Task: Predict test scores of students

Target: Predicting the posttest scores of students from 11 features

Solution Workflow:

Data Loading -> Preprocessing -> Feature Engineering -> Model Training -> Model Validation -> Model Inference

Solution Approach:

1. Manual Way of building models and validating it
2. Automated model building and validation

In Scope:

1. Preprocessing
2. Feature Engineering (Certain degree)
3. Model training
4. Model Validation
5. Model Inference

Out of Scope:

1. Model Monitoring
2. Model Deployment
3. Model Explainability
4. Model Interpretability
5. Model Versioning
6. Model Retraining
7. Model Optimisation

Other details:

Timeline: 3 Hours (Approx)
Date of Submission: 20th July 2021
Worked by: Lakshmikanth Rajamani

Insights from the task:

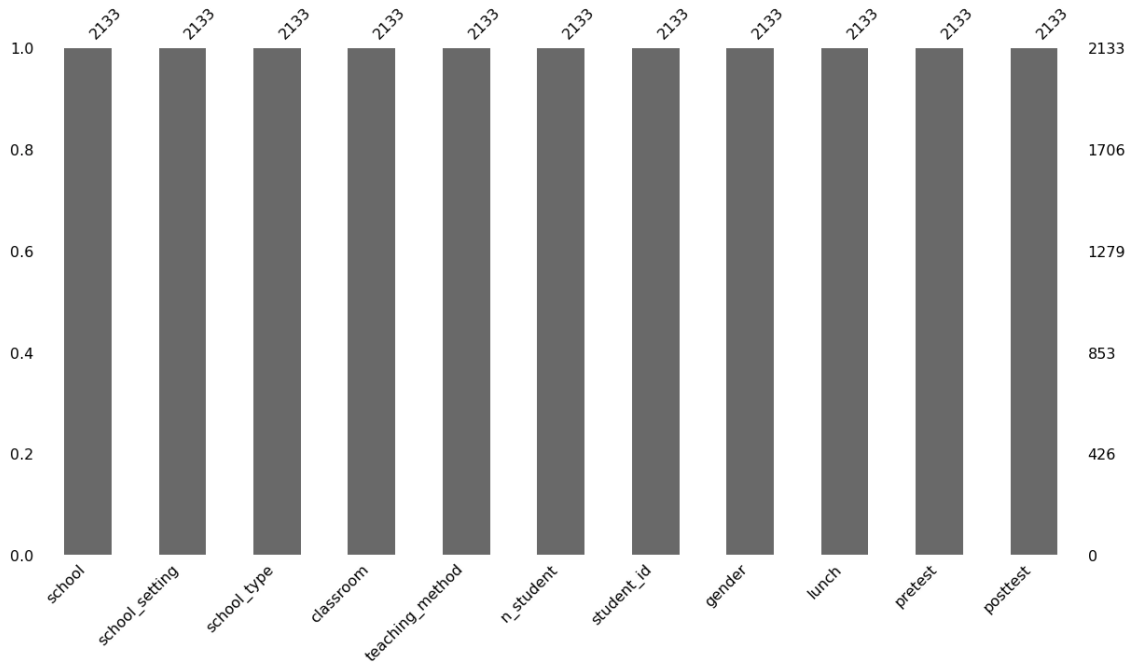
1. There are more public schools available than non-public schools

```
Public        1582
Non-public     551
Name: school_type, dtype: int64
```

