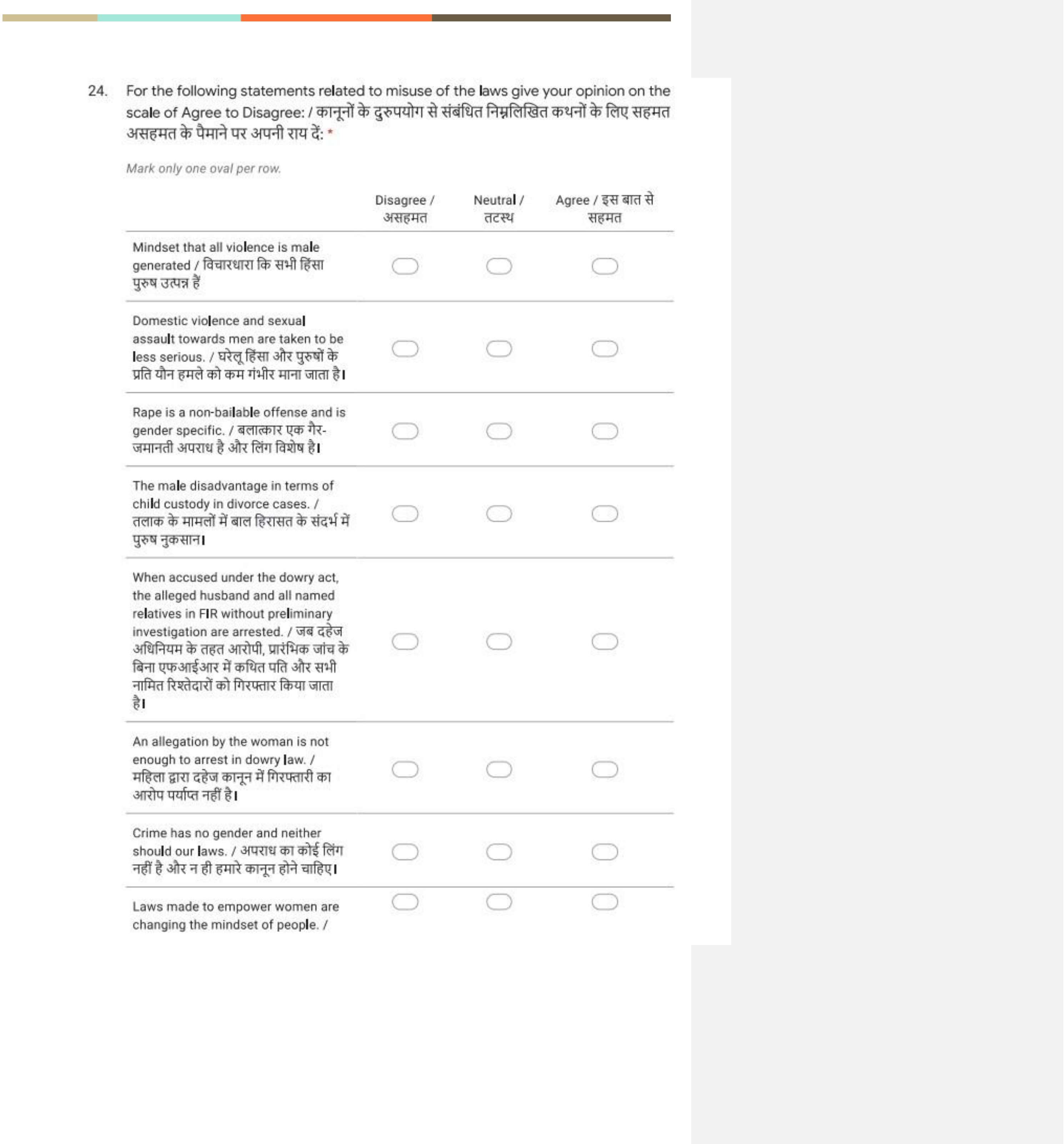
Check all that apply.

