# Hyperparameter Tuning-DB13

- Nikita Gawde
- Ira Pantbalekundri
- Purvang Jayesh Thakkar

# Abstract

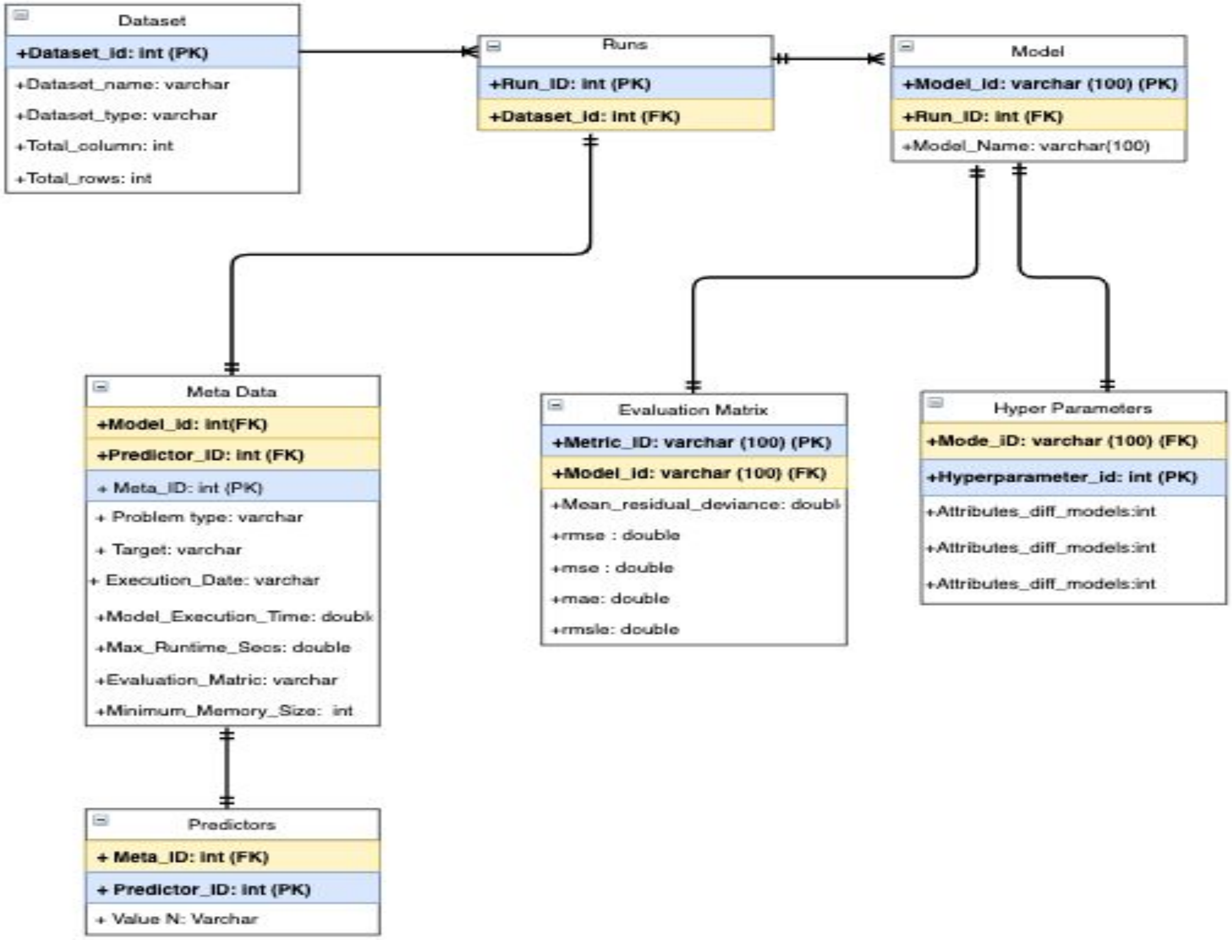
- The goal of this project is to provide a database which will store all the hyperparameters for a particular model for a given dataset.
- The hyperparameter database is a public resource with algorithms, tools, and data that allows users to visualize and understand how to choose hyperparameters that maximize the predictive power of their models.
- The hyperparameter database is created by running millions of hyperparameter values, over thousands of public datasets and calculating the individual conditional expectation of every hyperparameter on the quality of a model.
- The hyperparameter database also uses these data to build models that can predict hyperparameters without search and for visualizing and teaching statistical concepts such as power and bias/variance tradeoff.

# Dataset 1: Predicting Mortality Rate for Cancer

- The dataset was obtained from Dataworld and aggregated from multiple sources including American Community Service, cancer.org.
- The goal of the dataset is to determine the cancer mortality rate by using multiple regression models such as GBM, Deep Learning, Stacked Ensembles, DRF, etc.
- Our objective is to store the JSON files and analyse.
- The mortality rate is estimated using different variables of the dataset as predictors.
- These predictors are stored in metadata.

