Part - I

Dataset:

- Dataset used here from Data.Gov - https://catalog.data.gov/dataset/public-school-characteristics-2020-21
- The dataset is about Public School characteristics for 2020-2021
- Contains data about Male, Female enrollment, students from different ethnicities, student-teacher ratio, grade levels offered, mode of teaching etc.
- 100722 rows and 79 columns in the original dataset
- Aim to predict the type of school based on all the information available
- Dataset contains categorical values, continuous and geo-location data

Missing Values

- Dataset does contain missing values in several fields
- Some are dropped, as we have sufficient data to do so and we are unsure what missing data mean - like STUTERATIO - student teacher ratio
- For columns describing the demographics, it makes sense to replace NAs with zeros

Main Statistics

- Missing values are in demographics columns and student-teacher ratio column

```
OBJECTID               0
STABR                  0
LEAID                  0
SCH_NAME               0
CHARTER_TEXT           0
MAGNET_TEXT            0
VIRTUAL                0
GSLO                   0
GSHI                   0
FRELCH             26605
REDLCH             26605
SCHOOL_LEVEL           0
TITLEI                 0
STITLEI                0
STATUS                 0
SCHOOL_TYPE_TEXT       0
SY_STATUS_TEXT         0
ULOCALE                0
STUTERATIO          1216
AMALM              31433
AMALF              31687
ASALM              17784
ASALF              18097
BLALM              11599
BLALF              12288
HPALM              40038
HPALF              40567
HIALM               4686
HIALF               4907
TRALM               8943
TRALF               9128
WHALM               4354
WHALF               4607
dtype: int64
```

- Continuous values statistics (demographics)
    - Average students enrolled is WH - White category (M and F) - this makes sense because of the population
    - Second Highest is - HI - American Indian/Alaska Native category
    - Least enrolment is from the HP - Hawaiian/Pacific Islander category
    - Second minimum is AM - Asian American Category

```
df[norm_cols].describe().transpose()
```

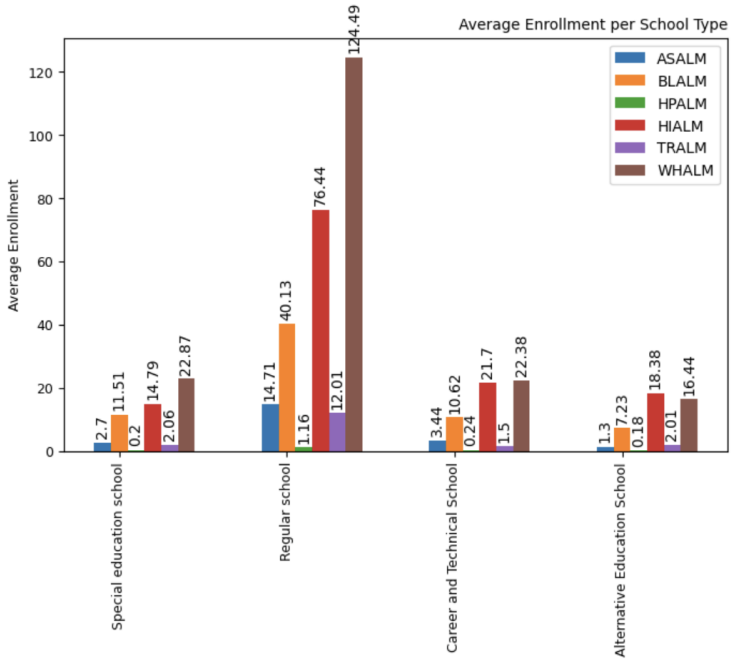
	count	mean	std	min	25%	50%	75%	max
AMALM	100722.0	2.496297	12.849060	0.0	0.0	0.0	1.0	618.0
AMALF	100722.0	2.398682	12.500261	0.0	0.0	0.0	1.0	690.0
ASALM	100722.0	13.560027	42.658770	0.0	0.0	2.0	9.0	2044.0
ASALF	100722.0	12.859455	40.235866	0.0	0.0	2.0	8.0	1473.0
BLALM	100722.0	37.284357	73.538664	0.0	1.0	7.0	40.0	2619.0
BLALF	100722.0	35.932934	72.888833	0.0	1.0	6.0	38.0	2741.0
HPALM	100722.0	1.075574	9.966851	0.0	0.0	0.0	0.0	689.0
HPALF	100722.0	1.011080	9.179666	0.0	0.0	0.0	0.0	612.0
HIALM	100722.0	71.161752	122.174094	0.0	5.0	24.0	87.0	2335.0
HIALF	100722.0	68.108199	117.958973	0.0	5.0	23.0	82.0	2214.0
TRALM	100722.0	11.097188	17.071544	0.0	1.0	6.0	15.0	1790.0
TRALF	100722.0	10.760867	17.271579	0.0	1.0	6.0	15.0	1748.0
WHALM	100722.0	114.918201	140.377863	0.0	16.0	79.0	164.0	6887.0
WHALF	100722.0	107.805812	136.099826	0.0	13.0	73.0	154.0	7410.0