□ Government Interference / सरकारी दखल
□ Enforced Paternity Leave / लागू पितृत्व अवकाश
□ Transparency in Salary / वेतन में पारदर्शिता
□ Promote Female Entrepreneurship / महिला उद्यमिता को बढ़ावा देना
□ Committing Flexible Working Hours / फ्लेक्सिबल वर्किंग आवर्स
□ Negotiation skills / बातचीत का कौशल
□ Encourage Remote working / रिमोट काम करने को प्रोत्साहित करें
□ Subsidize Childcare / चाइल्डकैअर को सब्सिडी दें
□ Amplify leaders who have an Equality mindset / उन नेताओं को प्रवर्तित करें जिनकी समता मानसिकता है

23. Do you believe there is exploitation of laws like dowry law in society? / क्या आप मानते हैं कि समाज में दहेज कानून जैसे कानूनों का शोषण है? *

Mark only one oval.

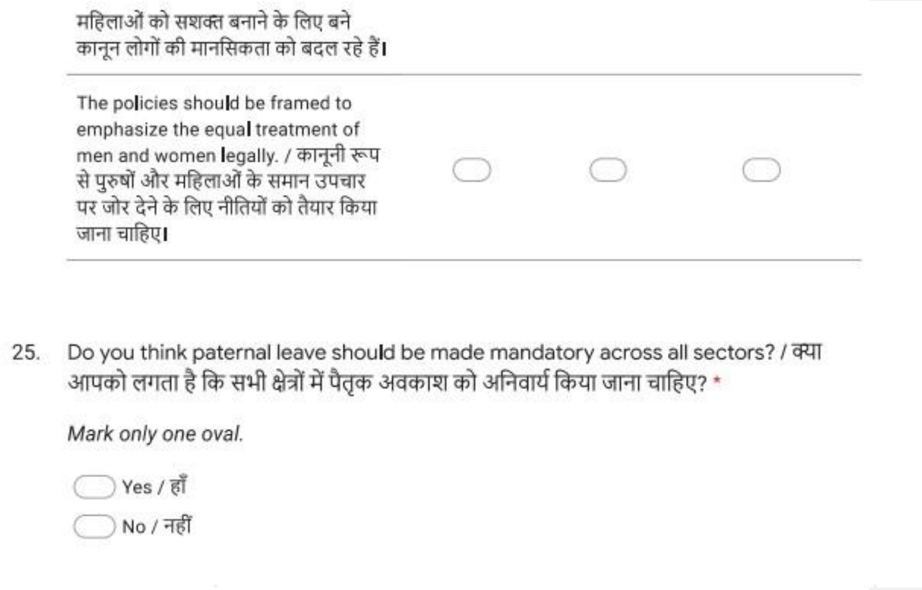
○ Yes / हाँ
○ No / नहीं



24. For the following statements related to misuse of the laws give your opinion on the scale of Agree to Disagree: / कानूनों के दुरुपयोग से संबंधित निम्नलिखित कथनों के लिए सहमत असहमत के पैमाने पर अपनी राय दें: *

Mark only one oval per row.

	Disagree / असहमत	Neutral / तटस्थ	Agree / इस बात से सहमत
Mindset that all violence is male generated / विचारधारा कि सभी हिंसा पुरुष उत्पन्न हैं	○	○	○
Domestic violence and sexual assault towards men are taken to be less serious. / घरेलू हिंसा और पुरुषों के प्रति यौन हमले को कम गंभीर माना जाता है।	○	○	○
Rape is a non-bailable offense and is gender specific. / बलात्कार एक गैर-जमानती अपराध है और लिंग विशेष है।	○	○	○
The male disadvantage in terms of child custody in divorce cases. / तलाक के मामलों में बाल हिरासत के संदर्भ में पुरुष नुकसान।	○	○	○
When accused under the dowry act, the alleged husband and all named relatives in FIR without preliminary investigation are arrested. / जब दहेज अधिनियम के तहत आरोपी, प्रारंभिक जांच के बिना एफआईआर में कथित पति और सभी नामित रिश्तेदारों को गिरफ्तार किया जाता है।	○	○	○
An allegation by the woman is not enough to arrest in dowry law. / महिला द्वारा दहेज कानून में गिरफ्तारी का आरोप पर्याप्त नहीं है।	○	○	○
Crime has no gender and neither should our laws. / अपराध का कोई लिंग नहीं है और न ही हमारे कानून होने चाहिए।	○	○	○
Laws made to empower women are changing the mindset of people. /	○	○	○



