DISABILITY INSURANCE(DI) CLAIM PREDICTION (STATE OF CALIFORNIA)

Name : Ramya Dhabade

Email : rd08289n@pace.edu

GitHub : https://github.com/U174749/Capstone-Project-2023

RESEARCH QUESTIONS

- Can one predict how many claims will be filed in the coming year based on the historical data so that a state department can allocate a certain amount in its budget?

- Can one predict the assigned funds will be sufficient if we increase the weekly benefit amount?

MOTIVATION

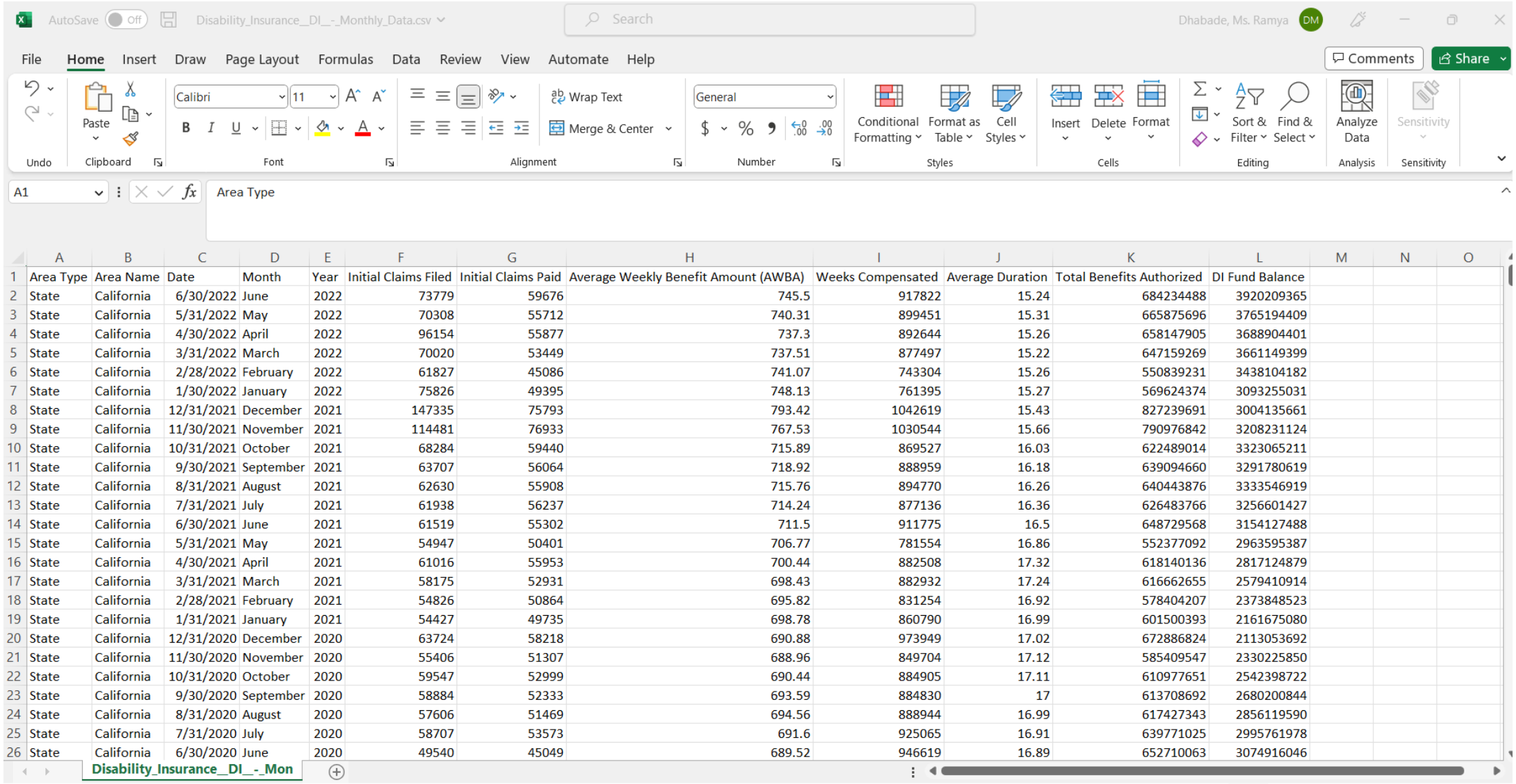
- This project will help me in working with real world problems and solve them using the knowledge gained through the master's program.
- This project will help in learning various factors while working towards it which includes, Complete work on time, Produce high quality work Focus on relevant task, Seek and clarify instructions, Adapt instructions to achieve project needs, Identify and tackle project problems, Pursue needed information, Improve ability to solve problems, Develop improved interpersonal skills, Learn and apply technical skills, Learn and apply professional skills.

