VIT-AP UNIVERSITY

Statistical Inference

Mini Project

"District wise Dowry Death Prediction"

Submitted to,

Prof. Dr. Santanu Mandal

Submitted by,

Acquin Joseph[22MSD7012]

Shivam Sen Gupta[22MSD7016]

Anjali K[22MSD7029]

Krishnendu K R[22MSD7030]

Sultan Fahmi [22MSD7002]

Place: VIT-AP University

Date: 25-01-2023

ABSTRACT

This mini project is based on dowry death of women in India. Dowry death is one of the most hideous or gruesome and burning issue in India. We decided to dig up on this topic so that we get an idea of the factors effecting and the changes that we could make which reduces dowry death. Our analysis is based on crime against women data between 2001-2020. For this project, which has a reference list given at the end the report, primarily two data set, the Indian census data collection and the Indian women crime data set were employed. A key hypothesis to explain the rise in dowry death is decrease in illiteracy and unemployment. Others include area wise population, working status of male and female, etc. In 2010 Supreme court judgement made it compulsory to establish prior harassment of a victim by the male spouse and his relatives arising from dowry. Besides, aggressive masculinity can't be curbed without addressing powerful influences of gender norms and system of inequality. In brief, the challenges of curbing the growing menace of dowry deaths are many but effective solutions are few. Imported the necessary libraries and completed all pre-processing on both sets of data, including district-wise grouping by mean in order to start connections between them. The statistical method used is correlation which helps in finding the strength of relationship between each attribute. We may then determine which elements have the greatest, minimal impact and act appropriately. This also includes data visualisation, which helps people interpret the information. We employed two techniques, including machine learning and time series analysis, for better dowry death prediction. The two machine learning models that we utilised for a precise prediction were the Huber regressor and the Bagging regressor. As a conclusion, we realised that unemployment and widespread illiteracy are the key factors that have the greatest impact on dowry deaths. It has also been demonstrated that beneficial adjustments to these aspects can reduce dowry deaths.

Key words: **Dowry Death, Illiteracy, crime, unemployment**

Table of Contents

INTRODUCTION

The death of women due to dowry is a major concern in India. The rate of dowry deaths varies greatly from region to region within India. The death of women is greatly endangered by the practise of dowry. In Indian culture, it is traditional for the bride's family to present the groom with a valuable gift or a sum of money at the time of the wedding. This tradition dates back to ancient times, when women were given cash and jewels by their families after their marriages, which they then saved as a means of achieving financial independence. Later, violence against women was blamed on the system. Dowry will become a challenge after marriage. If the wife is unable to give the wealth that the groom family expects, they will begin to treat her badly.

Even in the present period, dowry is given priority over a girl's aptitude, intelligence or moral character. The main cause of domestic violence is dowry. The dowry system has become a must-follow tradition in Indian society. Regardless of whether the bride's family is wealthy or not, every family is expected to provide the money or assets requested by the groom's family. When the woman's family does not provide a significant amount of dowry, domestic violence will result.

In the year 1961, the Dowry prohibition act was passed which prohibits the practice of accepting and giving dowry. Then also people in India are not considering dowry system as illegal. People in India who are aware of the law and its execution 60 years ago are still not following it. Many murders have continued to take place because of not giving dowry. Most of the victims of dowry death in India are young women who can’t withstand the harassment and torture and committed suicide by consuming poison or hanging themselves.

Through the project, we hope to demonstrate how and by what means we may reduce dowry deaths in India. What are the many elements that affect dowry deaths? and attempting to foretell dowry death by taking into account dependant elements. Through research and the consideration of influencing circumstances, we are attempting to forecast the number of dowry deaths that will occur in the upcoming year.

METHODOLOGY

CORRELATION

Correlation is a statistical technique that aids in determining the degree of relationship between various quality.

We attempt to establish a correlation between dowry deaths and a number of variables, including personal literacy rate, male and female literacy rates, main workers, marginal workers, nonworkers, female and male birth rates etc. Additionally, we used the Python __ Library to visualise the association between each factor and dowry death.

BAGGING REGRESSOR

Huber regression is a regression technique that is too robust to outliers. The idea is to use a different loss function rather than the traditional least squares.

A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

In this project we use bagging regressor to get the accurate prediction with Huber regressor as the beta estimator.

TOOLS AND ANALOGY

Python is the key tool used in our project. Data analysts and experts can use Python as a powerful tool to do complex statistical calculations, produce data visualisations, develop machine learning algorithms, manage, and analyse data.