- Stats of categorical columns
  - # of unique schools is 88788
  - Most of the schools are in Suburban Area
  - Most of the schools are in operational status
  - Most Schools offer classes - Not Virtual

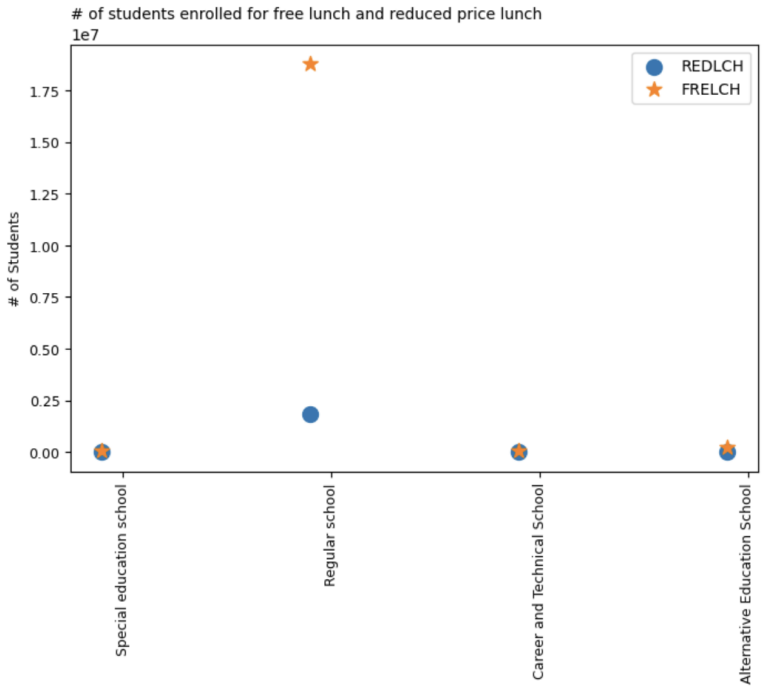
	count	unique	top	freq
GSHI	100722	19	05	28323
GSLO	100722	18	PK	32063
ULOCALE	100722	12	21-Suburb: Large	26452
SY_STATUS_TEXT	100722	7	Currently operational	98717
SCHOOL_TYPE_TEXT	100722	4	Regular school	91595
STITLEI	100722	4	1-Yes	50240
TITLEI	100722	4	1-Yes	60044
SCHOOL_LEVEL	100722	10	Elementary	53024
VIRTUAL	100722	6	Not Virtual	49156
MAGNET_TEXT	100722	4	No	62768
CHARTER_TEXT	100722	3	No	88489
SCH_NAME	100722	88788	Lincoln Elementary School	70
STABR	100722	57	CA	10330

Visuals:

