Prediction and Analysis of Crime against Women In India using Machine Learning Algorithms

# Ritika ,Vidushi Awasthi

Department of ComputerScience and Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India

**Abstract.**In India crimes against women are increasing at a significant rate as noticed in some recent years, and have reached unprecedented levels. By examining the data provided by NCRB we can notice there isa significant rise in crimes in 2022 as compared to previous years. This is a critical issue, which needs to be analyzed with relevant information regarding all types of crimes against women. For this purpose, one can take help of Artificial Intelligence and machine learning techniques that can be termed as the most important tool for analyzing the crime prone areas, conducting various studies on different crime types.

This research paper focuses solely on the crimes committed against women across different states of India spanning from 2017-2022 , covering various crimes like Rape, Kidnapping or abduction of women, Insulting modesty, dowry deaths, cybercrime against women and more state wise. Moreover, this paper also includes yearly analysis of persons arrested, convicted, discharged for that particular crime. A deep analysis has been conducted to gain deeper insights into the data and machine learning algorithms. Through this analysis we identified the most crime prone states in India, most frequently committed  crime across the state and many more aspects.

Our findings through this study not only understands the historic crime trends but also forecast the future crime incidents, which would be very helpful in taking preventive measures, law enforcement and policy making. This research highlights the integration of Machine Learning into the justice system to enhance its effectiveness in protecting vulnerable populations.

**Keywords:**Crime against Women, person arrested, convictions, machine learning techniques, random forest regression, XG boost.

1. **Introduction**

The number of cases filed each year in India for crime against women have been increasing each year. The accurate prediction for such crimes is not efficient as several cases go unreported because many women are not aware of their rights and the official reporting systems(Krishnan, 2017; Jhamb, 2011). Some social norms such as patriarchy, religious and cultural beliefs about marriage, and asymmetrical gender expectations(Koenig et al., 2006; Chaudhary, 2013; Chachra, 2017) are the reasons women feel uncomfortable disclosing their experience to the members of the community(Krishnan, 2017).

Data trends show that the recent increase in Cyber Crime also points toward the lack of awareness and education of the citizens regarding online safety despite the internet being an integral part of our lives. The availability of cheap smartphones and expanding internet connectivity has made it easier for perpetrators to exploit vulnerable individuals, particularly those who lack basic knowledge of digital security([**https://www.ijnrd.org/papers/IJNRD2312232.pdf**](https://www.ijnrd.org/papers/IJNRD2312232.pdf)).

According to the NCRB dataset, crimes against women, for example, Dowry deaths, Rape or Gang Rape, Acid Attacks, Domestic Violence, Child Abuse have been increasing every day. Hence, keeping that in view, the establishment of the gender equality principle was done in the Constitution of India, and to uphold and implement the Constitutional mandate, the state has created several laws as well as has taken various actions to guarantee equal rights, eradicate social injustice, and prohibit multiple forms of violence and massacres. (Hagan, Frank E. Crime types and criminals. Sage, 2009).

National Crime Records Bureau data shows a 4% rise in the number of crimes against women in 2022 over 2021, and nearly a 20% increase compared to the number of gender crimes reported in 2020. The data also shows that 4,45,256 cases were reported in 2022 alone. According to the NFHS-5 estimations, the proportion of women who experienced spousal violence has increased in 5 states. In Karnataka, it has risen from 21% to 44%. Women also lag behind men in literacy rate by 71.5% against 84.4% of men, which points towards high female dropout rates. From 2017-2022, Uttar Pradesh recorded the highest number of cases for crime against women. But when compared to the population, Assam stood on the top of the list.

The crime investigation framework should be fast and efficient in identifying patterns in criminal activity in order to predict and anticipate future trends, enabling swift addressing of the matter. In order to tackle such crimes, law enforcement agencies and state governments need to implement proactive measures to ensure a safe and secure environment for women. Researchers all around the world have been working to detect crimes against women, which is why the role of data analytics and Machine Learning models have become increasingly important in this effort. Such technologies provide valuable insights for criminal analysis, making them an emerging technique for research related Crime. To uphold safety, law enforcement agencies are turning towards AI and Data Mining techniques.Data Mining is the process of extracting knowledge or insights from large amounts of data using various statistical and computational techniques.[[Data Mining Tutorial - GeeksforGeeks](https://www.geeksforgeeks.org/data-mining/)] It is a key component of database knowledge discovery and helps to transform raw data into actionable information which can further be used for making accurate predictions and applying them to real world scenarios through analysis and inferences. Various scientific and statistical Machine Learning algorithms are being utilized for image and speech recognition and also for detecting medical conditions and classifying data. These algorithms can also be used for predicting crime rates by analyzing statistical data related to crime against women.

1. **Proposed System**

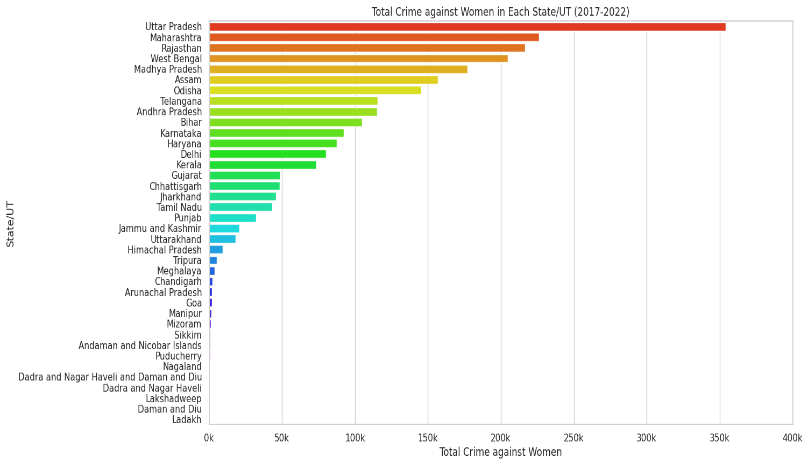
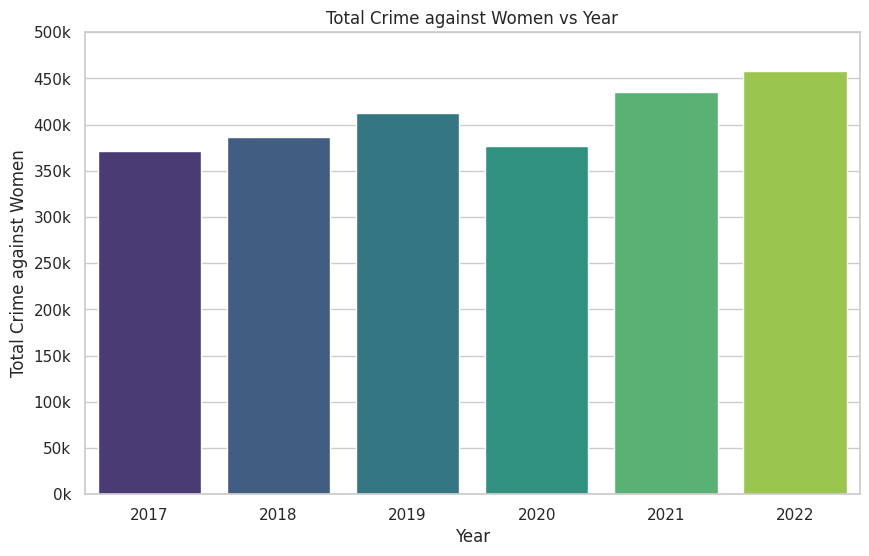
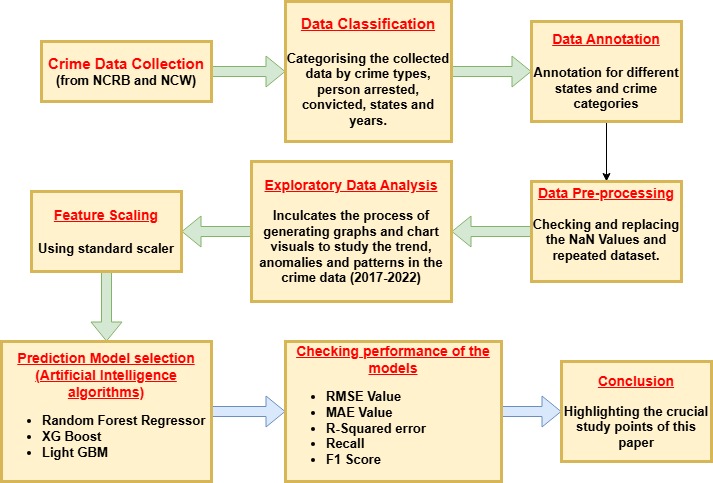


Fig.1Total crime committed against women year wise Fig.2 State wise crime committed against women (2017- 2022)

To investigate data about crimes against women in India, the Crime Data Provenance Model was created. By the support of annotations—subtext or information that offers more meaning to the original data—this framework aims to present an approach to facilitate the use of the lineage of crime data in order to identify distinct crime types. The crime data is classified and predicted using these crime types. Prediction models are implemented in accordance with the basic aspects identified by the framework, which are different crime sub-categories. After then, the outcomes are contrasted and examined. Interestingly, the framework has only been applied to one specific type of crime—crimes against women. The process of the framework is broken down into a number of subsections, each of which describes the particular goals and duties carried out at each level.



* 1. **Data Extraction**

The data for this paper was sourced from the National Crime Records Bureau (NCRB) website and the National Commission for Women (NCW). The NCRB was established on March 11, 1986, and the NCW was established on January 31, 1992. The primary goal of the NCRB website is to enhance the capabilities and bolster Indian police operations through information technology, thereby improving the effectiveness of law enforcement. The objective of the NCW is to raise awareness about women's rights in India and advocate for their issues and concerns. Our data spans from 2017 to 2022 and was available year-wise on these mentioned websites. Some books and research papers helped in gaining a deeper understanding of crimes, their trends, and other aspects. The report *Violence against Women in India*, created by the International Center for Research on Women, aimed to provide a detailed analysis of the trends and patterns in crimes against women.

This research paper focuses on three different datasets: ‘State/UT Wise Crime against Women from 2017-2022,’ ‘Crime Head Wise Disposal of Persons Arrested for Crimes against Women from 2017-2022,’ and ‘State/UT Wise Disposal of Persons Arrested for Crimes against Women from 2017-2022.’ At first glance, the datasets seemed complex, but they were later simplified to a great extent.

The dataset has been segregated on the basis of States and Union Territories. As we know India has 28 states and 9 Union Territories and so the data has been aggregated based on the number of cases reported for mentioned crime categories in these states or UTs.

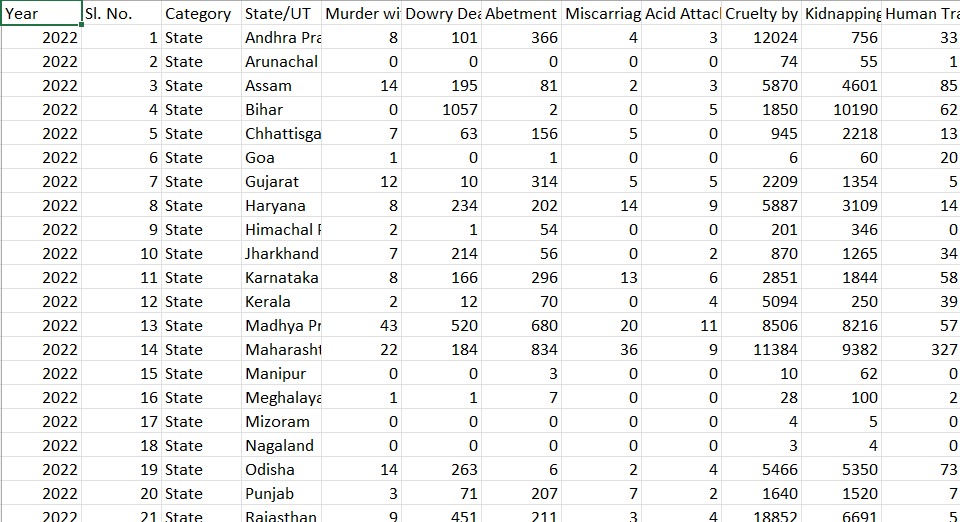
Now some of the important columns from the dataset ‘State/UT wise crime against women from 2017-2022’:

* State/UT
* Murder with Rape/Gang Rape
* Dowry Deaths
* Acid Attack
* Cruelty by Husband or His Relatives
* Kidnapping and Abduction of Women
* Rape Total
* Attempt to Commit Rape
* Assault on Women with Intent to Outrage Her Modesty
* Dowry Prohibition Act (1961)
* Cyber Crimes/Information Technology Act (Women Centric Crimes Only)
* Protection of Children from Sexual Offences Act (Girl Child Victims Only)
* Total Crime Against Women

 Here are some important columns from the dataset ‘State/UT Wise Disposal of Persons Arrested for Crimes Against Women from 2017-2022.’:

* State/UT
* Persons Arrested - Total
* Persons Charge Sheeted - Total
* Persons Convicted - Total
* Persons Discharged - Total
* Persons Acquitted - Total

This detailed classification facilitates a comprehensive analysis of crime trends and patterns.