Data Pre-processing:

- NumPy
- Pandas

Data Visualization:

- Matplotlib
- Plotly
- Seaborn

Machine Learning:

- Sci-kit Learn
    - Bagging Regressor
    - Huber Regressor

DISCUSSION AND FINDINGS

Data mining is the first phase. A crucial component of data analytics is data mining. We use two datasets for our study: a crime dataset and a census dataset. The crime dataset offers district-by-district information on the various crimes against women from 2001 to 2020 in India. Additionally, the census dataset offers district-level data for all censuses. The cleaning of the provided data is the second and possibly most crucial stage. Unsuitable features, such as those that are nominal, ordinal, numerical or that occasionally need to be transformed might cause problems with data. Step three involves cleaning up the data and making it flexible for our projects point. By providing a mean aggregative function, group the dataset according to district using groupby function in python.

However, considering that it is 2023, all women are speaking up against everything that affects them, including young children. Consequently, a negative connection might be detected when examining the relationship between year and dowry death. That is, the number of dowry deaths is declining each year. However, it still occurs most commonly, mainly in rural regions. The below-presented visualisation of the worked-on data enables the observer to see the outcome.

Code:

```
sns.barplot(x="YEAR",y="DOWRY DEATHS",data=df)
```

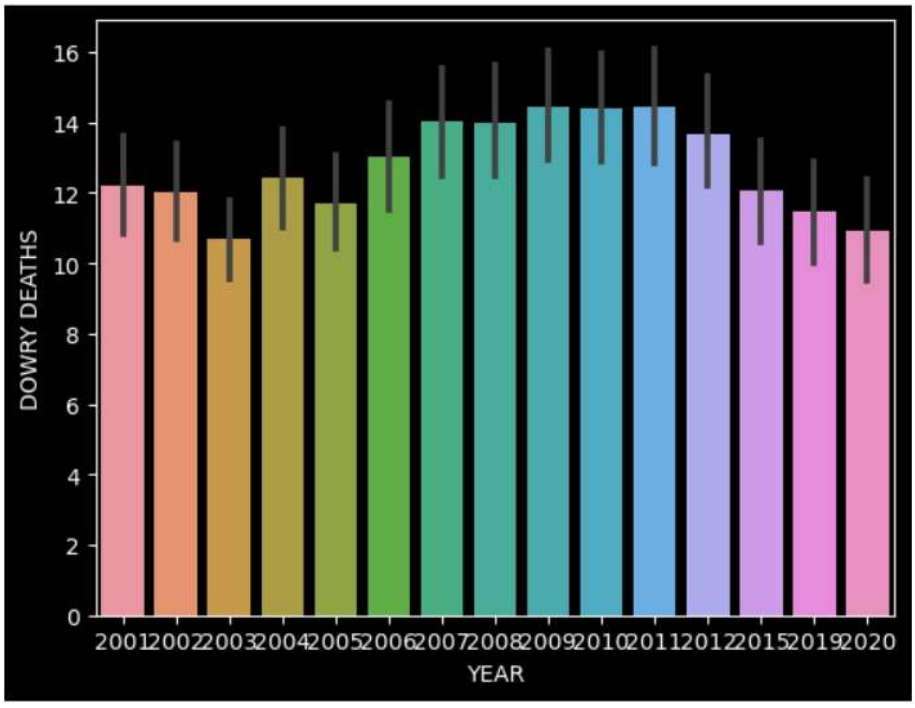

Fig 1.0

PRIOR 1

Identifying the contributing factors to dowry deaths that can be reduced in India and determine how other factors relate to dowry death.

- Visualizing relation between dowry death and female literacy rate in India by using

Code:

pt.scatter(df_vis,x='DOWRY DEATHS', y='Flit_Rate', color="state",log_x=True, size_max=30,trendline='lowess')