DATASET

Dataset:

- The monthly summary report is intended to provide the user with a quick overview of the status of the DI program at the state level. This summary report contains monthly information on claims activities, average weekly benefit amounts, average duration of claims, benefits authorized, the DI Fund balance and other statistics. This data is used in budgetary and administrative planning, program evaluation, and reports to the Legislature and the public.
- The dataset has a claim data for every month for over past 48 years.
- Data consists of 582 rows & 12 columns.
- Link: https://data.ca.gov/dataset/disability-insurance-di-monthly-data

DATASET-SNAPSHOT



Area Type	Area Name	Date	Month	Year	Initial Claims Filed	Initial Claims Paid	Average Weekly Benefit Amount (AWBA)	Weeks Compensated	Average Duration	Total Benefits Authorized	DI Fund Balance
State	California	6/30/2022	June	2022	73779	59676	745.5	917822	15.24	684234488	3920209365
State	California	5/31/2022	May	2022	70308	55712	740.31	899451	15.31	665875696	3765194409
State	California	4/30/2022	April	2022	96154	55877	737.3	892644	15.26	658147905	3688904401
State	California	3/31/2022	March	2022	70020	53449	737.51	877497	15.22	647159269	3661149399
State	California	2/28/2022	February	2022	61827	45086	741.07	743304	15.26	550839231	3438104182
State	California	1/30/2022	January	2022	75826	49395	748.13	761395	15.27	569624374	3093255031
State	California	12/31/2021	December	2021	147335	75793	793.42	1042619	15.43	827239691	3004135661
State	California	11/30/2021	November	2021	114481	76933	767.53	1030544	15.66	790976842	3208231124
State	California	10/31/2021	October	2021	68284	59440	715.89	869527	16.03	622489014	3323065211
State	California	9/30/2021	September	2021	63707	56064	718.92	888959	16.18	639094660	3291780619
State	California	8/31/2021	August	2021	62630	55908	715.76	894770	16.26	640443876	3333546919
State	California	7/31/2021	July	2021	61938	56237	714.24	877136	16.36	626483766	3256601427
State	California	6/30/2021	June	2021	61519	55302	711.5	911775	16.5	648729568	3154127488
State	California	5/31/2021	May	2021	54947	50401	706.77	781554	16.86	552377092	2963595387
State	California	4/30/2021	April	2021	61016	55953	700.44	882508	17.32	618140136	2817124879
State	California	3/31/2021	March	2021	58175	52931	698.43	882932	17.24	616662655	2579410914
State	California	2/28/2021	February	2021	54826	50864	695.82	831254	16.92	578404207	2373848523
State	California	1/31/2021	January	2021	54427	49735	698.78	860790	16.99	601500393	2161675080
State	California	12/31/2020	December	2020	63724	58218	690.88	973949	17.02	672886824	2113053692
State	California	11/30/2020	November	2020	55406	51307	688.96	849704	17.12	585409547	2330225850
State	California	10/31/2020	October	2020	59547	52999	690.44	884905	17.11	610977651	2542398722
State	California	9/30/2020	September	2020	58884	52333	693.59	884830	17	613708692	2680200844
State	California	8/31/2020	August	2020	57606	51469	694.56	888944	16.99	617427343	2856119590
State	California	7/31/2020	July	2020	58707	53573	691.6	925065	16.91	639771025	2995761978
State	California	6/30/2020	June	2020	49540	45049	689.52	946619	16.89	652710063	3074916046

LITERATURE REVIEW

The comparison of the result with Human prediction clearly shows that ANN outperforms in prediction. It reduced the error by 11.5%. It can be concluded that ANN can be used for medical claims prediction resulting in strong forecasting.

The gap between the allocated budget and realized expenditure in NDIS can be closed faster and at a reduced cost using an appropriate machine learning model compared to the current manual processes.

A comparison of 2013 to 2020 indicates reducing profits, premiums, and assets but increasing claims. However, comparison and forecast predicts a normalization of economic indicators from January 2021.

In the logistic regression, the factors most strongly associated with exhaustion of STDI benefits are age, diagnosis, and employer industry. Waiting to allow some claims to resolve without intervention improves the efficiency of targeting efforts.

To identify people with disability in claim data predictive models provide an improved tool when multiple claim-based indicators are considered. For age 65 or older sensitivity–specificity trade-off for the models is considerably better than for those ages 18-64.