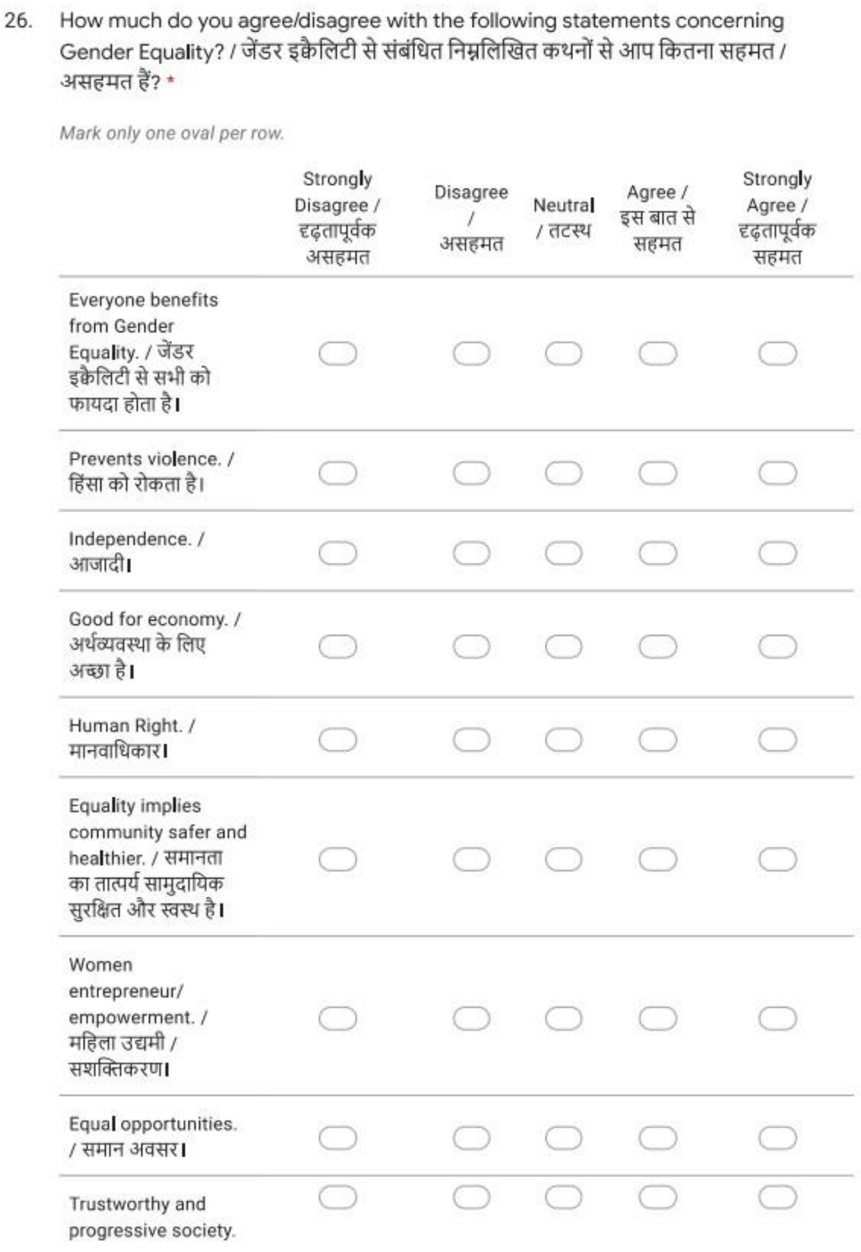
महिलाओं को सशक्त बनाने के लिए बने कानून लोगों की मानसिकता को बदल रहे हैं।			
The policies should be framed to emphasize the equal treatment of men and women legally. / कानूनी रूप से पुरुषों और महिलाओं के समान उपचार पर जोर देने के लिए नीतियों को तैयार किया जाना चाहिए।	○	○	○

25. Do you think paternal leave should be made mandatory across all sectors? / क्या आपको लगता है कि सभी क्षेत्रों में पैतृक अवकाश को अनिवार्य किया जाना चाहिए? *

Mark only one oval.

