# PROJECT REPORT

## Title : Optimizing Predictive Modeling

---

# OVER VIEW:

**This public data is about the medicare Department on how medicare prays for medical services.This data is prepared by Centers for Medicare & Medicaid Services(CMS). This data includes info about on what services and procedures doctor and other health care providers have given to this medicare patients and the finalised claims for services provided outside of the hospital and does not cover any things during medical equipment.**

**In the physician and other Practitioners data comes from the medicare claim records.These records are for the medicare patients who are part of the fee for sevice program where medicare pays for each service individually.These medical services are organized by National Provider Identifier.**

# ABOUT THE DATA SET:

**This dataset consists of 73 columns.**

---

- **This dataset gives the total information of the provider including name,gender,address .**
- **It also gives the information of the patients from various age groups**
- **This dataset tells about the services and utilizations**
- **This data also give the information about the utilizations by HCPCS code (I.e -Place of the service)**
- **It also gives the info whether the Service is a drug**
- **This dataset has the information in detail about the Service counts,Beneficiary counts,Provider charges,Medicare allowed amounts and their payments and Place of the Servive Indicator.**
- **This data has also informations of patients such as Gender,race and Health conditions.**

---

# TARGET COLUMN:

**In this case I choose Tagert column as "Tot_Mdcr_Payment_Amt",which defines about the total payment paid by the medicare for there respective services.By this we can understand and predict medicare payment amounts can help in optimizing the health care costs and identifying areas for cost savings so this is directly tied to financial outcomes,making it crucial for budgeting and financial planning in healthcare systems.So by analyzing and predicting these amounts,healthcare providers and policymakers can devise strategies to reduce unneccesary spending and improve efficiency.**

# EXPLORATORY DATA ANALYSIS(EDA):

**Before model building doing the basic EDA is necessary for the better results.some basic EDA steps should be followed before model building are:**

---

- **Knowing all the datatypes of all columns i.e (numerical,categorical,strings etc..)**
- **So understanding the data in form of numbers we have to know about the statistical information of the data set such as mean,median,mode,maximum,minimum,variance,standard deviation etc..**
- **Handling the missing values by mean,median and knn imputation methods.**

- Detecting outliers and duplicate values
- Data Visualizations based on the target column.

---

# MACHINE LEARING MODEL(ML):

**Based on my target column Iam building the regression model**
**Here are the steps to be followed:**

---

- Importing all the necessary Libraries
- Loading the data
- Splitting the data into Train and Test
- Doing all the Pre-processing and data transfromations for fitting the data. For numerical columns we have to perform the standardization and for categorical columns we have to use One hot encoding or Label encoding
- Building the model using ML algorithms (I trained my model for SGDRegressor and XgBoostegressor
- Making the predictions on respective models
- Evaluating the model with the respective metrics
- The model which gives the better results on the test data is considered as the suitable model for the data

---

# ARTIFICIAL NEURAL NETWORK(ANN):

**Here are the steps to be followed:**

---

- Import Libraries: TensorFlow/Keras for building the ANN, and others for data manipulation.
- Load and Prepare Data: Load your dataset, handle missing values, and scale features.
- Build the Model: Define the architecture of the ANN.
- Compile the Model: Specify optimizer, loss function, and metrics.
- Train the Model: Fit the model to the training data.
- Evaluate the Model: Assess performance on test data.
- Save and Load the Model: Save the trained model for future use.

---

# OBERVATIONS:

**Performance of SGDRegressor by metrics like Mean Sqaure Error and R^2 score**

---

- Mean square Error : **1.4165441298742975e+20**
- R^2 score : **-227590258.5704113**

---

**Performance of XgBoost Regressor by metrics like Mean Sqaure Error and R^2 score**

---

- Mean sqaure Error : **352018173293.1132**
- R^2 score : **0.4344270274136036**

---

**From this two models XgBoost Regressor will be best model because it has high R^2 score.**

# ANN USING DIFFERENT OPTIMIZERS

1. Adam
2. RMSprop
3. Adagrad
4. SGD

1.Training the model with different hyperparameter gives accurate results
2.Use Dropout and BatchNormalzation for better outcomes
3.Train the model with different learning rates and drop out rates to compare which gives the best result

L1= 0.001 D1 = 0.3 L= learning rate

L2= 0.01 D2 = 0.5 D = dropout rate

| Optimizers | MSE with L1&D1 | MSE with L1&D2 | MSE with L2&D1 | MSE with L2&D2 | R^2 with L1&D1 | R^2 with L1&D2 | R^2 with L2&D1 | R^2 with L2&D2 |
|---|---|---|---|---|---|---|---|---|
| Adam | 63550731 1431.8636 | 63550661 8118.8811 | 59841283 6805.9395 | 58774545 3967.5618 | -0.02104 3191788 926663 | -0.02104 2077871 76504 | 0.038554 9593077 9396 | 0.055693 8000817 7516 |
| RMSprop | 63571129 6304.2091 | 63573389 8080.2516 | 57211247 5438.3943 | 57434579 1388.9279 | -0.02137 0925807 70128 | -0.02140 7239142 159362 | 0.080810 6571303 9837 | 0.077222 4811875 1562 |
| Adgrad | 63606090 3842.3008 | 63606121 6741.227 | 63601851 2075.335 | 63601749 8580.7992 | -0.02193 2625712 872245 | -0.02193 3128434 576066 | -0.02186 4516622 27662 | -0.02186 2888282 716986 |
| SGD | 62320530 6272.9775 | 62577990 1611.4836 | 30743635 1801.6576 5 | 34268092 7646.0364 | -0.00127 8071251 5534365 | -0.00541 4566606 847027 | 0.506054 7876180 281 | 0.449428 7920298 5675 |

1.Adam

Best MSE: 587745453967.5618 (L2&D2) Best R^2 : 0.05569380008177516 (L2&D2)

2.RMSprop

Best MSE: 572112475438.3943 (L2&D1) Best R^2 : 0.08081065713039837 (L2&D1)

3. Adagrad

Best MSE: 636017498580.7992 (L2&D2) Best R^2 : -0.021862888282716986 (L2&D2)

4.SGD

Best MSE:307436351801.65765 (L2&D1) Best R^2 : 0.5060547876180281 (L2&D1)

# ML VS ANN

**By comapring all the ML algorithms and the Different Optimizers of the ANN we can conclude that :**

- **Best MSE: 307436351801.65765 at model ANN (SGD)**
- **Best R-squared: 0.5060547876180281 at model ANN (SGD)**
- **Best Optimizer based on MSE: SGD**
- **Best Optimizer based on R-squared: SGD**

# CONCLUSION

**In analyzing Medicare payment amounts, predicting these values accurately holds significant importance for financial and operational efficiency in healthcare systems. The focus on the target column, which represents Medicare payment amounts, underscores its relevance for cost management and budgeting in healthcare.**

**Based on the results, the best performance was achieved with the ANN model using the SGD optimizer. Specifically:**

- **Best Mean Squared Error (MSE): 307,436,351,801.66307,436,351,801.66307,436,351,801.66**
- **Best R-squared Score ($R^2$): 0.510.510.51**

**The SGD optimizer outperformed others in both metrics, highlighting its effectiveness in minimizing prediction errors and improving model accuracy for this target. This suggests that, for the given dataset, the SGD optimizer is the most suitable choice for models aiming to predict Medicare payment amounts. Overall, these insights can guide healthcare providers and policymakers in leveraging predictive analytics to better manage financial resources, optimize costs, and enhance decision-making processes.**