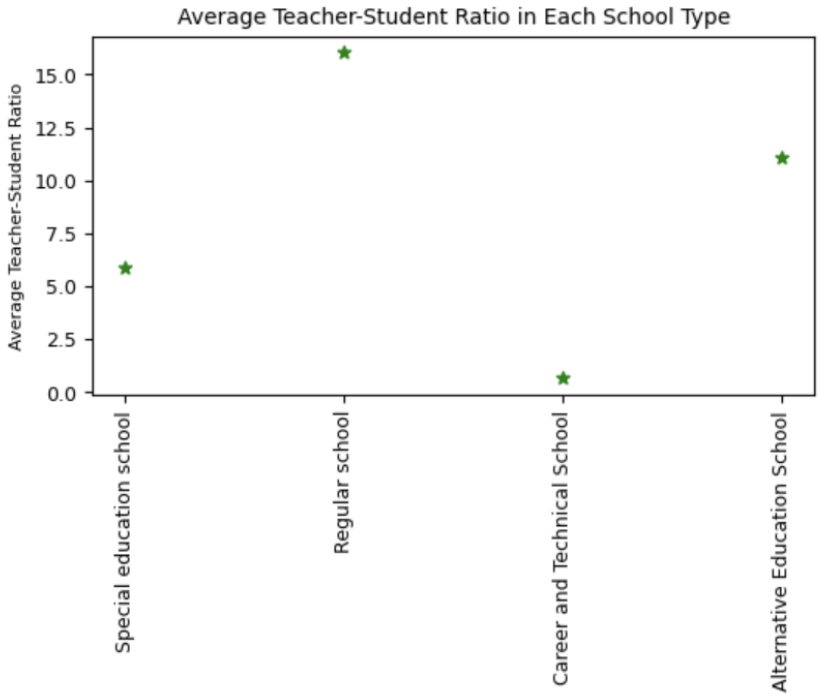
1. Average Enrollment from different ethnicities in different school types - most of the population goes to Regular School. And majority of them belong to White Category and second highest are American Indians



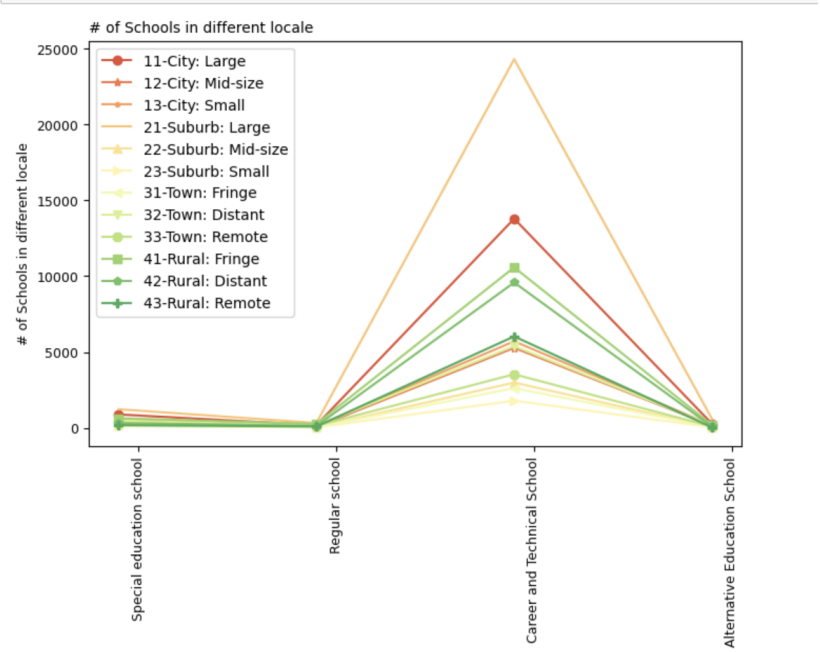
2. Students enrolled for free lunch and reduced price lunch
    a. In Regular schools, a large number of students are enrolled in free lunch programs.
    b. In other schools, average number of students enrolled to free lunch and reduced price lunch programs are comparable