○ Yes / हाँ

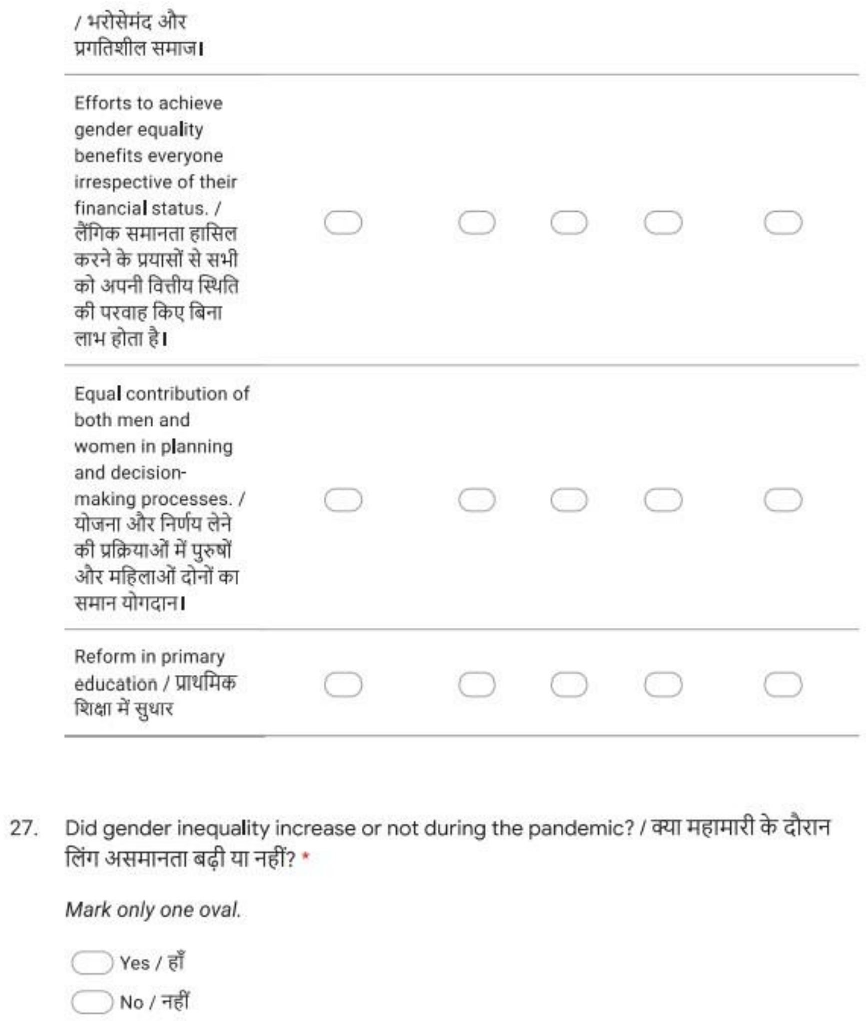
○ No / नहीं



26. How much do you agree/disagree with the following statements concerning Gender Equality? / जेंडर इक्वैलिटी से संबंधित निम्नलिखित कथनों से आप कितना सहमत / असहमत हैं? *

Mark only one oval per row.