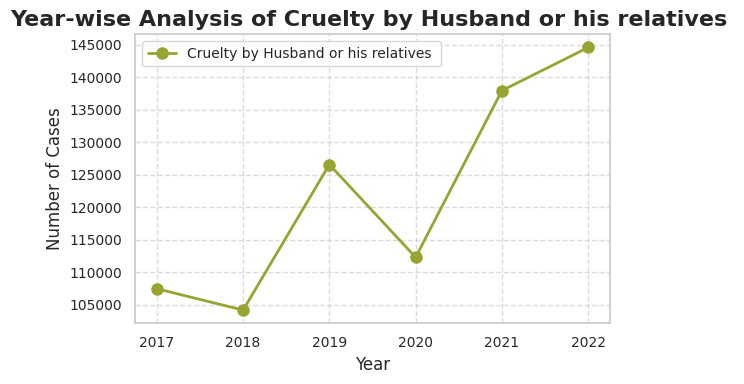
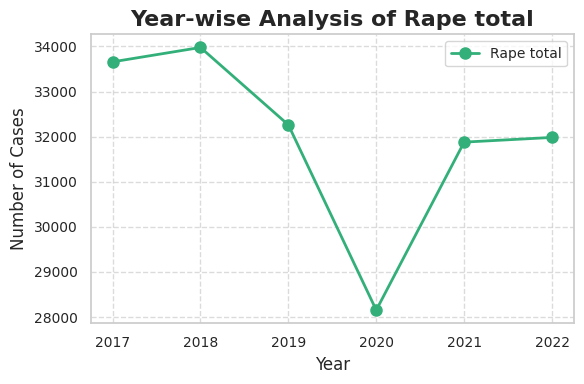


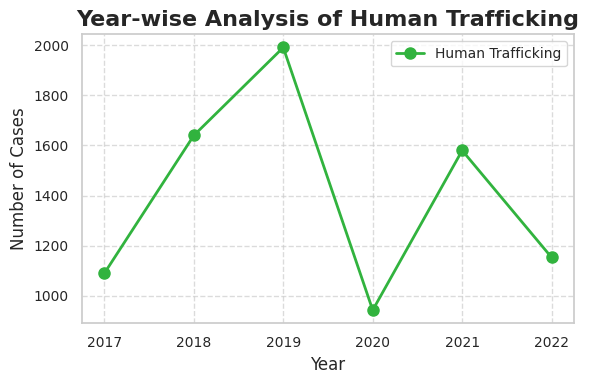
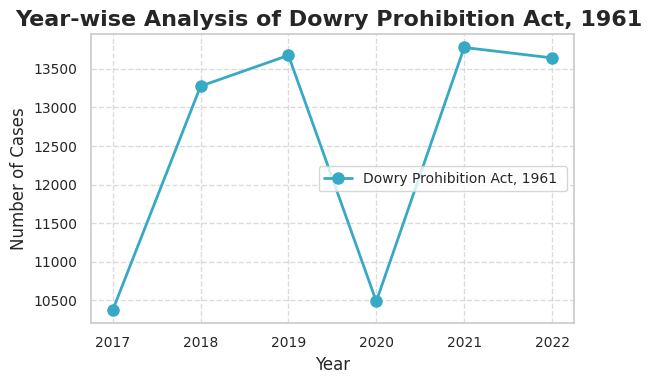
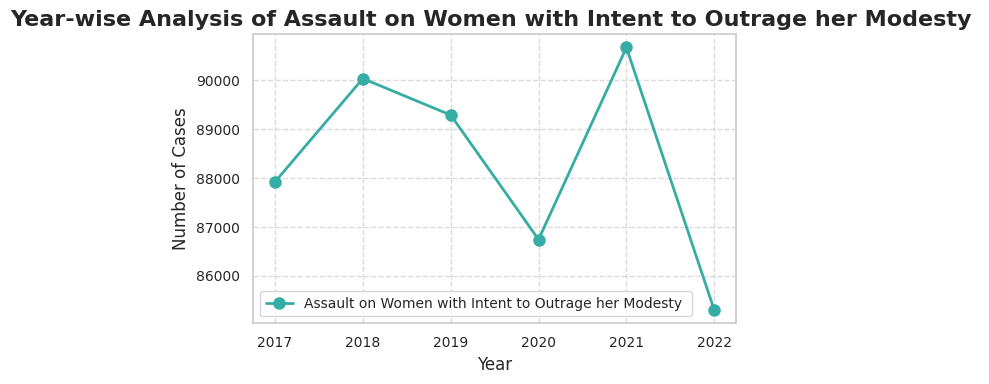
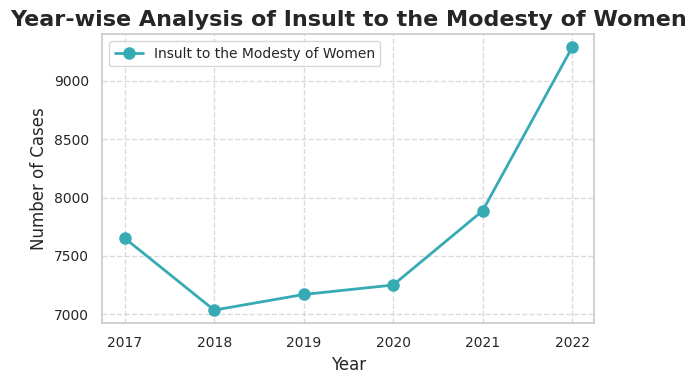
* 1. **Data Preprocessing**

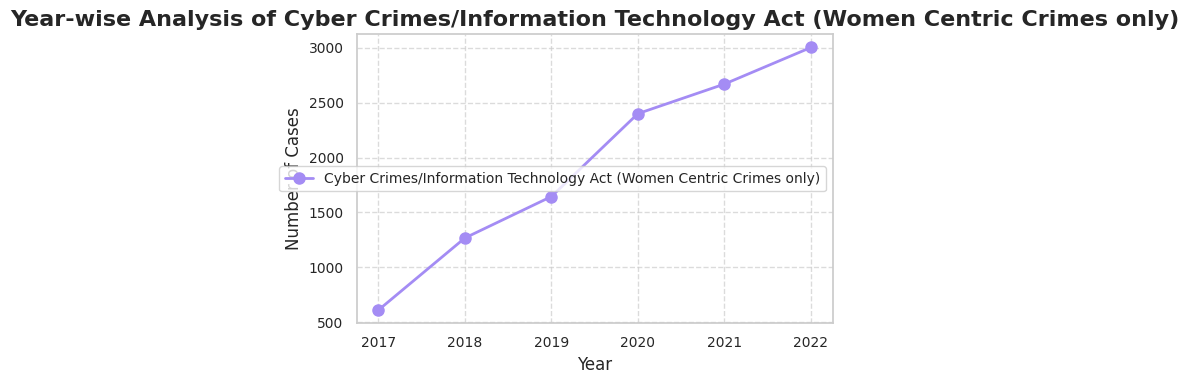
Data preprocessing becomes a very crucial step in cleaning and preparing the dataset for the use of machine learning models and thus escalating the accuracy of the training models. The first step in preprocessing is to set a platform like ‘Jupyter notebook’ or ‘Google collab’ and then the installation of different libraries of python like Matplotlib, NumPy, scikit-learn, itertools, SimpleImputer, Seaborn, Math, and Pandas which would be helpful in carrying out different types of operation on datasets like handling of missing values, identifying categorical features , dealing with inconsistencies and many more.[ <https://www.researchgate.net/publication/357810730_A_comparative_study_of_crimes_against_women_based_on_Machine_Learning_using_Big_Data_techniques> ]

[ <https://www.researchgate.net/publication/353414128_Exploratory_Data_Analysis_And_Crime_Prevention_Using_Machine_Learning_The_case_of_Ghana> ]

1. **Exploratory Data Analysis**
   1. **Crime trend year wise**

****

****

****

Graphs illustrating different major crimes against women from 2017-2022.

In this we have generated various graphs and visuals showcasing the year -wise frequency of particular crimes (in number) which will help in analyzing the anomalies and pattern of crimes against women over the year 2017-2022 offering the valuable insights into temporal dynamics of women safety.

Additionally some more graphs were plotted to analyze the % increase in crimes. Here one gets to see the year on year comparison of the crimes over the specified period. It actually reveals the severity, falling and rising of a particular crime. Clear and detailed analysis of different categories of crimes is very necessary to take preventive measures.

By plotting these graphs we observe some fluctuation and interesting patterns:

1. If we would focus on graph illustrating 'Rape Total' reveals a marked decline of around -12% in rape cases in 2020 as compared to 2019. This major change could be attributed to the pandemic COVID-19 when nationalwide lockdown was imposed all over India. Citizens were confined in their homes and thus the possibility for the rape cases reduced hence decline in reported cases during the period.
2. Now the graph illustrating the 'Insult to the modesty of women'  shows a significant rise but contrary , the graph 'Assault on women to outrage her modesty' shows a deep decline.Insult to modesty includes non-physical harassment but Assault on women involves physical abuse and harassment.

             The reasons could be:

* Social stigma and fear: As the cases of physical assaults are often more traumatic and are underreported due to fear, social stigma, the fear of being judged or blamed by our society. This stigma leads to impacting their personal lives. Sometimes women fear retaliation if the perpetrator is a family member or someone with social power. They feel that reporting such crimes could endanger their loved ones.
* Legal complexities:  If women report physical assault she has to navigate a complex legal system.This process involves multiple interactions with law enforcement, court appearances which could be time-consuming and re-traumatizing.
* Reporting Non-physical abuse and verbal insults:  verbal insults often occur in public places or digital platforms, and hence there are great chances of witnesses and evidence in the form of video recordings. This creates high chances for the perpetrator for being held accountable.
* In such cases there is less invasive investigation and procedures and a higher level of support for the victim.

3. The significant rise in cases of 'Kidnapping and Abduction of Women' and violations of the 'Protection of Children from Sexual Offences (POCSO) Act' highlights different reasons:

* Increased vigilance : there is improvement in reporting and vigilance mechanisms of such crimes against women and children as authorities have become more focused and proactive in registering these types of cases and thus reflecting an increase in the number of cases.
* Increased exposure to social media : on social media they usually get connected with many strangers and usually share real time or geotagging locations, personal activities and travel plans, which gives predators a chance to gain access to potential victims. Trends on social media encourage young girls engaging in meeting strangers and increasing vulnerability to abduction or sexual abuse.

4. There is a steep increase in cybercrime, particularly the crimes which are targeting women, from 2017-2022. This could be attributed to several factors:

* Growth of digital platforms and social media:  with more easy access to the internet, digital platforms, smartphones and social media, the number of potential targets for cybercrime has increased a lot. More use of these platforms leads to rise in cyberstalking, online harassment and cyberbullying.
* Lack of digital literacy: Many internet users from rural or semi-urban areas lack awareness about online security and privacy. This gives power to cybercriminals to exploit the potential targets and commit crimes like fraud, identity theft, online abuse, harassing and blackmailing women victims.
* Moreover COVID-19 pandemic led to a massive shift to digital platforms and that has increased the number of cybercriminals as well.

In 2020 total crime rate against women shows a significant decline because of lockdowns and restricted movement of people during the pandemic. But cybercrime continues to rise and thus reflecting a shift to digital offenses and vulnerability of women to cyber threats.