# Conceptual Model

# CONCEPTUAL DIAGRAM



Dataset

+Dataset_id: int (PK)

+Dataset_name: varchar

+Dataset_type: varchar

+Total_column: int

+Total_rows: int

Runs

+Run_ID: int (PK)

+Dataset_id: int (FK)

Model

+Model_id: varchar (100) (PK)

+Run_ID: int (FK)

+Model_Name: varchar(100)

Meta Data

+Model_id: int(FK)

+Predictor_ID: int (FK)

+ Meta_ID: int (PK)

+ Problem type: varchar

+ Target: varchar

+ Execution_Date: varchar

+Model_Execution_Time: doubl

+Max_Runtime_Secs: double

+Evaluation_Matric: varchar

+Minimum_Memory_Size: int

Evaluation Matrix

+Metric_ID: varchar (100) (PK)

+Model_id: varchar (100) (FK)

+Mean_residual_deviance: doubl

+rmse : double

+mse : double

+mae: double

+rmsle: double

Hyper Parameters

+Mode_iD: varchar (100) (FK)

+Hyperparameter_id: int (PK)

+Attributes_diff_models:int

+Attributes_diff_models:int

+Attributes_diff_models:int

Predictors

+ Meta_ID: int (FK)

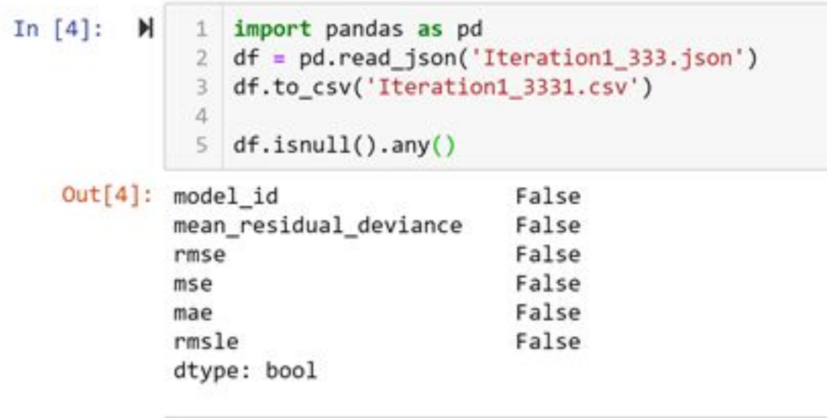
+ Predictor_ID: int (PK)

+ Value N: Varchar

# Data Processing

- The Data Science team had curated the dataset and removed all of the outliers and null values
- The iterated data was acquired in JSON format which is converted to csv files for easy processing in MySQL workbench.
- We received the JSON files for: Runs, Evaluation Matrix, Hyperparameters for every model, the predictors and the metadata file.
- We are storing evaluation metrics for each and every model and every
- run of it.
- We have normalized the Dataset upto 3NF which supports referential integrity ie every table is linked to the others via keys.

# Data Preprocessing:Checking null values



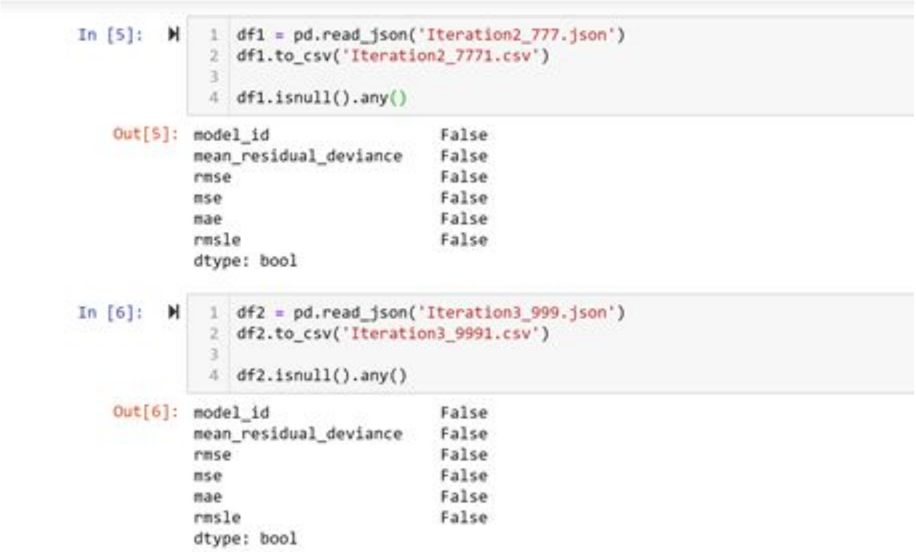
In [4]:

```
import pandas as pd
df = pd.read_json('Iteration1_333.json')
df.to_csv('Iteration1_3331.csv')

df.isnull().any()
```

Out[4]:

```
model_id                  False
mean_residual_deviance    False
rmse                      False
mse                       False
mae                       False
rmsle                     False
dtype: bool
```



In [5]:

```
df1 = pd.read_json('Iteration2_777.json')
df1.to_csv('Iteration2_7771.csv')

df1.isnull().any()
```

Out[5]:

```
model_id                  False
mean_residual_deviance    False
rmse                      False
mse                       False
mae                       False
rmsle                     False
dtype: bool
```

In [6]:

```
df2 = pd.read_json('Iteration3_999.json')
df2.to_csv('Iteration3_9991.csv')

df2.isnull().any()
```

Out[6]:

```
model_id                  False
mean_residual_deviance    False
rmse                      False
mse                       False
mae                       False
rmsle                     False
dtype: bool
```

# What ‘s left ?

- Generating the CSV files for all the JSON files left out.
- SQL Use cases which determine the best hyperparameters for a particular model .
- Stored Procedures, Functions and Indexes which will reduce the execution time and be essential for querying the hyperparameters.
- Documentation for the entire process.

# THANK YOU :)