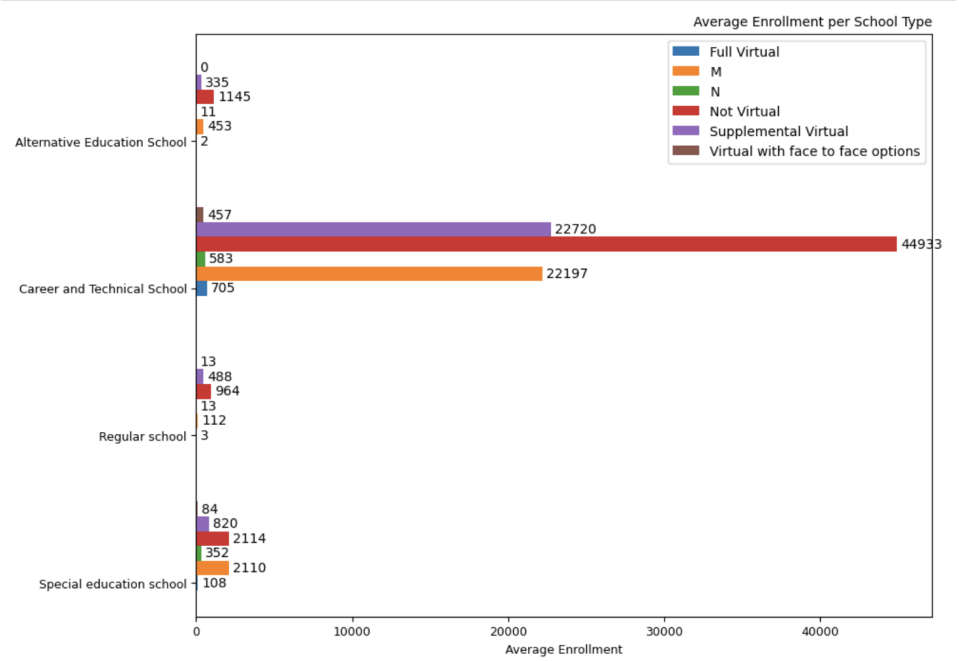
3. Average Student-teacher ratio in each school type
    - We can see highest student-teacher ratio 1:16 in Regular schools
    - Lowest in career and technical school 1:2, which is good for student's growth
    - Alternative Education School has about 1:10
    - Special Education Schools have 1:5



4. Number of schools in different locale
    a. Most of the schools are Suburban-Large and Cities-Large and Rural Fringe areas
    b. Least of them are in Suburban-Small and Town-Fringe area



5. Average Enrollment per School Type - Most of the students are enrolled to non-virtual classes and supplemental mode of teaching.