**Year wise analysis of crimes in states**

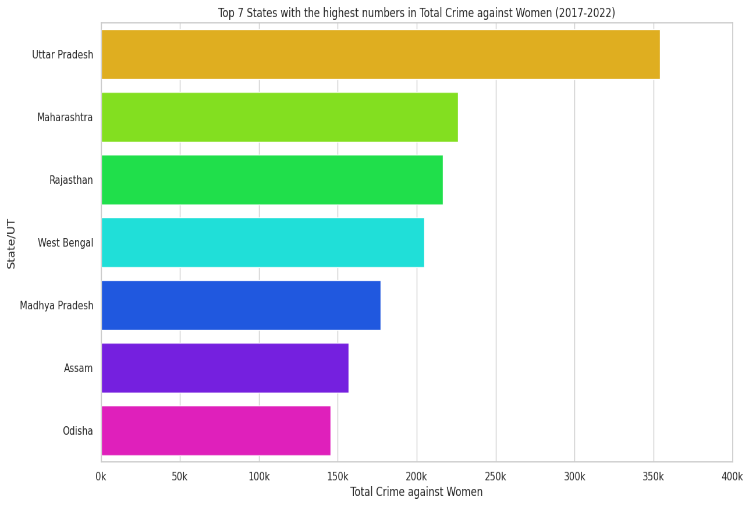
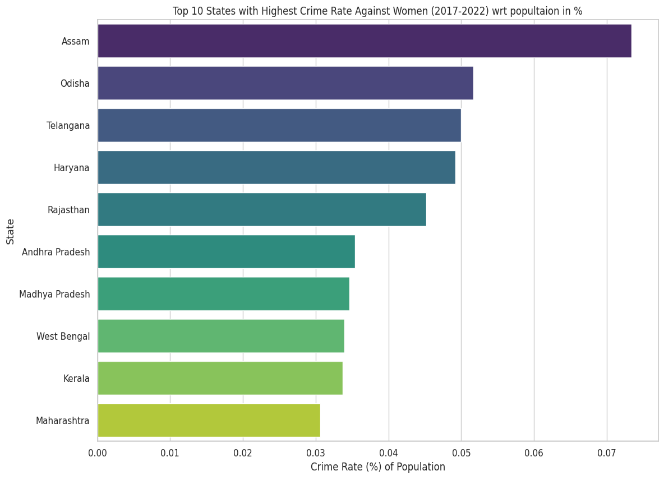
1. *2017-2019: Pre-Lockdown Period*
   * In terms of crime against women, Uttar Pradesh, Maharashtra, Rajasthan, and West Bengal are the states which have been consistently ranked at the top. There was a steady increase in the number of reported cases across these states in this period, which is an indication that either there was an actual rise in the number of incidents or the reporting mechanisms have improved over the years.
   * States like Madhya Pradesh and Odisha also show a similar trend of increasing incidents, however the rate is slower compared to the top states.
   * Due to smaller populations, smaller states and union territories like Lakshadweep, Mizoram, and Dadra and Nagar Haveli have consistently reported the lowest numbers.
2. *2020: COVID-19 Lockdown Impact*
   * The lockdown implemented in 2020 in India is the major reason for the initial dip in the number of cases observed in the graph. The main reason being lockdown measures and restricted movement.
   * Even though there was a general decrease in the number cases, still states like Uttar Pradesh and Maharashtra reported higher numbers compared to the others, although there was a significant drop compared to previous years.
   * Another very important point to be notes is that lockdown might also have led to underreporting due to limited access to support systems and law enforcement agencies.
3. *2021-2022: Post-Lockdown Recovery*
   * In 2021, when lockdown restrictions eased, the crime rates rose again. States like Uttar Pradesh, Maharashtra, Rajasthan, and West Bengal regained or even surpassed their pre-2020 numbers. The rebound not only indicates the return to normal reporting systems but also actual increase in crime rates due to various factors like social and economic pressure in the post-lockdown period.
   * In 2022, the continuous increase in the crime numbers across several states is clearly evident. Uttar Pradesh still remained the stated with the most reported cases, followed by Maharashtra and Rajasthan. However, this may be due to high populations.
   * The growth of crimes in 2021 and 2020 suggests that issues faced by women have intensified over the years due to socio-economic impacts of the pandemic, which includes loss of jobs or income, increased domestic tensions, an reduced social mobility.
   * The number of reports across smaller states/UTs like Lakshadweep, Mizoram, and Dadra and Nagar Haveli have continued to remain low but the consistent growth pattern aligns to that of larger states in terms of trends.

**Considerations:**

* The year 2020 stands out in the graphical representation as an anomaly due to the pandemic and the subsequent lockdown imposition with likely underreporting across the states.
* The following years (2021 and 2022), clearly show a rebound and even growth in the number of reported cases, indicating towards the increased socio-economic burdens and other lingering effects of the pandemic and the lockdown collectively.

This analysis of the number of reported cases across the states gives us an overview of the trends of crime rates over the years 2017 to 2022, with finding particular attention on the anomaly seen in 2020 due to the lockdown and the pandemic and the reasons for growth of crime in the subsequent years 2021 and 2022.

**3.2 Analysis of Crime Rates Relative to Population in Indian States (2017-2022)**

****

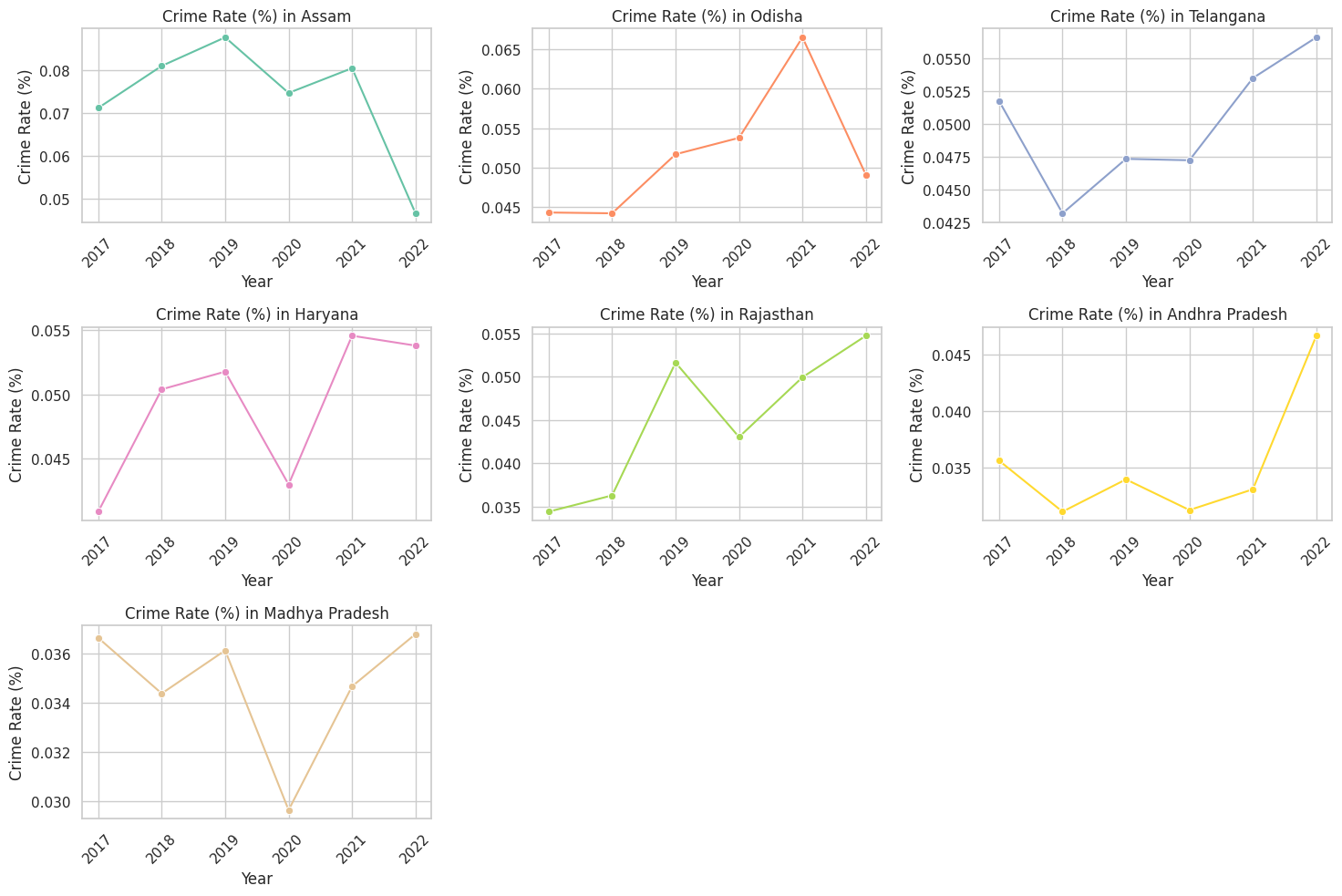
During our analysis of the number of reported cases across the states and UTs, states like Uttar Pradesh, Maharashtra, Rajasthan, and West Bengal consistently stayed on the top of the graph. However, a major reason for this was the high population of these states. So, in order to find the states with the highest crime rate in correspondence to their population, we divided the number of cases reported by the population of that state in that particular year. The state names that now emerged were not expected. States like Assam, Odisha, Telangana, Haryana, Rajasthan, Andhra Pradesh, Madhya Pradesh, West Bengal, Kerala, and Maharashtra have the highest crime rates per capita.

**Unexpected Findings.**

These state names are typically not associated with high crime against women rates which makes this particular finding very significant in our analysis. For example, despite the small population of Assam, it still tops the list, which reflects a higher crime rate in Assam as compared to its population size.

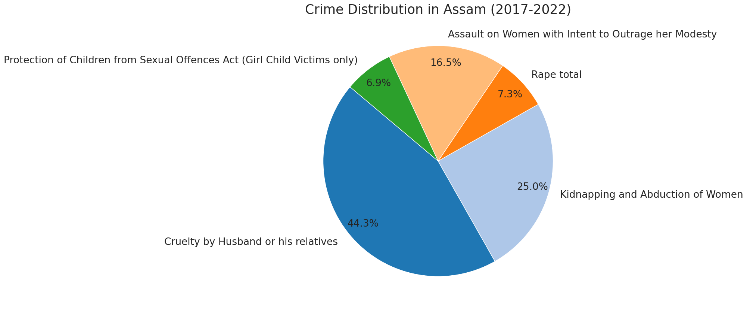
**Importance of This Analysis**

This analysis challenges the common perceptions and highlights the states with the actual highest crime rates making them the states where women’s security is the most at risk. The unexpected high criminal rates signify that there is an urgent need for further investigation into the cultural and socio- economic factors at play, making this analysis crucial for developing targeted policies and interventions

**Top 7 States with Highest Crime Rates:**

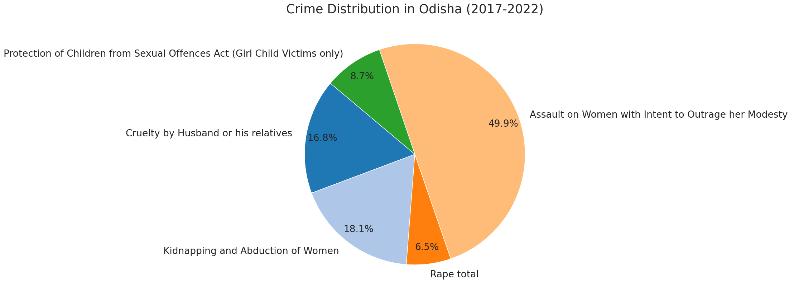
1. *Assam***:** The crime rate in Assam saw a slight increase from 2017 to 2019, followed by a drop in 2020 and a subsequent rise in 2021 and 2022.
2. *Odisha***:** Odisha witnessed a fluctuating trend with an initial increase from 2017 to 2019, a dip in 2020, and a rise in 2021, culminating in the highest crime rate in 2022.
3. *Telangana***:** Telangana experienced a relatively stable crime rate from 2017 to 2020, followed by a slight increase in 2021 and a drop in 2022.
4. *Haryana***:** Haryana's crime rate remained relatively consistent throughout the period, with minor fluctuations between 2017 and 2022.
5. *Rajasthan***:** Rajasthan saw a fluctuating trend with an initial decrease from 2017 to 2019, a rise in 2020, followed by a decline in 2021 and an increase in 2022.
6. *Andhra Pradesh***:** Andhra Pradesh witnessed a significant drop in crime rates from 2017 to 2020, followed by a slight increase in 2021 and a further drop in 2022.
7. *Madhya Pradesh:* Madhya Pradesh experienced a general decline in crime rates from 2017 to 2022, with minor fluctuations throughout the period.