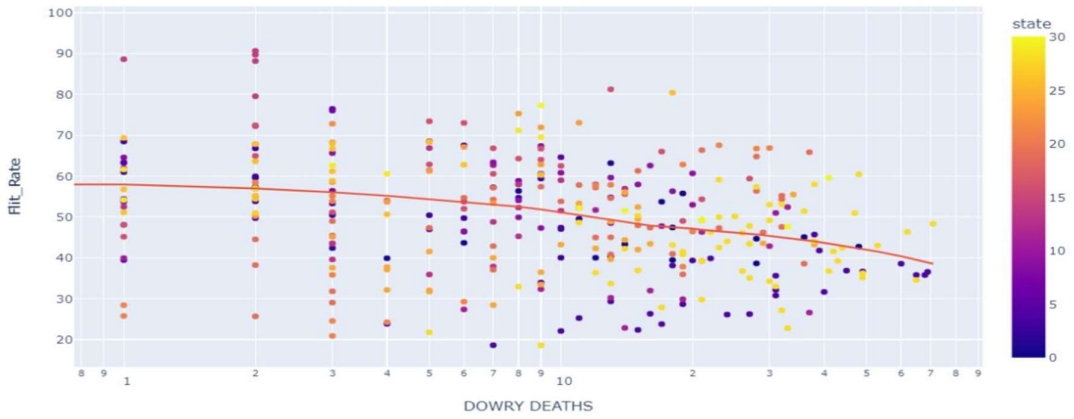

Fig 1.1

We can deduce from Fig. 1.1 that as female literacy increases, the number of dowry deaths declines. Thus, it is clear that dowry death is decreased by a high female literacy rate.

- Relation between dowry death and male literacy rate by using

Code:

pt.scatter_3d(df_vis, x='DOWRY DEATHS', z='Mlit_Rate', y='state',color='state', opacity=0.7)

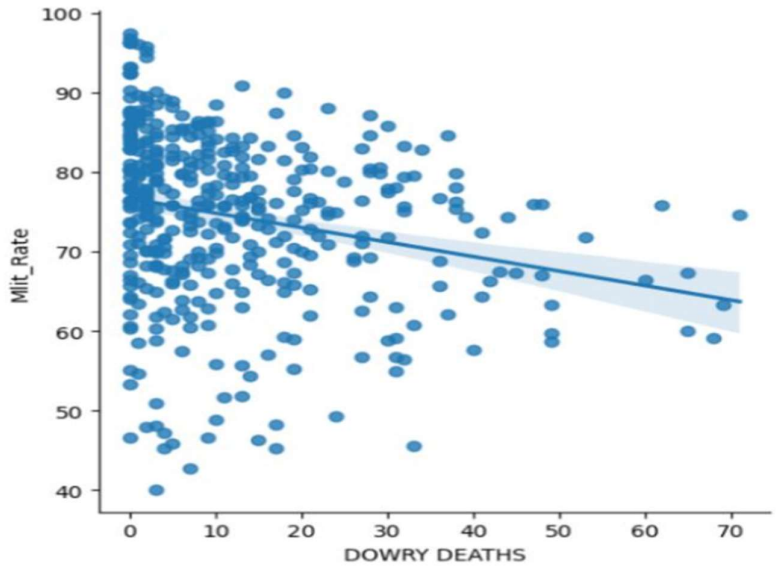

Fig 1.2

We can deduce from Fig. 1.2 that as male literacy increases, the number of dowry deaths declines. Thus, it is clear that dowry death is decreased by a high male literacy rate.

- Visualising relation between dowry death and nonworkers.

Code:

```
pt.scatter(df_vis,x='DOWRY DEATHS', y='Nonworkers', color="state",log_x=True, size_ma
x=30,trendline='ols')
```

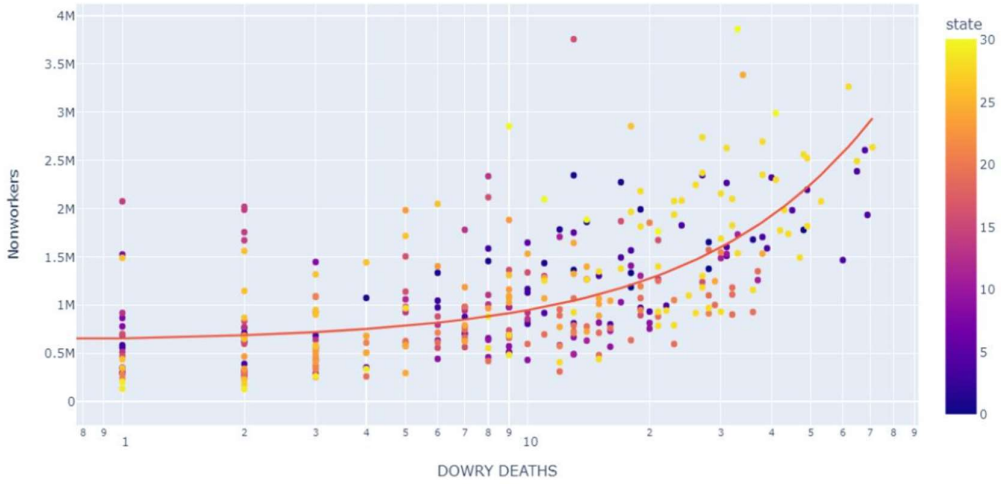

Fig 1.3

Figure 1.3 demonstrates that when the number of non-workers rises, dowry deaths also rise. In India, unemployment is a growing problem that will cause a spike in dowry deaths.