Part - 2

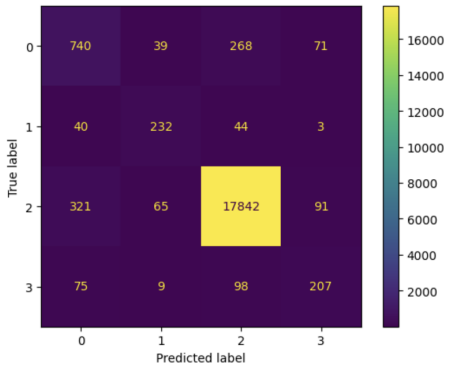
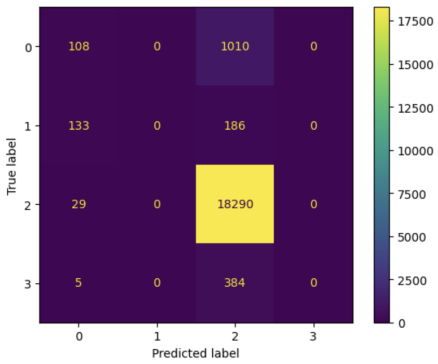
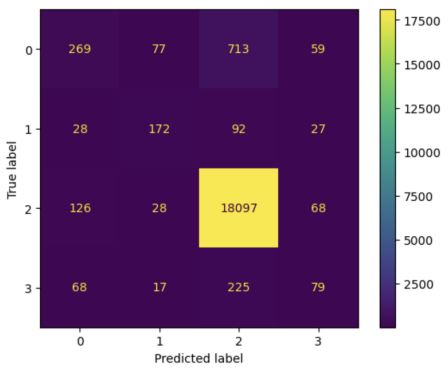
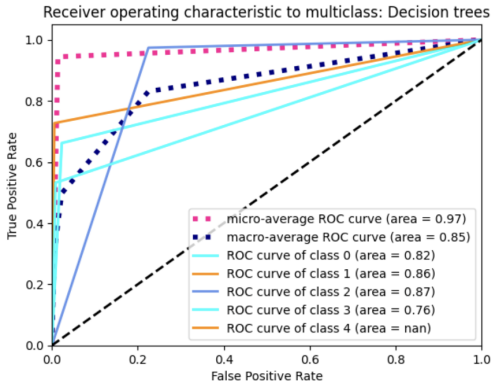
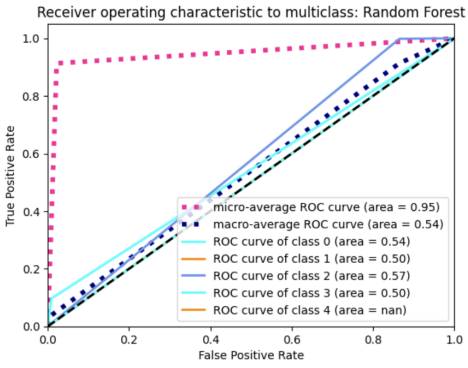
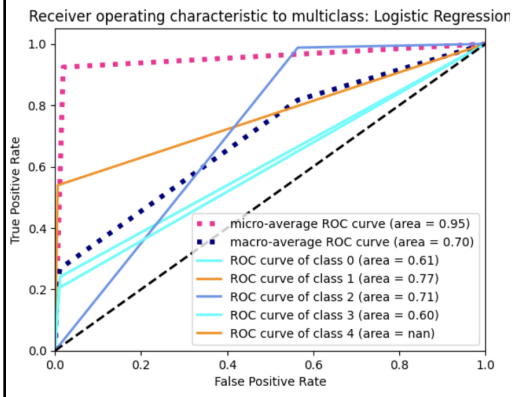
School Type is the target

1. Models used:
    1. Decision Trees:
        a. Partitions the data into subsets based on the values of different features and recursively builds a tree-like structure of decisions
        b. Nodes are based on maximizing the information, goal here is to minimize the number of nodes, leaves give out the outcome (class A or B)
        c. Splits happen based on the data it sees not on the universal data
        d. Advantages: Faster to train, easy to interpret, can handle both numerical and categorical data
        e. Disadvantages: Bigger trees are need to capture complex relations between the data, this can lead to overfitting, sensitive to small changes in the data

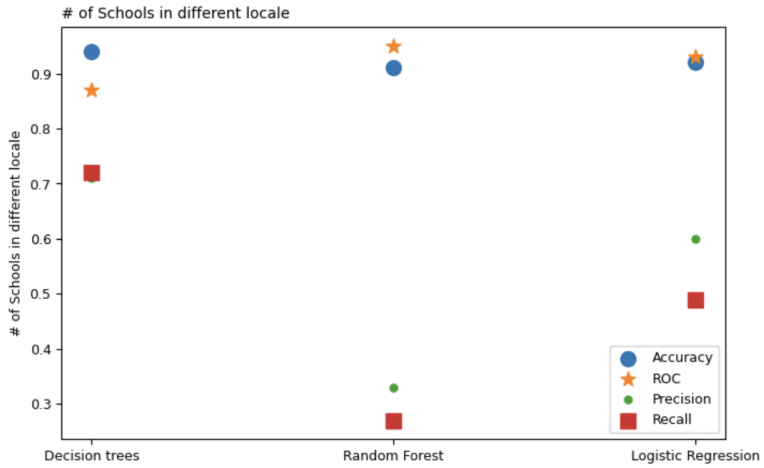
    2. Logistic Regression
        a. Uses Regression to predict the classes
        b. It gives the probability of belonging to a particular class
            i. Linear regression: y = ax + b
            ii. Logistic function ensures the result is between 0 and 1
                1. $y = e^{(ax+b)} / (1 + e^{(ax+b)})$
                2. $log(y / (1 - y)) = e^{(ax+b)}$
                3. Maximize the likelihood: $l(a, b) = y * (1 - y)$
        c. Advantages - easy to interpret, gives probabilistic outcome and computational efficiency
        d. Disadvantages - outcome depends on the data and sensitive to bias