**Analysis of Crime Distribution in Assam (2017-2022)**



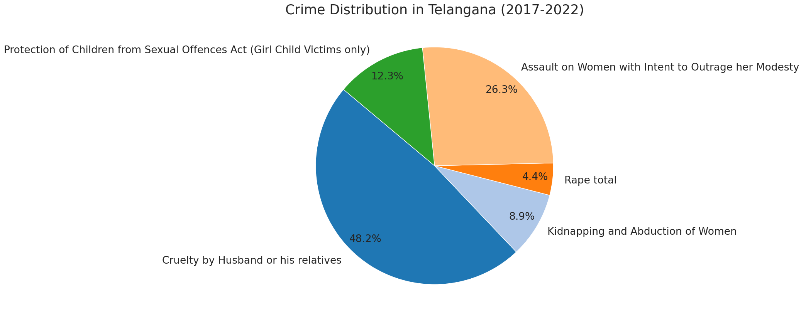
Above graph showing crime data analysis in Assam reveals a concerning situation of domestic violence against women (44.3%). “*Cruelty by husbands or relatives*” emerged as the most prevalent crime, illustrating deep-rooted societal issues. Additionally, sexual violence, including *rape* and *assault*, and *kidnapping and abduction*, pose significant safety concerns for women in the state. Kidnapping is also a major crime pervading here (25%) and hence demanding enhanced law enforcement.

**Analysis of Crime Distribution in Odisha (2017-2022)**



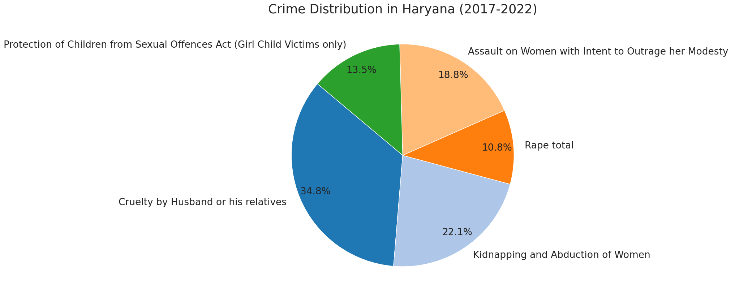
This chart shows “*Assault on women with the intent to outrage her modesty*” comprises the largest proportion of crime against women at 49.9%. This highlights a significant concern for the safety of women in public places. The next most committed crime is “*kidnapping and abduction of women*”, which comprises 18.1% of total crime showing issue around forced abduction.

**Analysis of Crime Distribution in Telangana (2017-2022)**



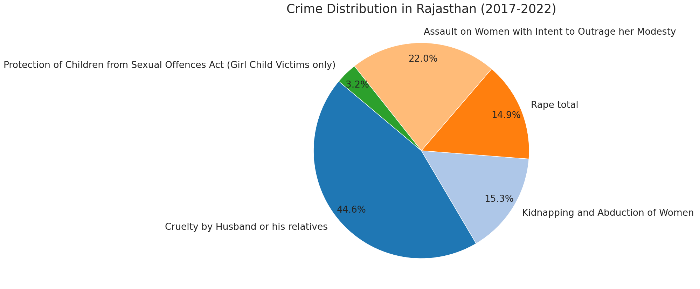
In Telangana, the chart shows that “*Cruelty by Husband or his Relatives*” as the most prevalent crime against women, comprising 48.2%. This against shows high rate of domestic violence and hence reflecting cultural factors and possibly high rate of reporting of this crime. "*Assault on Women with Intent to Outrage her Modesty*" follows with 26.3%.

**Analysis of Crime Distribution in Haryana (2017-2022)**



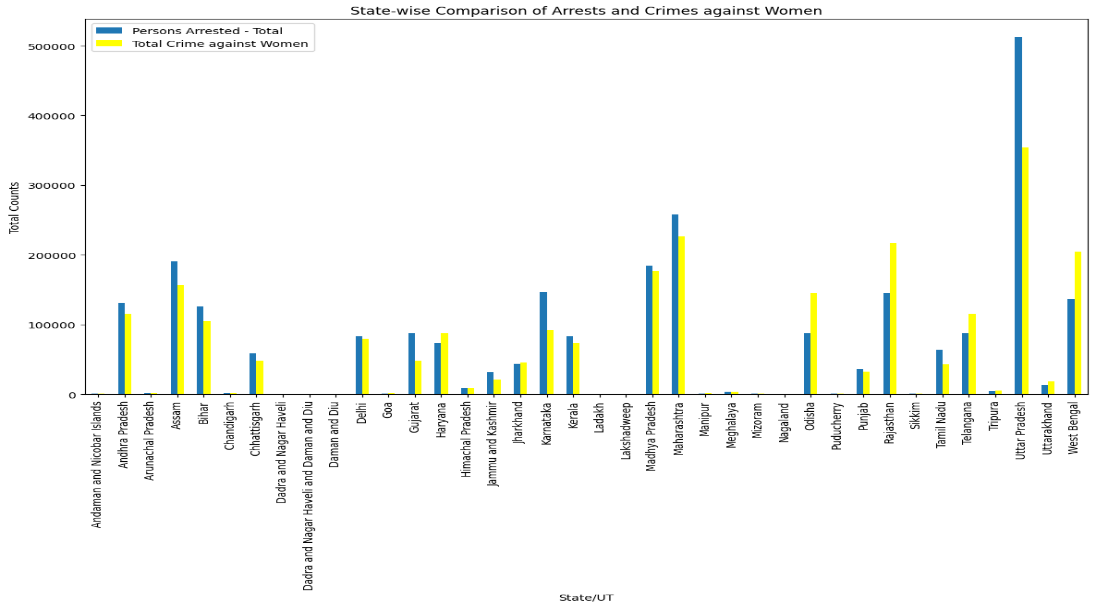
The provided pie chart reveals the prevalence of crimes against women in Haryana during the period 2017-2022. Domestic violence (34.8%), as evidenced by the high percentage of *cruelty by husbands or relatives*, is a significant issue. Sexual offenses, including *rape* (10.8%) and *assault* (18.8%), are also prevalent, highlighting the need for increased awareness and support for victims. *Kidnapping and abduction* (22.1%) pose additional safety concerns for women in the state. These findings underscore the urgent need for comprehensive measures to address violence against women in Haryana

**Analysis of Crime Distribution in Rajasthan (2017-2022)**

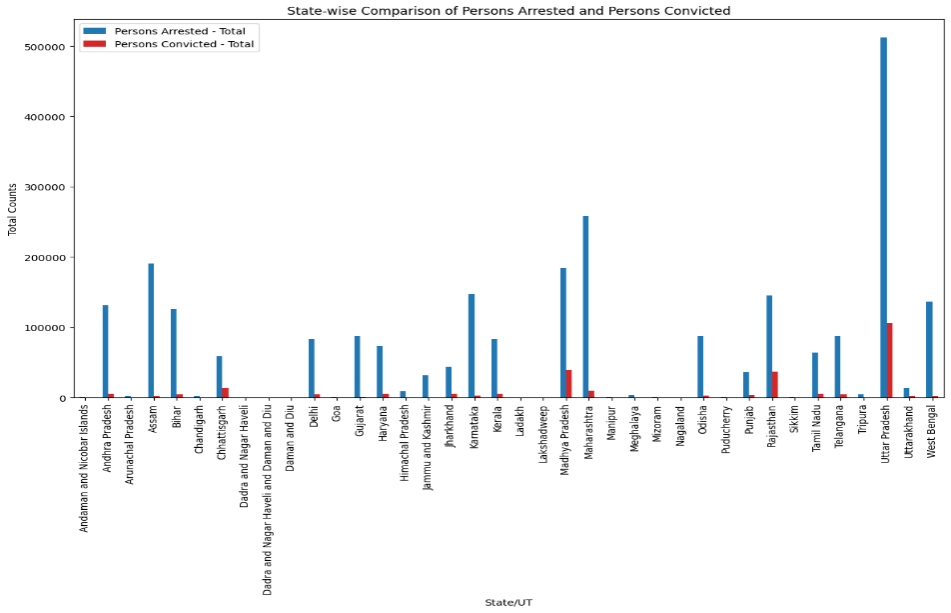


The data presented shows concerning situation of crimes against women in Rajasthan during the 2017-2022 period. High prevalence of *cruelty by husbands or relatives* (44.6%) is a significant issue showing the societal norms. Additionally, sexual offenses, including rape (14.9%) and assault (22.0%), are prevalent, highlighting the need for increased awareness and support for victims. *Kidnapping and abduction*, which comprises 15.3% of total crimes, pose further safety concerns for women in the state.

**3.3 Comparison of crimes against women Vs arrests and conviction (State- wise)**

****

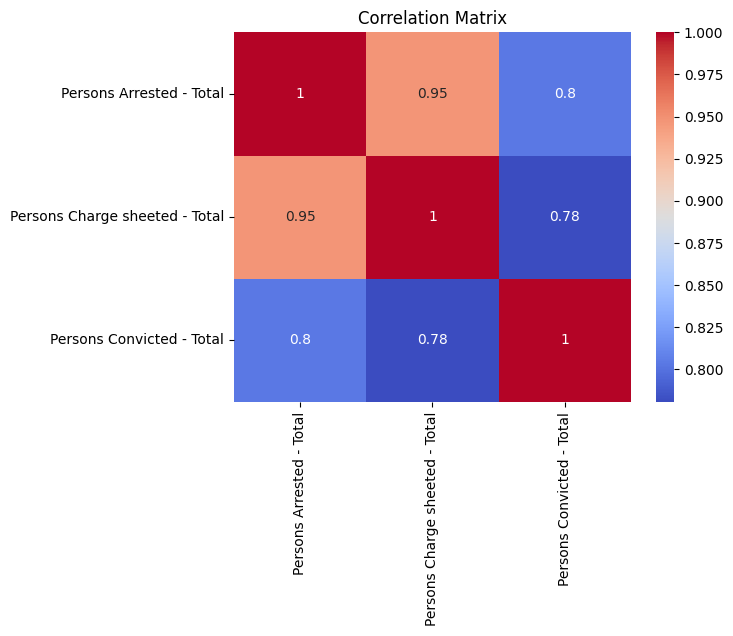
The provided graph illustrates a comparison of arrests and crimes against women across various states and union territories in India. While there is a general correlation between the number of arrests and reported crimes, the data reveals significant fluctuations between states. This suggests that factors such as law enforcement efficiency, reporting rates, and the severity of crimes can influence the discrepancy between arrests and reported cases. Further analysis is necessary to understand the underlying reasons for these variations and to inform effective strategies for addressing crimes against women in India.

****

The provided graph illustrates a significant disparity between the number of persons arrested and those convicted for crimes against women across various Indian states. While arrests often serve as initial steps in the legal process, securing convictions can be a complex and time-consuming endeavor. Factors such as inadequate evidence, delays in the judicial system and the influence of social and political pressures can contribute to the low conviction rates.

The disparity between arrests and convictions is particularly concerning for crimes against women, as it can undermine victims' trust in the justice system and perpetuate a culture of impunity for perpetrators. Addressing this issue requires comprehensive reforms, including strengthening the judicial system, investing in forensic science, and implementing measures to protect victims and witnesses.

**Correlation of different stages of criminal justice**

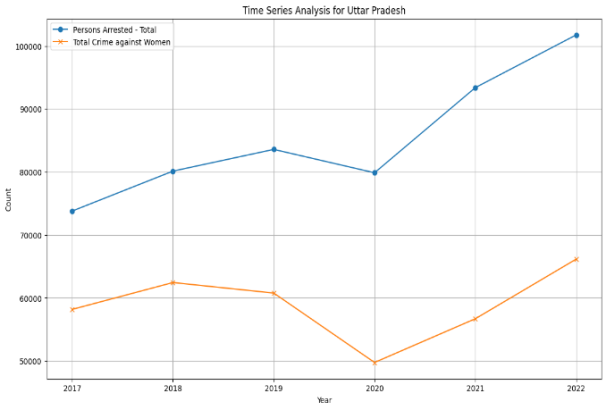
****

The shown correlation matrix provides us a detailed analysis of the relationship between the different stages of the criminal justice process in India and these are ‘number of persons arrested’, ‘number of persons charged-sheeted’ and ‘number of persons convicted’. In the above correlation matrix we get a detailed insight of how each value illustrates the strength of association between these different stages. However It is crucial to note that this doesn’t indicate equality in numbers at these stages.