- Visualizing relation between dowry death and marginal workers.

Code:

```
pt.scatter_3d(df_vis, x='DOWRY DEATHS', z='Marginalworkers', y='state',
              color='state', opacity=0.7)
```

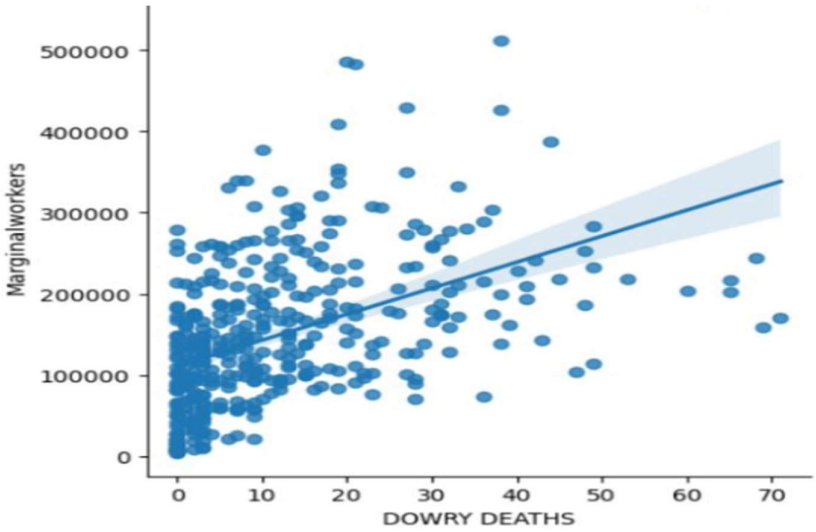

Fig 1.4

Figure 1.4 illustrates that when the number of marginal workers rises the dowry deaths also rises.

- Visualizing relation between Dowry death and personal literacy rate.

Code:

```
pt.scatter(df_vis,x='DOWRY DEATHS', y='Personlit_rate', color="state",log_x=True, size_max=30,trendline='lowess')
```

Fig 1.5

We can deduce from Fig. 1.1 that as personal literacy increases, the number of dowry deaths declines. Thus, it is clear that dowry death is decreased by a high personal literacy rate.

CORRELATION

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable. Two or more variables considered to be related, in a statistical context, if their values change for one the variable i.e., if one of the variable increase or decreases so does the value of the other variable (although it may be in the opposite direction).

Code:

```
sn.heatmap(corr,linewidth=.5,cmap="crest")
```

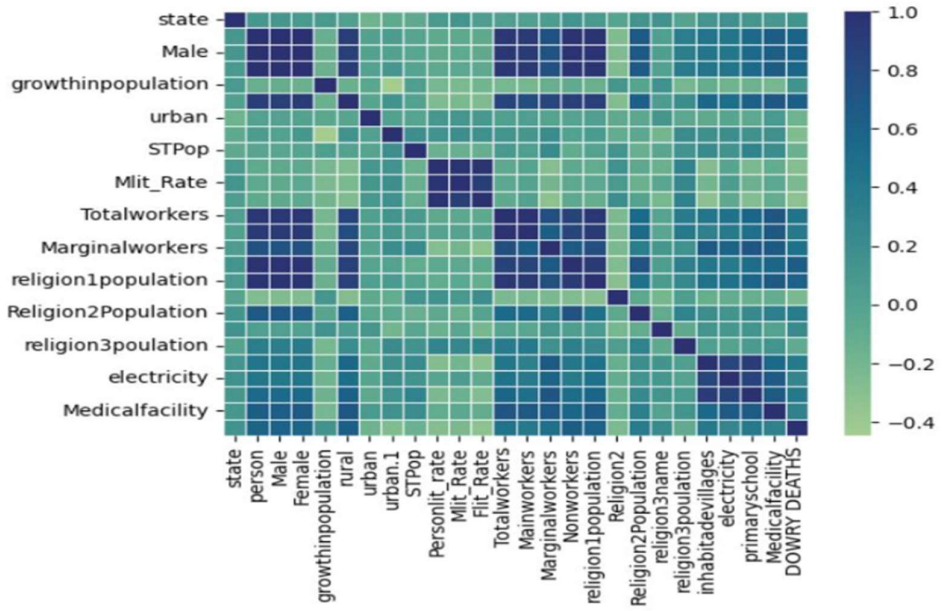

Fig 2.0

The correlation between each column in the census dataset is shown in Fig. 2.0. Many attributes have a negative correlation with one another, and many attributes have a positive correlation with one another.