	Strongly Disagree / दृढ़तापूर्वक असहमत	Disagree / असहमत	Neutral / तटस्थ	Agree / इस बात से सहमत	Strongly Agree / दृढ़तापूर्वक सहमत
Everyone benefits from Gender Equality. / जेंडर इक्वैलिटी से सभी को फायदा होता है।	○	○	○	○	○
Prevents violence. / हिंसा को रोकता है।	○	○	○	○	○
Independence. / आजादी।	○	○	○	○	○
Good for economy. / अर्थव्यवस्था के लिए अच्छा है।	○	○	○	○	○
Human Right. / मानवाधिकार।	○	○	○	○	○
Equality implies community safer and healthier. / समानता का तात्पर्य सामुदायिक सुरक्षित और स्वस्थ है।	○	○	○	○	○
Women entrepreneur/ empowerment. / महिला उद्यमी / सशक्तिकरण।	○	○	○	○	○
Equal opportunities. / समान अवसर।	○	○	○	○	○
Trustworthy and progressive society.	○	○	○	○	○



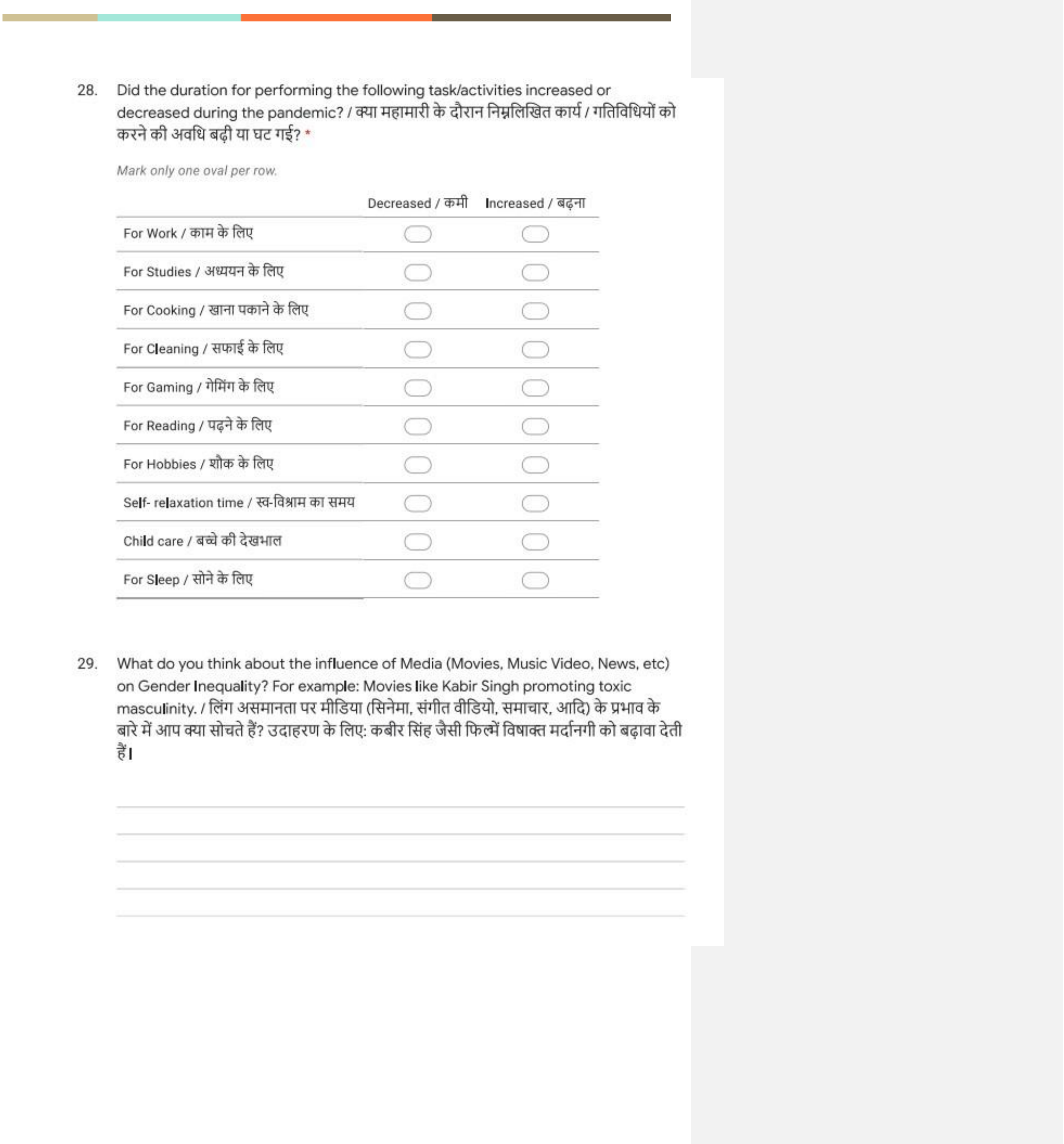
/ भरोसेमंद और प्रगतिशील समाज।					
Efforts to achieve gender equality benefits everyone irrespective of their financial status. / लैंगिक समानता हासिल करने के प्रयासों से सभी को अपनी वित्तीय स्थिति की परवाह किए बिना लाभ होता है।	○	○	○	○	○
Equal contribution of both men and women in planning and decision-making processes. / योजना और निर्णय लेने की प्रक्रियाओं में पुरुषों और महिलाओं दोनों का समान योगदान।	○	○	○	○	○
Reform in primary education / प्राथमिक शिक्षा में सुधार	○	○	○	○	○