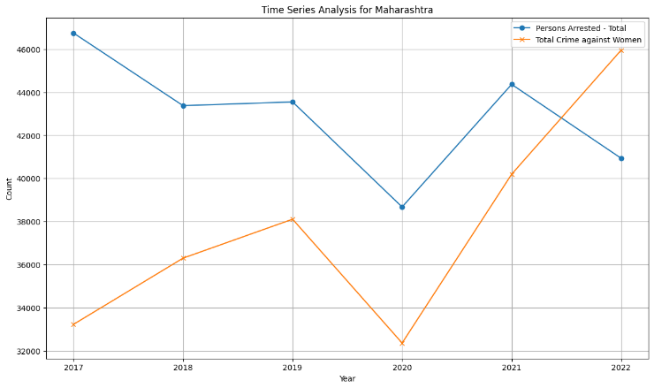
* After doing core analysis one can see a high positive correlation of 0.95 between persons arrested and charge-sheeted. This suggests that if the number of arrests increases then, chances of these cases advancing to charge-sheeting also increases proportionally. It particularly indicates strong dependency on their ratios. This also signifies that most arrested individuals finally led to the charge-sheeting stage.
* Likewise the correlation between arrest and conviction (0.80) and between charge-sheeting and convictions (0.78) suggests the similar pattern , that if arrest increases, conviction increases but the actual number of convictions remain much lower as compared with charge-sheeting and arrests. This  basically suggests the significant attrition rate in the system where not all the charge-sheeted cases result in convictions, possibly due to challenges in proving guilt beyond a reasonable doubt, withdrawal of cases.

Overall, this correlation matrix highlights strong interrelation between the stages of arrest, charge sheeting, and conviction in India; the actual numbers are revealing a different story with very less convictions as compared to arrest or charge sheeting.

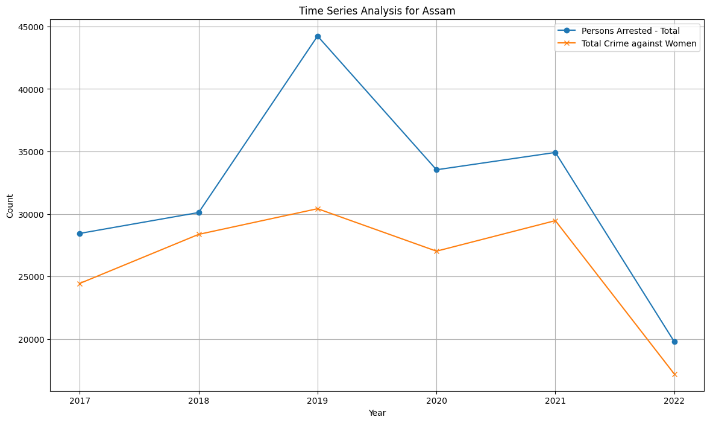
**Time series analysis of crime reported vs person arrested in different states**

****

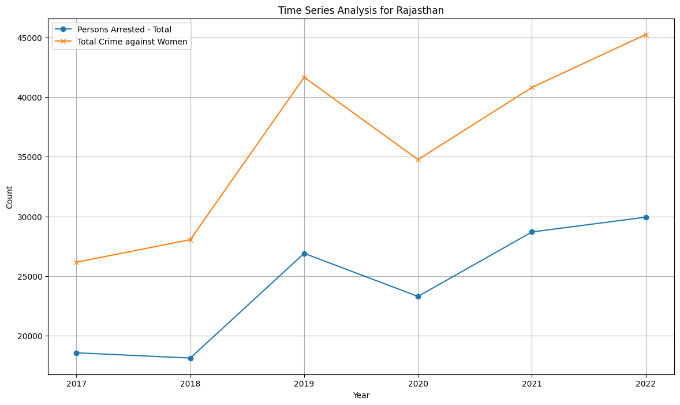
*Uttar- Pradesh:*The graph illustrates a strong growing trend particularly in the year of 2022 in the number of persons arrested and total crime cases against women. This shows an increased focus and concern for the crime against women. It could be due to better reporting mechanism, increased awareness. But this uprising trend calls for further investigation whether this increase in arrests is really helpful in reducing crime activities.



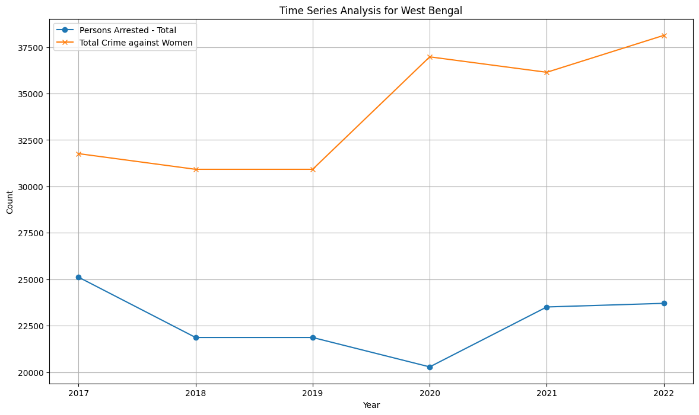
*Maharashtra***:**Graph of Maharashtra indicates a diverging trend between the reported crimes and persons arrested. The crimes against women have increased dramatically,especially after 2020 with peaking in 2022 by 42% as compared with 2020 whereas the arrests have decreased particularly in 2022. This clearly signifies the challenges in law enforcement. This gap needs to be addressed and also a targeted intervention is necessary to lessen this disparity.



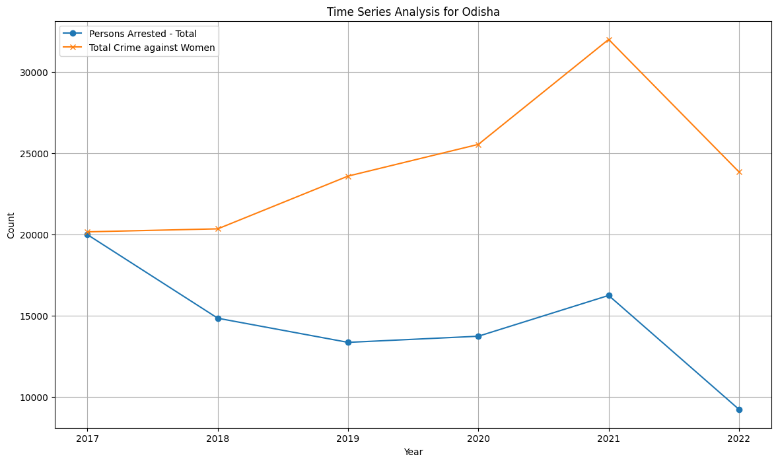
*Assam:*As the graph shows, reported crimes and persons arrested peaked in 2019 but by then it significantly declined in 2022 by 50 %. This decline suggests the good strategy of Assam police of reviewing crimes and better investigation to ensure improved criminal justice system. Another aspect could be whether this represents a true decrease in crime or reflects other factors such as changes in policing, public reporting, or societal conditions.



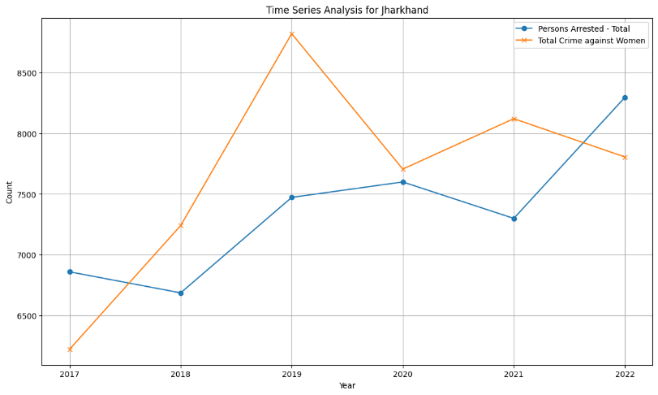
*Rajasthan***:**In Rajasthan the graph trend shows that the number of crimes committed always remained higher than the number of persons arrested with a peak in 2019 and 2022  highlighting the years with heightened crime activity. There is an urgent need for focus and enhancing the arrest rate response.



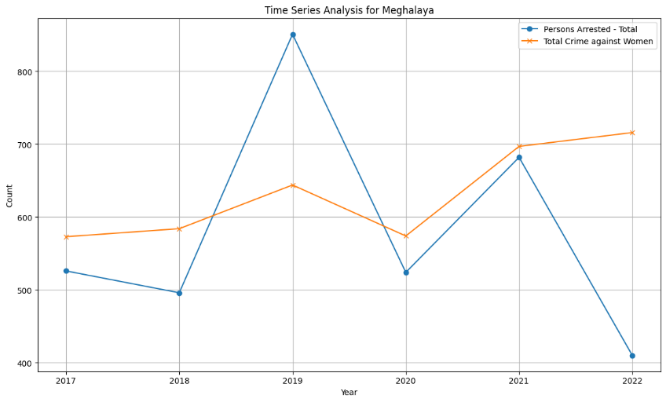
*West Bengal:*Crimes against women in Bengal shows an upward trend, especially from 2020 onwards and in 2020 it increased by 19%, peaking in 2022 around 38000 cases. The number of people arrested follows a stable trend, with a slight dip in 2018. This huge disparity between the increasing trend of crimes and the stable rate of persons arrested showcases the huge challenges in law enforcement efficiency, and increase in crime to such a level that it outpaces the ability to make an arrest. Rise in crime in the post COVID era could be due to social and economic reforms after the pandemic.



*Odisha:*The graph shows an increase in crime until 2021 and then a noticeable decline in 2022 by 27%. The spike in arrest in 2017 and 2021 shows the focused law enforcement and some policy reforms.  In 2022 the decline in crime could be due to improved preventive measures or it can be due to underreporting of crimes.

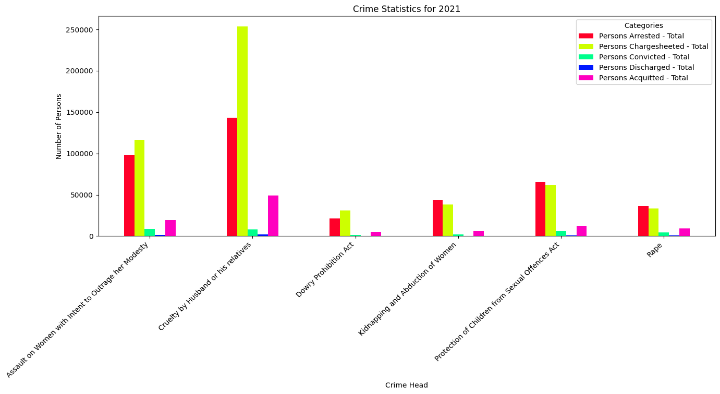
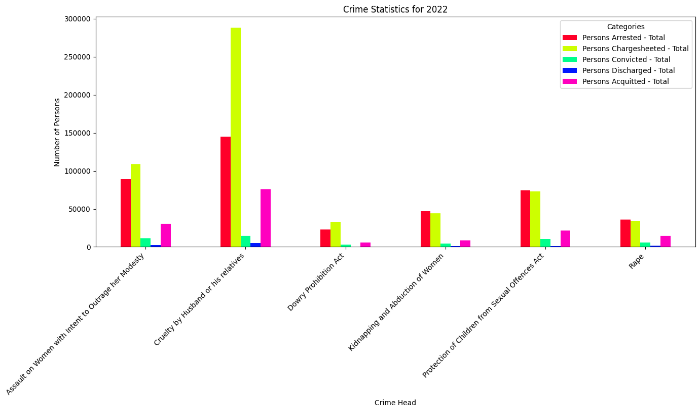