3. Random Forest
    a. Random Forest - bagging technique, creates an ensemble of independent decision trees. First trains each tree on separate samples in the data.
    b. For each tree, at each split, randomly select a set of predictors, choose an optimal predictor and optimal threshold for the split
    c. Advantages: Reduces the risk of overfitting, can handle numerical and categorical features (as Decision trees) added advantage is that it provides feature importance
    d. Disadvantages: not as interpretable as a decision tree, could be computationally expensive

2. Classification problem: ROC and confusion Matrix

Decision Trees	Random Forest	Logistic Regression
		
		
One-vs-One ROC AUC scores (dt): 0.827196 (macro) One-vs-Rest ROC AUC scores (dt): 0.842456 (macro)	One-vs-One ROC AUC scores (rf): 0.786522 (macro) One-vs-Rest ROC AUC scores (rf): 0.944237 (macro)	One-vs-One ROC AUC scores (lr): 0.857768 (macro) One-vs-Rest ROC AUC scores (lr): 0.937483 (macro)

Model Performance Report				
Models	Acc	Prec	Rec	ROC
Decision Trees	0.94	0.71	0.72	0.87
Random Forest	0.91	0.33	0.27	0.95
Linear Regression	0.92	0.60	0.49	0.93



3. Results and comparison:

a. Decision Trees: with not only good accuracy score but also higher precision and recall.
b. Random Forest: Equally good accuracy, but lower precision and recall
c. Logistic Regression: Good accuracy, but precision and recall lower than decision trees but better than random forest.
d. So, Decision trees with best performance, then Logistic regression and then Random Forest.
e. ROC - Receiver Operating Characteristics: plots true-positive rate and false-positive rate, helps visualize trade-off between TP rate and FP rate.
    i. Because we are working with a multi-class classification problem, we are plotting separate line for each class
    ii. Decision trees - all the lines are closer to top-left corner indicating better performance
    iii. Random forest - only 1 line close to top-left, indicating this model is predicting only one-class or biased.
    iv. Logistic Regression - Better than Random Forest with curves closer to left edge
f. AUC - represents area under the ROC curve, closer to 1, better
    i. Decision Tree's OVO and OVR scores are comparable
    ii. Whereas the same for Random Forest and Logistic Regression are quite far part

PART - 3

1. Results of the Neural Network with own setup

Hyper-parameters	Original Setup	Setup-1	Setup-2	Setup-3
Optimizer	SGD	Adam	SGD	SGD
Epochs	2	5	5	5
Learning rate	0.0001	0.001	0.001	0.001
Linear layers	3	3	3	4
Accuracy	36%	54%	61%	59%

   a. There is an increase in accuracy from Adam Optimizer to SGD optimizer (Setup-1 and Setup-2)
   b. Decrease in accuracy with 4 linear layers than 3 (Setup-2 and Setup-3)
   c. But with increased number of Epochs, and reduced learning rate, there is higher accuracy (Original Setup and setup-1)

2. Neural Network for Public School data -
   a. Linear layer (with 64 output features)
   b. Lineary layer (32 input and 4 output features)

```
Net(
  (ll1): Linear(in_features=26, out_features=64, bias=True)
  (ll3): Linear(in_features=32, out_features=4, bias=True)
  (pool): MaxPool1d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
)
```

   c. Max Pooling layer with kernel size of 2
   d. Accuracy: 0.916
   e. Precision: 0.715
   f. Recall: 0.492
   g. ROC - even though micro-average, accuracy shows good results, individual class performance is not as good as the performance of Decision Trees (with the current Network setup)