27. Did gender inequality increase or not during the pandemic? / क्या महामारी के दौरान लिंग असमानता बढ़ी या नहीं? *

Mark only one oval.

○ Yes / हाँ

○ No / नहीं



28. Did the duration for performing the following task/activities increased or decreased during the pandemic? / क्या महामारी के दौरान निम्नलिखित कार्य / गतिविधियों को करने की अवधि बढ़ी या घट गई? *

Mark only one oval per row.

	Decreased / कमी	Increased / बढ़ना
For Work / काम के लिए	○	○
For Studies / अध्ययन के लिए	○	○
For Cooking / खाना पकाने के लिए	○	○
For Cleaning / सफाई के लिए	○	○
For Gaming / गेमिंग के लिए	○	○
For Reading / पढ़ने के लिए	○	○
For Hobbies / शौक के लिए	○	○
Self- relaxation time / स्व-विश्राम का समय	○	○
Child care / बच्चे की देखभाल	○	○
For Sleep / सोने के लिए	○	○

29. What do you think about the influence of Media (Movies, Music Video, News, etc) on Gender Inequality? For example: Movies like Kabir Singh promoting toxic masculinity. / लिंग असमानता पर मीडिया (सिनेमा, संगीत वीडियो, समाचार, आदि) के प्रभाव के बारे में आप क्या सोचते हैं? उदाहरण के लिए: कबीर सिंह जैसी फिल्में विषाक्त मर्दानगी को बढ़ावा देती हैं।

BIBLIOGRAPHY

References for research on the topic

- https://en.wikipedia.org/wiki/Gender_inequality_in_India#/media/File:India's_Global_Rank_on_Selected_Gender_Indequality_Indices.jpg
- https://www.weforum.org/agenda/2019/02/india-s-inequality-crisis-hurts-girls-and-women-the-most
- https://www.statista.com/statistics/983020/female-labor-force-participation-rate-india/
- https://www.drishtiias.com/daily-updates/daily-news-analysis/global-gender-gap-index-2020-wef
- The difference between men and women: How we view gender equality | Ipsos

For Factor Analysis;

- https://www.spss-tutorials.com/spss-factor-analysis-tutorial/
- https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/

For K-Modes;

- va4newphysics.wordpress.com/2016/10/26/into-the-world-of-clustering-algorithms-k-means-k-modes-and-k-prototypes/comment-page-1/
- http://www.legalserviceindia.com/legal/article-3095-how-women-misuse-their-rights.html
- https://timesofindia.indiatimes.com/readersblog/riyable/the-gender-advantage-women-who-misuse-it-men-who-bears-it-5475/

For Decision Trees;

- https://www.geeksforgeeks.org/decision-tree/
- https://www.geeksforgeeks.org/decision-tree-implementation-python/

For A-priori;

- https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html