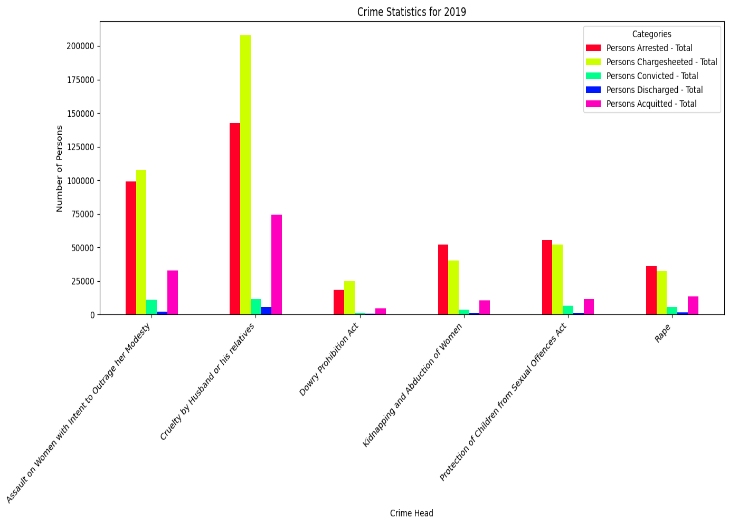
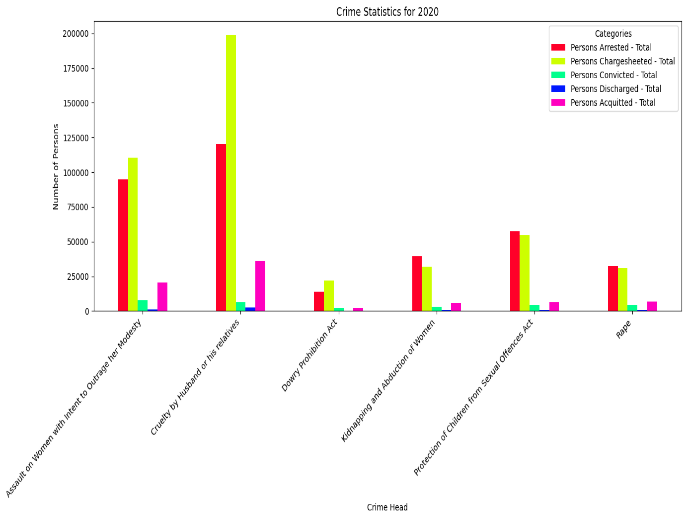
*Jharkhand:* The increase in crime in 2019 by 25%  shows a rise in reporting or could be linked with some events. The decrease in arrest in 2021 could be possibly due to some operational challenges or lack of resources due to pandemic. But the recovery in arrest in 2022 by 13% clearly indicates the restoration of efficiency in law enforcement operation.



*Meghalaya:*The sharp decline in the graph of persons arrested after 2019 around 37-38%, despite the continued rise in crime against women shows some alteration in law enforcement strategy, some external social movement which affected crime rates and inefficiency of policy makers and police of the state. This inconsistent trend also signifies lack of public cooperation.

**3.4 Comparison of crimes against women Vs arrests and conviction (Crime- wise)**





The given bar charts illustrating the crime against women statistics have several unique insights.

**Dominance of some crimes:** Crimes like *“cruelty by husband or his relatives”* with the highest arrest of 2,80,000 and conviction of around 10,000 persons in 2022  stands out as the most dominant and significant category, having the highest number of persons arrested. This indicates domestic violence as a pervasive issue in Indian society and hence it requires forced legal procedures. It clearly suggests, it as a deep-seated issue in our society and has been affecting a number of households, but the high number in cases of domestic violence also shows the increased awareness and willingness among victims to come forward.

*“Insult to the modesty of women”* with approximately 1,10,000 individually arrested and sharp decline in conviction standing at around 10,000 shows a critical gap in legal procedures. Factors like ambiguity in legal definition, evidence challenges like lack of physical evidence or witnesses can be responsible for this. Understanding of gender sensitivity is necessary.

**Discrepancy in legal procedures**: the graph shows noticeable disparity between the person arrested and person conviction in every major crime category. The number of arrests are high but the number of convictions are remarkably low. This also indicates inadequate collection of evidence, issue of witness protection or procedural delays.

Impact of societal attitudes: Above graph shows moderate number of *“dowry prohibition act”*, *“Protection of Children from Sexual Offences Act”* and “*rape cases”*. Over the years of reporting in these crimes has shown growth but still it is significantly lower as –

*Dowry prohibition act*: Here, It is difficult to get concrete evidence as transactions during the dowry usually happen privately and involve complex cultural and family dynamics. This, itself complicates the legal procedures as sometimes there is reluctance in family members to testify about the crime.

*Protection of children from sexual offenses act (POCSO):* Here the low conviction rate is due to judicial process, which sometimes doesn’t include child-friendly procedures, lack of specialized training for police and officers to handle such sensitive cases or fear of societal stigma which usually prevent the gathering of evidence.

*Judicial gaps in rape cases:* decline in conviction can be because of social stigma and societal and family pressure, which might result in weak testimonies. Sometimes there are instances where victims are blamed which prevent victims from aggressively pursuing their cases.

The data clearly suggests that the judicial system is proactive in terms of arrest and chargesheets but there is utmost need to make some judicial reforms to bridge the gap between arrests and conviction.

**3.5 Gender Analysis of the crimes committed against women**

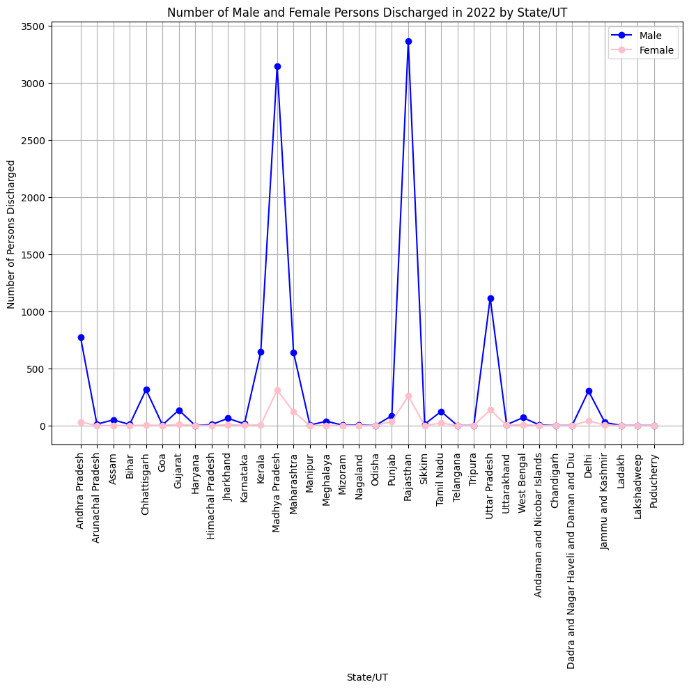
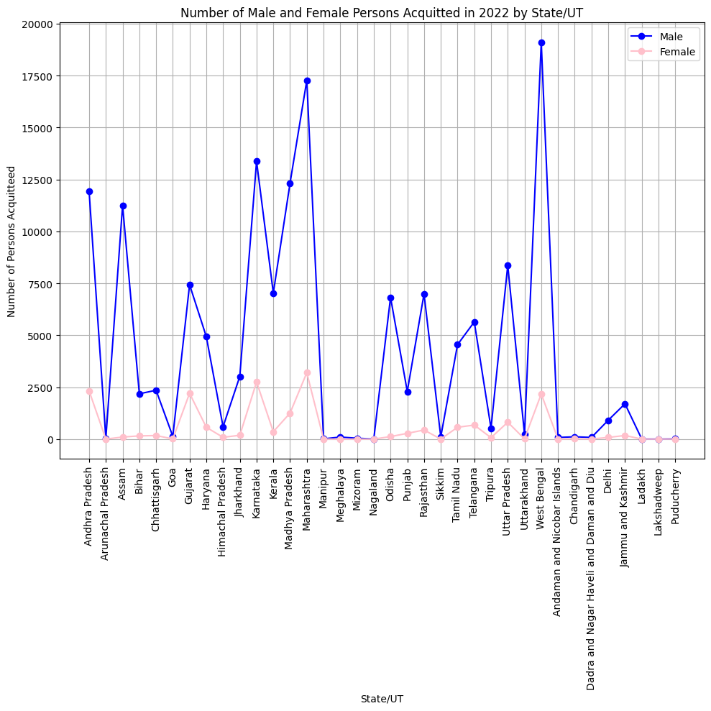
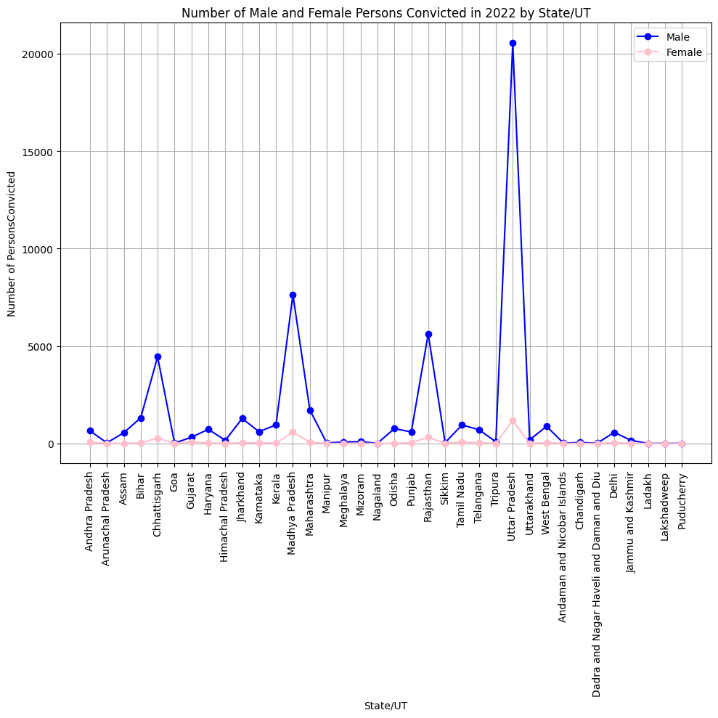
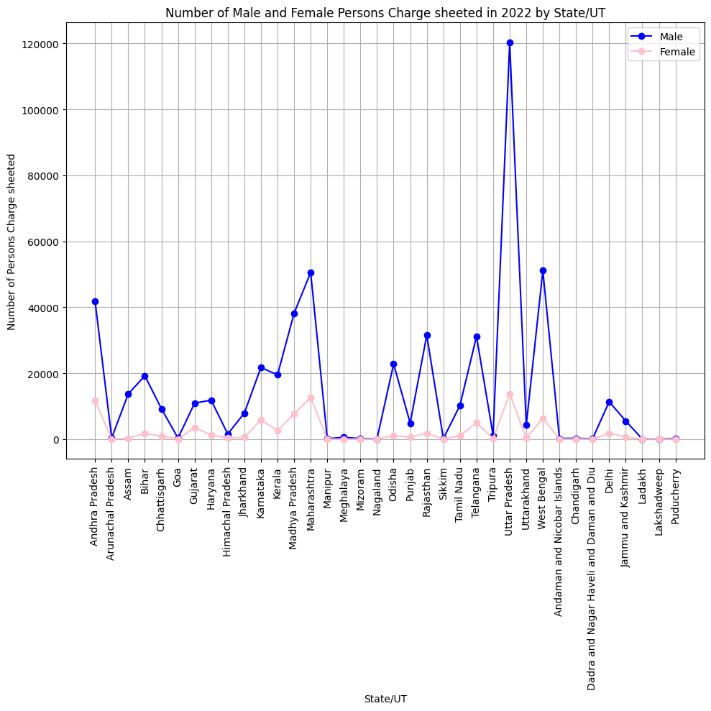
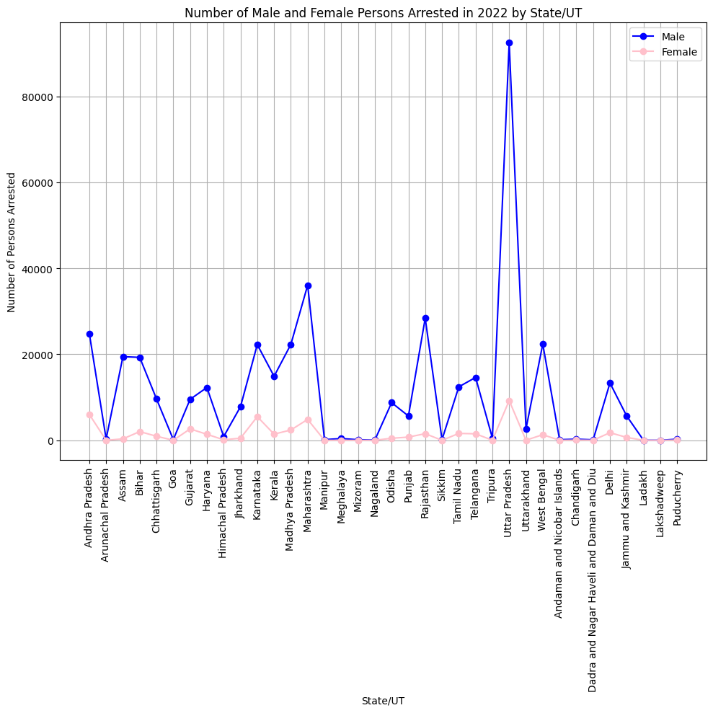
These graphs show the male(blue) and female(pink) arrested, charge sheeted, convicted, acquitted and discharged during 2022. It also shows that both men and women are involved in committing crimes against women, however the men the disproportionally far more involved.