Correlation table:

	FEMALE LITERACY RATE	MALE LITERACY RATE	PERSONAL LITERACY RATE	NON WORKER	MARGINAL WORKER
DOWARY DEATH	-0.323218	-0.232595	-0.284065	0.643093	0.470345

Table 1.0

According to Table 1.0, there is a positive association between the number of dowry deaths and the number of marginal employees and non-workers, while there is a negative correlation between female literacy rate, male literacy rate and personal literacy rate. We must raise the literacy rates of men, women, and individuals in order to reduce dowry deaths. We also need to reduce the number of unemployed people and low-wage workers. There are other factors that are related to dowry death, although those listed in table 1.0 are more dependent.

PRIOR 2

PREDICTION OF DOWRY DEATH

MACHINE LEARNING

Machine learning (ML) is a high level programming algorithm that allows software applications to become more accurate at predicting outcomes.

We imported all sklearn model in machine learning for analysis.

" " "from sklearn.model_selection import train_test_split

from sklearn.metrics import r2_score

from sklearn.linear_model import HuberRegressor

from sklearn.ensemble import BaggingRegressor

from sklearn.preprocessing import StandardScaler" " "

First split the data into training and testing data.

" " "X_train,X_test = train_test_split(df1_new,test_size=0.3,random_state=42)

y1_train,y1_test = train_test_split(df2_final,test_size=0.3,random_state=42)" " "

Fitted standardscalar for both training and testing data and then did the transform (Standardscalar is used to Standardize features by removing the mean and scaling to unit variance whereas transform perform standardization by centering and scaling.)

" " "std = StandardScaler()

std.fit(X_train)

std.fit(X_test)

X_train = std.transform(X_train)

X_test = std.transform(X_test)" " "

Fit the bagging regressor and Huber regressor model for training data.

“ “ “model =
BaggingRegressor(base_estimator=HuberRegressor(epsilon=1.065,max_iter=3000, alpha=0.00001, warm_start=False),n_estimators=10, random_state=0)

model.fit(X_train,y1_train) ” ” ”

Now using the testing data to predict the dependant variable for the unseen testing data and finding the accuracy of the predicted values using the actual values.

“ “ “pred = model.predict(X_test)

Accuracy = r2_score(y1_test,pred)

round(Accuracy*100,2)” ” ”

The final accuracy is 73.41%.

TESTING OF HYPOTHESIS

H0: Male, Female, Personal literacy rate is directly proportional to dowry death and number of marginal workers and non-workers inversely proportional to dowry death.

HA: : Male ,Female, Personal literacy rate inversely proportional to dowry death and number of marginal workers and non-workers directly proportional to dowry death.

CODE:

```
post_cor = ['Nonworkers','Marginalworkers']
neg_cor = ['Flit_Rate','Mlit_Rate','Personlit_rate']
df_test1 = df1_new.iloc[5:10]
df_test = df_test1.copy()
for i in df_test.columns:
    y=df_test[i].mean()
    z=df_test[i].min()
    if i in post_cor:
        df_test[i] = df_test[i].apply(lambda x:x-z)
    if i in neg_cor:
        df_test[i] = df_test[i].apply(lambda x:x+y)
    else:
        pass
df2_test = df2_final.iloc[5:10]
#Fitting the manipulated data in the trained model
std = StandardScaler()
std.fit(df_test)
df_test = std.transform(df_test)
Test_pred = model.predict(df_test)
Test_pred
#Predicted Value
DISTRICT
5    16
6    11
7     3
8    -1
9    32
Name: DOWRY DEATHS, dtype: int64
#Actual Value
df2_test
DISTRICT
5    36
6    18
7     6
8     4
9    13
Name: DOWRY DEATHS, dtype: int64
```

We attempted to predict the number of dowry deaths that would occur when certain conditions changes, using a machine learning model. In this model, we increased the values of the variables that showed a negative correlation, such as Male literacy rate, female literacy rate and personal literacy rate. We reduced the value of variables that showed a positive correlation, such as the number of non-workers and marginal workers.