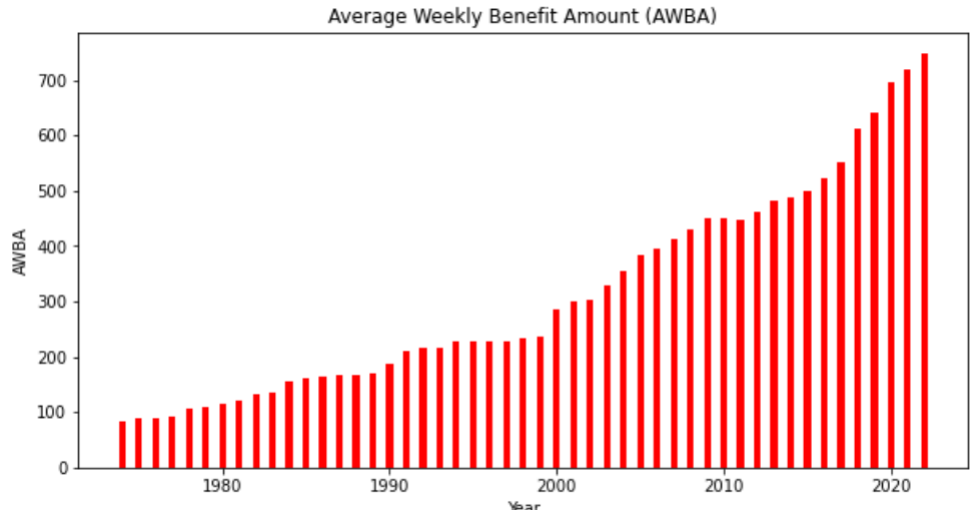
DATASET VARIABLES

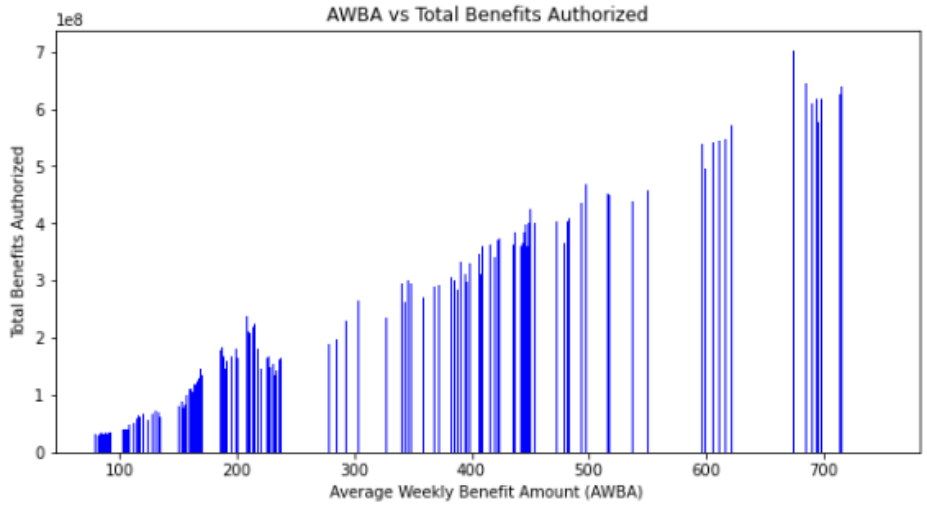
- Area Type : State Information is provided in this field
- Area Name : Name of the State is provided (California)
- Date : Date in DD/MM/YYYY format
- Month : Month of the year in categorical values.
- Year : 1974 - 2022
- Initial Claims Filed : Number of claims filed that year for the date.
- Initial Claims Paid : Number of claims paid that year for the date.
- Average Weekly Benefit Amount (AWBA) : The average amount that is paid on a weekly basis.
- Weeks Compensated : Number of weeks compensated as per the date.
- Average Duration : Average Duration calculated.
- Total Benefits Authorized : Amount that has been authorized so far.
- DI Fund Balance : Amount that is available in Disability Insurance Funds.