**Key Findings:**

* There is an overwhelming male dominance in crimes committed across all states and union territories. The number of male committed crimes far exceeds that of the female committed crimes which suggests a strong correlation between gender and criminal activity.
* It is significantly visible that the pink line remains notably below the blue line, implying consistent gender disparity in crimes committed. This trend is particularly pronounced in the states with higher number of crimes committed.
* States like Uttar Pradesh show an overwhelming number of male arrests (over 80,000) compared to the female arrests (which are mostly below 10,000 in any state) revealing crimes against women are primarily perpetrated by men. Andhra Pradesh, Maharashtra, and Karnataka also have a relatively high number of arrests but still male arrests are disproportionately higher than the female arrests.
* The states specific variations of the magnitude of the gender gap suggest that we must consider regional differences in socio-cultural factors and law enforcement practices.
* To quantify the gender gap, the crimes committed by women constitute for less than 10% of the total crimes committed.

**Implications of the Findings:**

* **Sociocultural Factors:** The male dominance in criminal rate may be attributed to gender roles, societal norms, and cultural expectations which play a significant role in shaping criminal behaviour.
* **Law Enforcement Practices:** The given graphs also raise arguments about potential biases in law enforcement practices, such as selective enforcement or dissimilar treatment of male and female suspects. There might also be difference in victimization patterns and leniency in sentencing
* **Underreporting:** There is also the possibility of underreporting in crimes involving female offenders, particularly in cases of sexual assault and domestic violence.



**Yearly gender trends of crimes committed against women**



The number of males convicted in 2022 is higher than that of 2017 implying that crime committed by males has only increased over the years. Several factors may be responsible for this increase such as change in law enforcement practices, legal system reforms like changing sentencing guidelines of bailing policies, more awareness about reporting systems, some eradication socio-cultural taboos, and increase in population.

However, on the other hand, the number of female convicts in 2022 is lower than that 0f 2017 and 2018, despite the increase in socio-economic burdens and domestic tensions post lockdown. Analysis suggests that the number of female convicts has decreased, it is essential to consider the possibility of underreporting in cases involving female offenders, particularly in instances of domestic violence or sexual assault. This could potentially skew the data. The decline in female convictions might also be due to some law enforcement biases leading to dissimilar treatment of male and female offenders, leading to leniency in sentencing, etc.

1. **Feature Scaling**

Feature scaling plays a very important role in preparing the dataset for ML algorithms. It ensures that all independent variables are standardized within a particular range. This needs to be implemented as it will be very helpful in improving the performance of our models: random forest regressor, XG boost, and Light GBM. Here the technique of standard scaling has been used to achieve the standardization.

After this standardization, the models have been applied on various crime columns in our dataset like 'Cruelty by Husband or His Relatives', 'Kidnapping and Abduction of Women', 'Rape Total', 'Assault on Women with Intent to Outrage Her Modesty', 'Insult to the Modesty of Women’, ‘Protection of Children from Sexual Offences Act (Girl Child Victims s Only','Total Crime Against Women'.

Formula of standard scaling:

With the given formula, data is transformed such that it has a mean of 0 and standard deviation of 1.

‘z’ is the normalized value.

‘x’ is the original data point

‘mean’ is the mean of the dataset

‘std\_dev’ is the standard deviation of the dataset.

After applying feature scaling, the dataset has been divided into two parts(75% and 25%). 75% for training the model and 25% for testing the various machine learning models.

1. **Prediction of crime against women using various machine learning models and algorithms.**

Here we are using three cutting-edge machine learning models – Random forest Regressor, XG Boost and Light GBM- to predict the patterns of crime against women. Each model has its own advantages. Random Forest Regressor has the ability to handle complex datasets and is renowned for its robustness, XG Boost is very helpful in optimizing the predictive performance by using gradient boosting and Light GBM offers scalability and efficiency especially with big datasets.With these models, we aim how well they predict different crime types and which is the best and reliable model in case of our dataset.

Some formulas, which will help in analyzing the performance of Machine learning models:

* RMSE = , where n signifies the number of data points, x(j) denotes the actual value for jth measurement, and (j) denotes the predicted value of jth measurement.
* R-Squared error =

=

Precision : Ratio of positive classifiers and total number of positives.

* F1 Score = (F1 Score is the harmonic mean of precision and recall)
* Recall = (the proportion of all actual positives that were classified correctly as positives)

**5.1Random Forest Regressor**

In order to model and predict the crime against women, a robust learning method i.e Random Forest Regressorhas been employed. This model works by building an extensive set of decision trees and give mean prediction of each tree. Random forest is ideal choice for the complicated datasets and is known for its resilience to overfitting.

Here, Random Forest has the efficient ability to identify the non-linear relationship between the input variables such as States/UT, year and categories and the target variable. This model has the potential to deal with the high dimensional dataset. Its embedded feature important capacity helps in identifying the most crucial factors .

In accordance with the evaluation metrics, including  R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE), the Random Forest Regressor has shown good performances in variety of crime categories.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Crimes committed against women | RMSE value (mean squared error) | MAE value (mean absolute error) | R-squared score | Cross- validation (R-squared) |
| Cruelty by Husband or his relatives | 2816.67 | 1869.77 | 0.71 | 0.88 |
| Kidnapping and Abduction of Women | 1908.72 | 1168.53 | 0.67 | 0.87 |
| Rape total | 648.84 | 423.57 | 0.69 | 0.84 |
| Assault on Women with Intent to Outrage her Modesty | 1837.13 | 1327.00 | 0.64 | 0.87 |
| Insult to the Modesty of Women | 154.18 | 106.77 | 0.71 | 0.92 |
| Protection of Children from Sexual Offences Act | 1082.59 | 730.90 | 0.53 | 0.81 |
| Dowry Prohibition Act, 1961 | 630.72 | 333.52 | 0.69 | 0.84 |
| Total Crime against Women | 7485.75 | 5231.02 | 0.73 | 0.92 |

**5.2 XG Boost**

XGBoost algorithm is a powerful learning method which has been utilized in order to predict and model crimes against women. XGBoost constructs a series of decision trees sequentially, and in it each tree aims to correct the errors of its predecessor. This model uses gradient boosting to optimize its predictive performance, making it well-suited for datasets which are complex and have intricate patterns.

XGBoost captures the non-linear relationships between input variables such as States/UT, year, and crime categories, and the target variable. It identifies the most influential factors required for making the predictions, using its advanced feature importance mechanism. XGBoost has demonstrates strong performance across various crime categories as it is evaluated using metrics like R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE), which showcase the robustness and reliability of this predictive model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Crimes committed against women | MAE value (mean absolute error) | RMSE value (mean squared error) | R-squared score | Cross- validation (R-squared) | F1 Score | Recall |
| Cruelty by Husband or his relatives | 1156.72 | 1655.49 | 0.90 | 0.91 | 0.79 | 0.87 |
| Kidnapping and Abduction of Women | 905.65 | 1348.25 | 0.83 | 0.88 | 0.79 | 0.85 |
| Rape total | 255.01 | 384.52 | 0.89 | 0.88 | 0.88 | 0.85 |
| Assault on Women with Intent to Outrage her Modesty | 958.38 | 1405.30 | 0.79 | 0.90 | 0.71 | 0.76 |
| Insult to the Modesty of Women | 51.09 | 80.11 | 0.92 | 0.93 | 0.83 | 0.88 |
| Protection of Children from Sexual Offences Act | 461.78 | 655.33 | 0.82 | 0.83 | 0.64 | 0.81 |
| Dowry Prohibition Act, 1961 | 271.71 | 533.63 | 0.75 | 0.79 | 0.85 | 0.85 |
| Total Crime against Women | 3148.06 | 4446.97 | 0.90 | 0.93 | 0.80 | 0.83 |