For testing, we used data from five districts. For the columns that showed a negative correlation, we added the mean value of the values in the column to each individual value, and for the columns that showed a positive correlation, we subtracted the minimum value of the values in the column from each individual value.

By altering the values of the factors, we were able to reduce the frequency of dowry deaths.

This outcome enables us to defend an accuracy rate of 73.14%.

TABULATION OF RESULT

SL NO	ACTUAL DOWRY DEATH	RESULTANT DOWRY DEATH
1	36	16
2	18	11
3	6	3
4	4	-1
5	13	32

Table 2.0

VISUALIZATION OF RESULT

CODE:

```
mtl.plot(range(1,6),Test_pred,'g',marker='o')
mtl.plot(range(1,6),df2_test.values,'m',marker='o')
mtl.xlabel('INDEX')
mtl.ylabel('DOWRY DEATHS')
mtl.legend(['Predicted Value', 'Actual Value'], loc=3)
```

GRAPH:

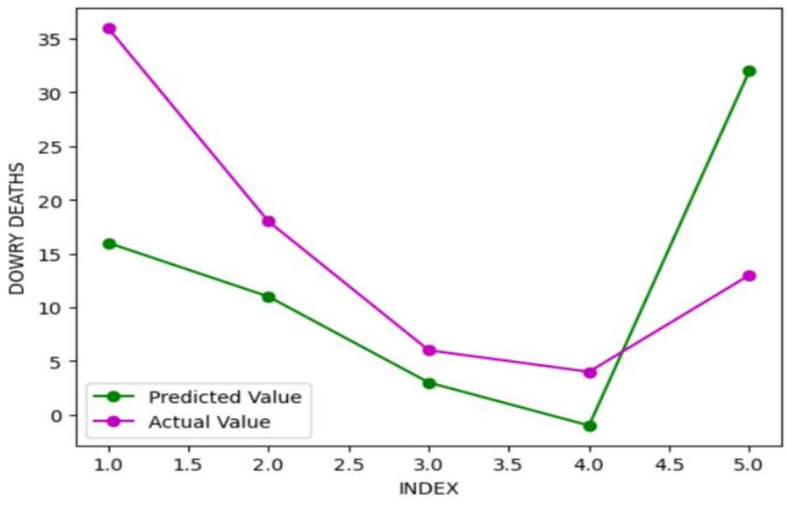

Fig 2.0

We got sufficient evidence to accept alternative hypothesis. That is, if Male, Female, Personal literacy rate increases the dowry death decreasing and number of marginal workers and non-workers decreases the dowry death decreases.

RESULTS

One of the results is from the machine learning model which we have implemented in our crime data. We have used Bagging Regressor along with Huber Regression to get a maximum prediction accuracy of 73.41%. The test which we have conducted by change the data with respect to the correlation it has with the depended data was successful. When the independent variable with positive correlation is reduced and the variable with negative correlation is increased, we can see that the dowry death is also reducing.

Secondly the graph (Fig 2.0) finding demonstrates the type of correlation between the independent and the depended variable in the data and also helps use to visually identify the underlaying relation between the attributes.

CONCLUSION

We may conclude from the statistics and the analysis that there may be measures to lower the rate of dowry deaths. Here, unemployment, rural area, and male and female literacy rates were the key influences on dowry deaths. According to the data, both male and female illiteracy has increased the number of dowry deaths, which makes sense. On the other hand, an increase in non-workers has also been shown to increase dowry deaths. Dowry deaths are also higher in rural areas, which is related to the fact that literacy declines there. Although the number of dowry deaths is declining each year which is depicted in the graph (Fig 1.0) it is evident from the graph (Fig 1.3, Fig 1.4) that there is a positive correlation and in graph (Fig 1.1, Fig 1.2, Fig 1.5) that there is a negative correlation and any modifications made in our favour to these elements will undoubtedly reduce dowry deaths which has been proven in the graph (Fig 2.0) according to our prediction accuracy.

REFERENCES

- Patel, A.(2022 February 21).Crime Against Women Dataset for Dark Spot identification in India.Mandeley Data. https://data.mendeley.com/datasets/whrdh8c5zb/1

- Dang, G.Why Dowry Deaths Have Risen in India?. https://crawford.anu.edu.au/acde/asarc/pdf/papers/2018/WP2018-03.pdf

- Chaudhary,S.(2022 April 15).Dowry and dowry death.Times Of India. https://timesofindia.indiatimes.com/readersblog/politiclaw/dowry-and-dowry-death-42574/