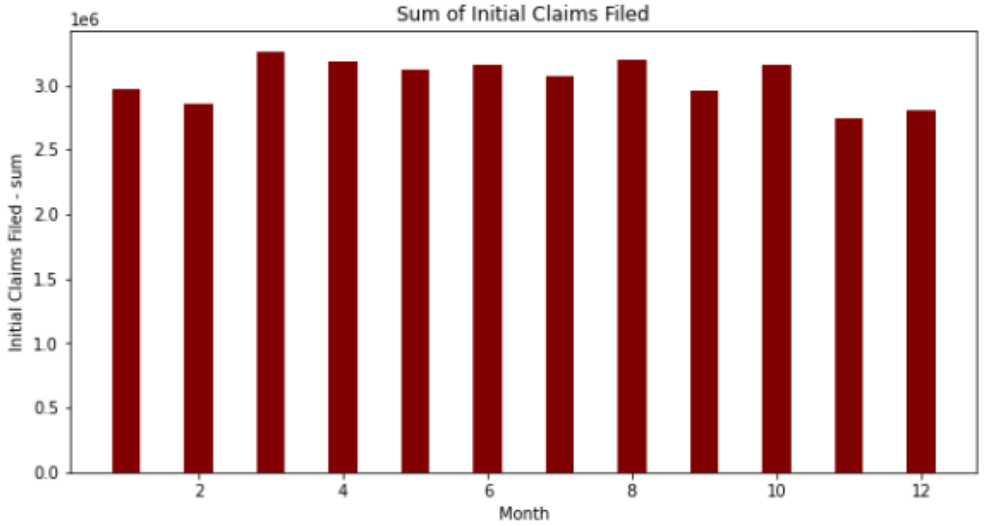
DATA PREPROCESSING

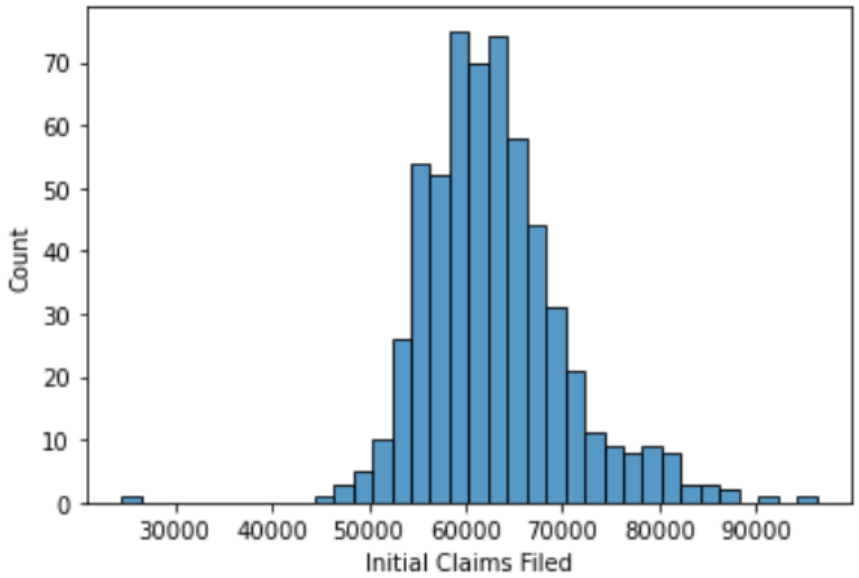
- Loaded the dataset into JNB in CSV format.
- Data consists of both categorical and numerical data.
- Found 12 null values in Average duration column.
- Removed the Nan values and replaced it with Mean values for Average Duration Data.
- Dropped 3 columns area type, area name & date as they are not applicable for the analysis.
- Replaced the month string values with the numeric values for better analysis.
- No duplicates found.

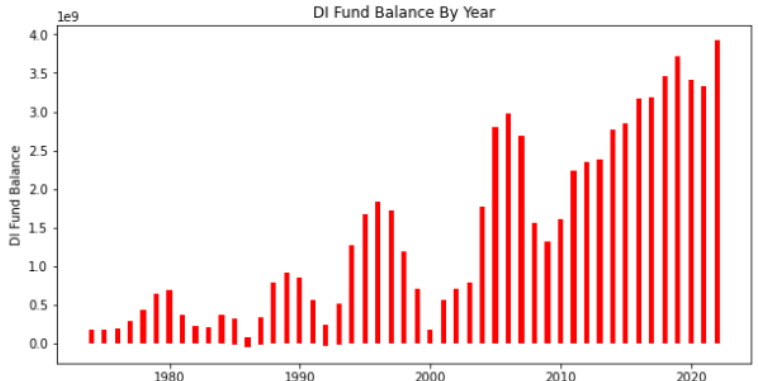
EDA


Average Weekly Benefit Amount (AWBA)
AWBA
Year


AWBA vs Total Benefits Authorized
1e8
Total Benefits Authorized
Average Weekly Benefit Amount (AWBA)


Sum of Initial Claims Filed
1e6
Initial Claims Filed - sum
Month


Count
Initial Claims Filed


DI Fund Balance By Year
1e9
DI Fund Balance

INSIGHTS

- There is a continuous increase in the Average Weekly Benefit Amount (AWBA) every year.
- Benefit paid have proportionally increased with an increase of a weekly benefit amount from year 1974 to 2022.
- We have observed that March & August are the top two months where maximum claims were filed.
- On an average 50k to 70k claims are filed every month.
- DI fund balance is constantly increasing from year 2010 to 2020

Derived evaluation metrics for different regression algorithms on a Disability claim Insurance dataset.