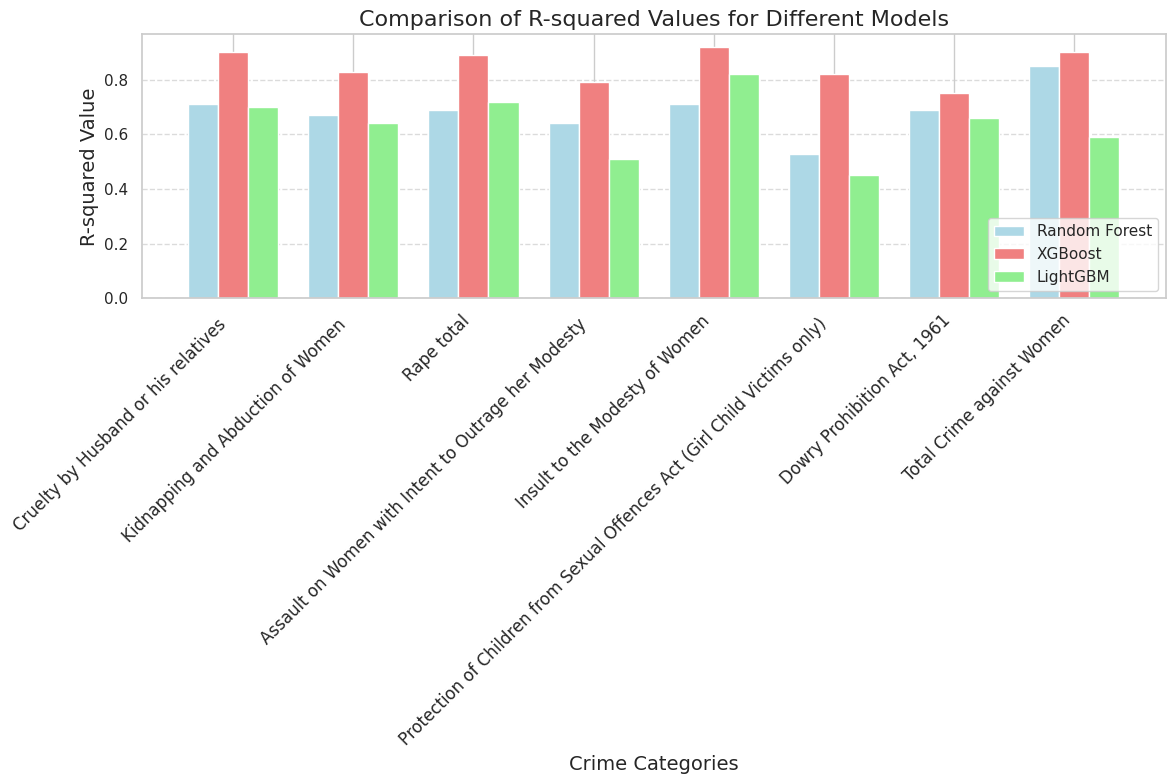
**5.3 Light GBM**

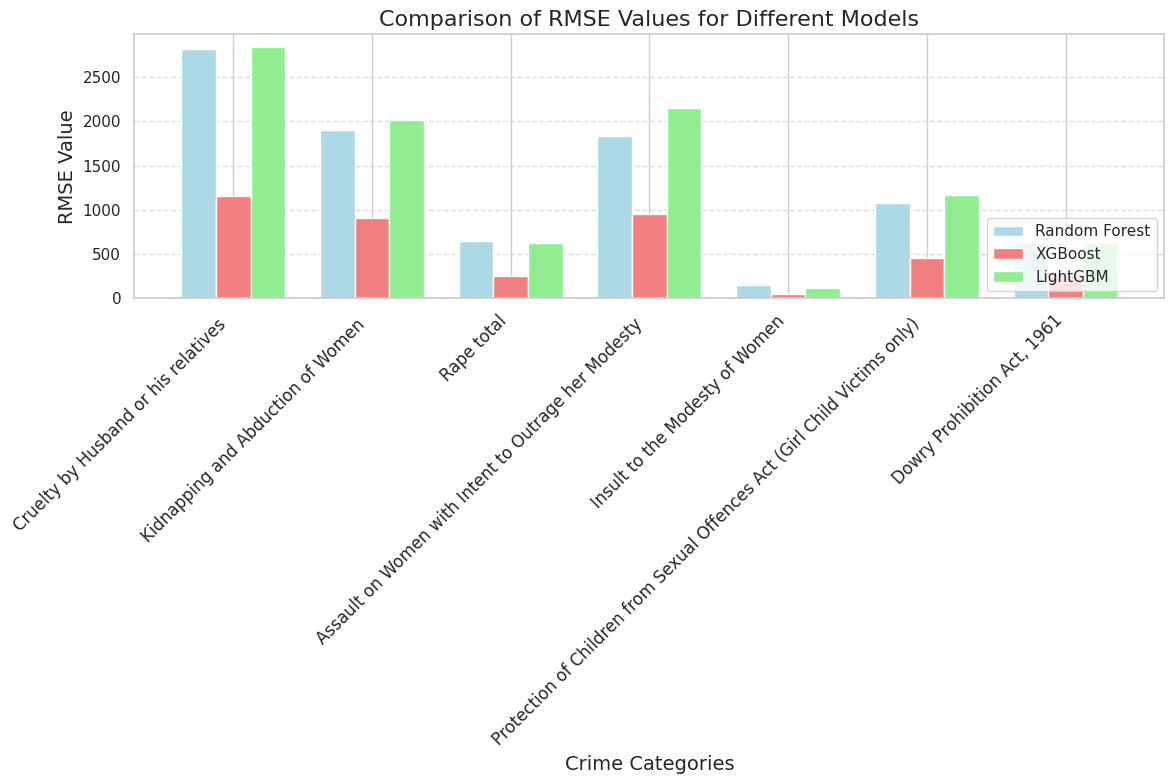
The LightGBM algorithm is an efficient learning model which has been employed to analyze and predict crimes against women. This algorithm is designed to handle big datasets with high-dimensional features. It uses a technique called gradient boost in order to build decision tree. LightGMB is for its scalability and speed, making it an optimal choice for processing or vast criminal dataset.

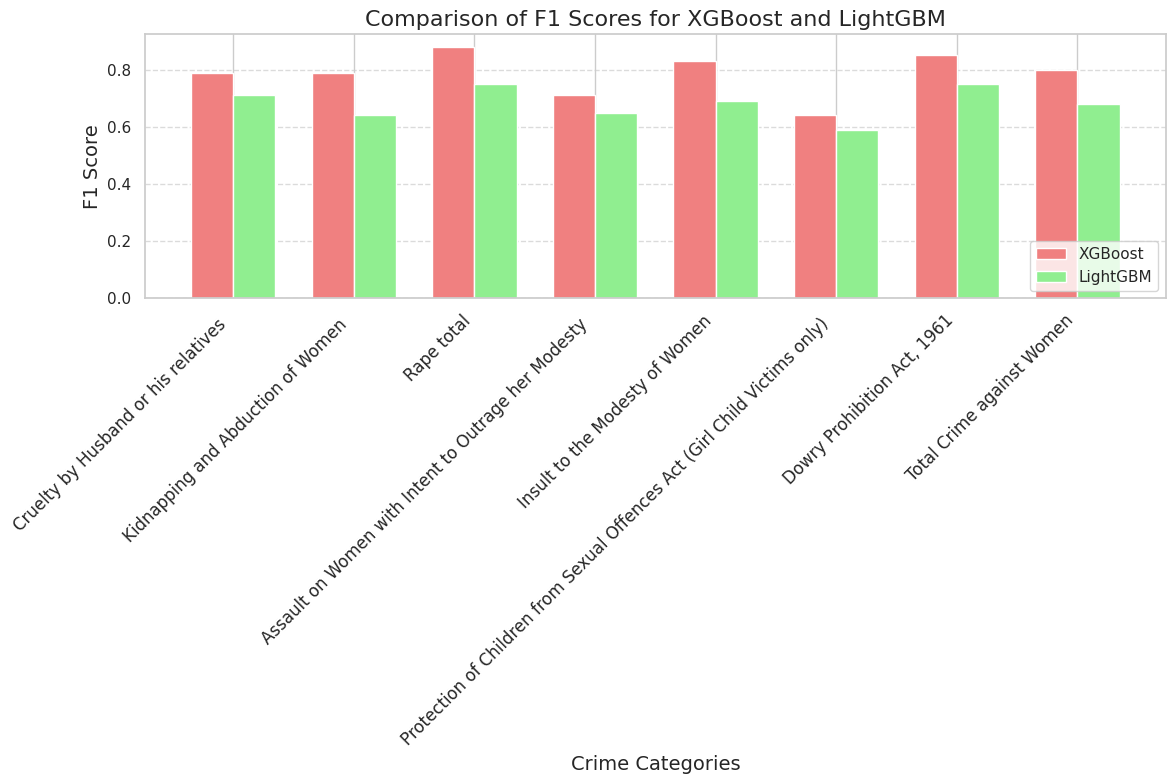
LightGBM effectively identifies complex, non-linear relationships between input variables such as States/UT, Year, and other crime columns, and the target variables. It has an embedded feature importance mechanism which gives it the ability to prioritize important features which further enhances the model’s accuracy. LightGBM delivers an excellent performance across different crime categories, as it is evaluated using key metrics like R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE), proving its effectiveness in predictive modeling.

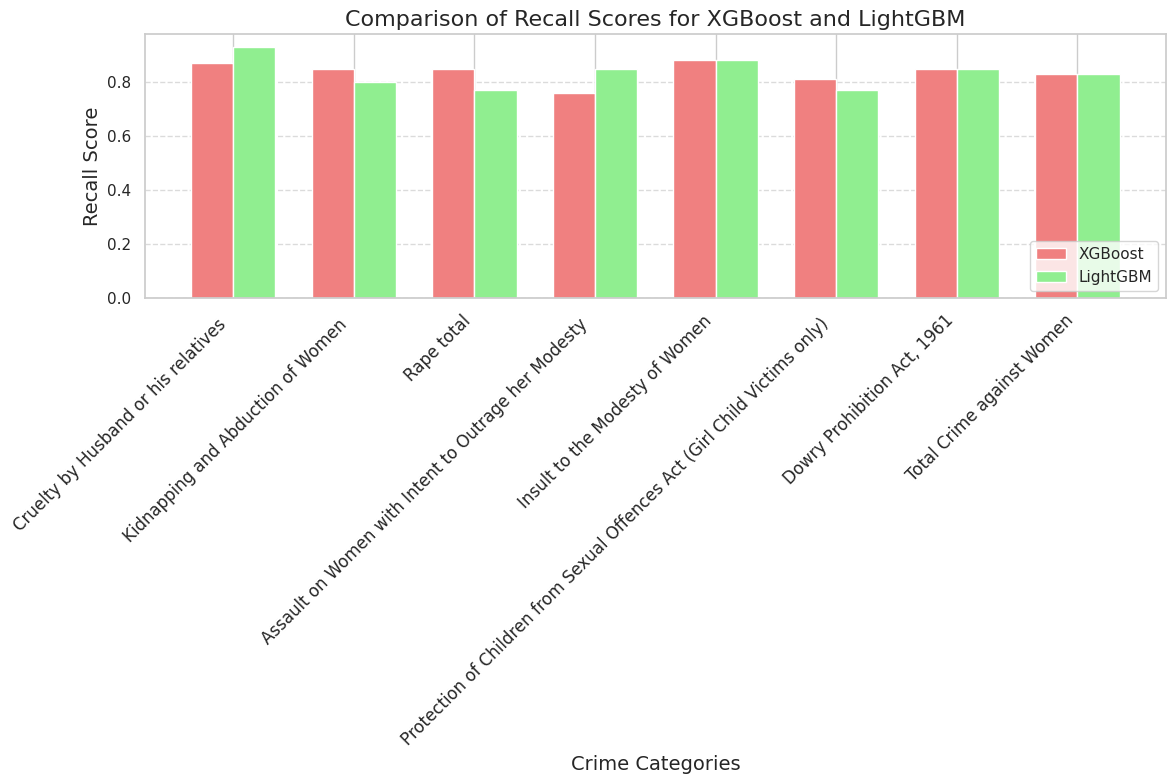
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Crimes committed against women | RMSE value (mean squared error) | MAE value (mean absolute error) | R-squared score | Cross- validation (R-squared) | F1 Score | Recall |
| Cruelty by Husband or his relatives | 2842.58 | 1951.77 | 0.70 | 0.80 | 0.71 | 0.93 |
| Kidnapping and Abduction of Women | 2014.91 | 1331.80 | 0.64 | 0.82 | 0.64 | 0.80 |
| Rape total | 623.34 | 440.23 | 0.72 | 0.77 | 0.75 | 0.77 |
| Assault on Women with Intent to Outrage her Modesty | 2155.47 | 1599.51 | 0.51 | 0.81 | 0.65 | 0.85 |
| Insult to the Modesty of Women | 120.07 | 91.34 | 0.82 | 0.83 | 0.69 | 0.88 |
| Protection of Children from Sexual Offences Act | 1173.31 | 849.26 | 0.45 | 0.78 | 0.59 | 0.77 |
| Dowry Prohibition Act, 1961 | 629.58 | 377.22 | 0.66 | 0.78 | 0.75 | 0.85 |
| Total Crime against Women | 9287.77 | 6495.33 | 0.59 | 0.83 | 0.68 | 0.83 |

**5.4 Comparison of used machine learning models**

****

****

****

****

1. **Challenges faced**

During this research, the data collection aspect was quite challenging, as it was essential and crucial for our different crime-based studies. The three major issues we encountered were:

* **Incompatibility in the data**: The data collected between 2017 and 2022 was presented in different format and includes multiple parameters. Moreover, the crime data of 2017 was less detailed. There were some unreported crimes which leaves gap in the record.
* **Inconsistencies in the data formats**: The data of 2020 was available in PDF format and had to convert that in excel format. Rest of the datasets were in excel format itself.
* **Crime data of large volume**: The datasets consist of numerous columns encompassing 19 sub-categories of crimes against women. The given information was collected for all states and UTs of India. So, making it compatible for analysis was a challenging task.
* **Dealing with multiple datasets**: This made the research work more complex, as each dataset has its own structure, format and some underlying issues that made integrating and analyzing them together a difficult task.

1. **Conclusion**

This research paper provides an insightful analysis of various types of Crimes against Women in India from 2017 to 2022 and makes usage of different kinds of machine learning algorithms to predict crime trends with the identification of high-risk regions. This study focuses on the importance of integrating data analytics and machine learning models such as XG Boost, Light GBM and Random Forest Regressor with the system of legal and criminal justice.. The study also illustrates the specified important findings:

* The highest numbers of reported cases are in Uttar Pradesh but according to our analysis Assam is on the top of the list if we consider crime rates relative to population, showcasing unexpected risk in the region.
* Crime committed against women is lowest in 2020 i.e. during the COVID-19 Pandemic but after it, crimes continued to rise. Crimes rebounded in 2022, driven by the social-economic tensions spreading in India post-pandemic.
* There is notable rise in crime committed by both men and women after 2020, with increase in male-perpetrated crimes being more pronounced as compared to 2017-2019. In contrast crimes committed by women decreased in comparison to that timeframe highlighting to potential bias in law enforcement such as treating male and female suspect differently, or underreporting of cases.
* The number of cases reported and the number of arrests was relatively equal but there was much lower conviction rate.
* Arrests were notably higher is northern India as compared with southern and north-eastern states and Union Territories which highlights regional disparity in law enforcement.

This research makes an effort in assisting the policy makers to implement efficacious measures to provide women a safer environment by predicting the future crime patterns. All the findings reveal an urgent need for targeted efforts and some reforms in law and policy for women safety. These advance cutting-edge machine-learning methods showcase their potential that how they can assist in proactive decision making in the battle of crimes committed against women.

1. **References**