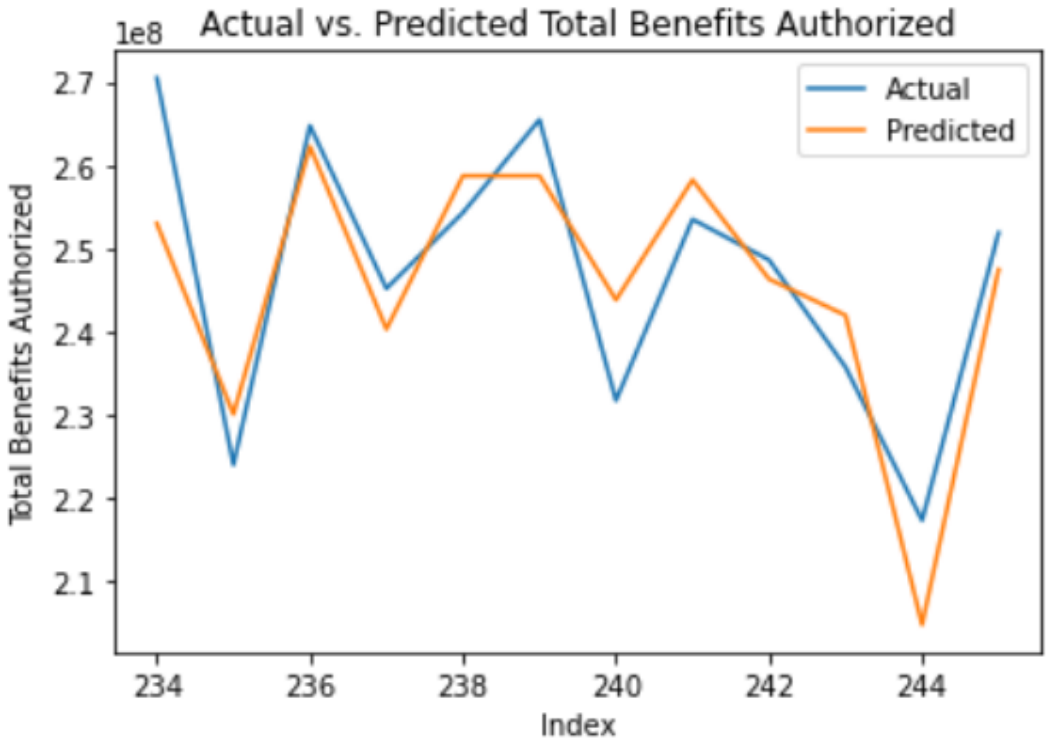
To compare the performance of different regression algorithms on this dataset, we have used below five regression models.

- The **linear regression model** has the highest MSE and the lowest R-squared, indicating that it does not fit the data well.
- The **decision tree regression** has a lower MSE and a slightly higher R-squared than linear regression, indicating a better fit than linear regression.
- The **random forest regression** has a lower MSE and a higher R-squared than decision tree regression, indicating a better fit than decision tree regression.
- The **gradient boosting regression** has the lowest MSE and the highest R-squared, indicating the best fit among all the models.
- The **support vector regression** has a high MSE and a negative R-squared, indicating that it performs worse than the other models and does not fit the data at all.

METHODOLOGY

- Based on the results , the mean squared error (MSE) of the gradient boosting model for the Insurance Disability Claim dataset is 14,241,186.19 and the R-squared value is 0.72.
- The MSE measures the average squared difference between the predicted values and the actual values, and a lower value indicates a better fit of the model to the data. In this case, the MSE of 14,241,186.19 suggests that the model has a relatively low error rate in predicting the target variable.
- The R-squared value represents the proportion of variance in the target variable. A higher R-squared value indicates a better fit of the model to the data, and a value of 0.72 suggests that the model explains 72% of the variance in the target variable, which is a good fit.
- Overall, these results suggest that the gradient boosting model is effective in predicting the target variable in the Insurance Disability Claim dataset.

RESULTS


Actual vs. Predicted Total Benefits Authorized
1e8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
Total Benefits Authorized
234
236
238
240
242
244
Index
Actual
Predicted