2. Out of 11 features - only 3 are numeric and the rest are categorical

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2133 entries, 0 to 2132
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   school           2133 non-null   object
 1   school_setting   2133 non-null   object
 2   school_type      2133 non-null   object
 3   classroom        2133 non-null   object
 4   teaching_method  2133 non-null   object
 5   n_student        2133 non-null   float64
 6   student_id       2133 non-null   object
 7   gender           2133 non-null   object
 8   lunch            2133 non-null   object
 9   pretest          2133 non-null   float64
 10  posttest         2133 non-null   float64
dtypes: float64(3), object(8)
memory usage: 183.4+ KB
```

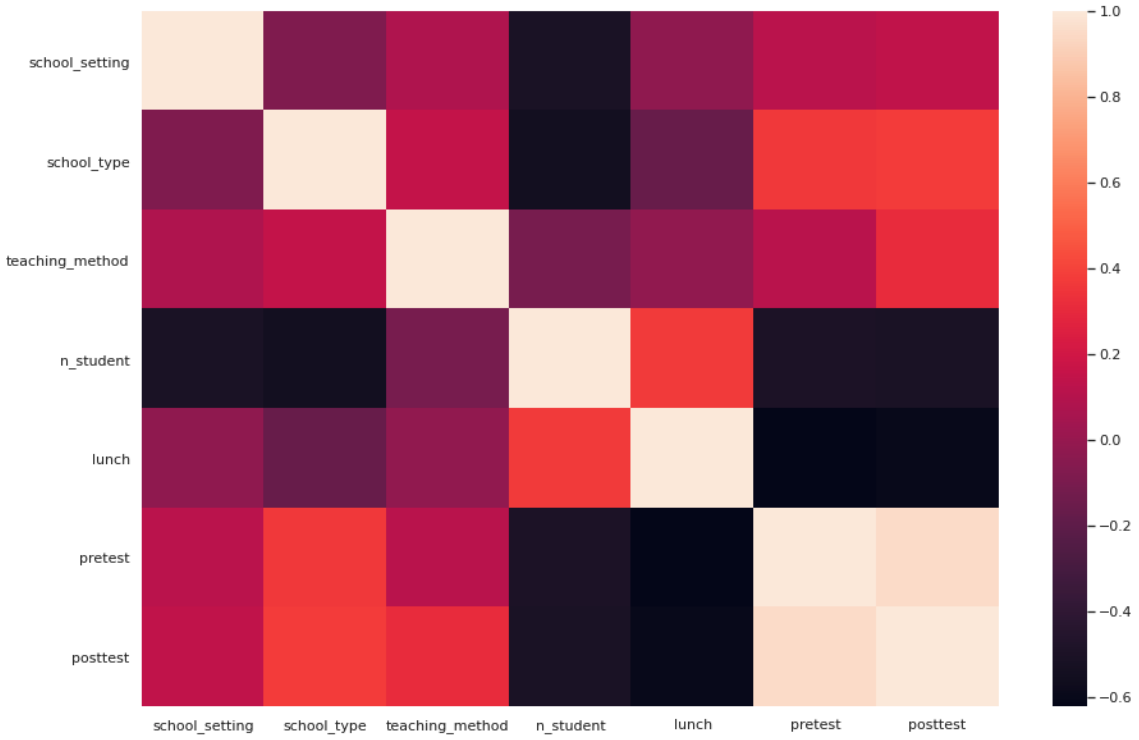
3. The dataset is quite good with no missing data but some useless features



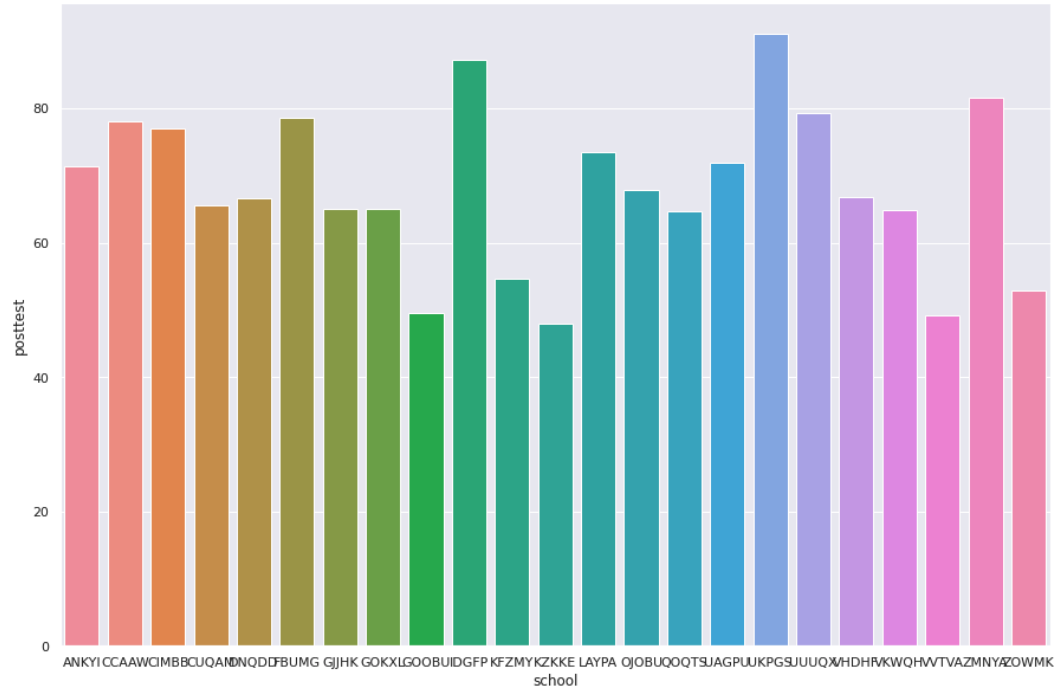
```
school == ['ANKYI' 'CCAAW' 'CIMBB' 'CUQAM' 'DNQDD' 'FBUMG' 'GJJHK' 'GOKXL' 'GOOBU'
 'IDGFP' 'KFZMY' 'KZKKE' 'LAYPA' 'OJOBU' 'QOQTS' 'UAGPU' 'UKPGS' 'UUUQX'
 'VHDHF' 'VKWQH' 'VVTVA' 'ZMNYA' 'ZOWMK']  and the count is 23


school_setting == ['Urban' 'Suburban' 'Rural']  and the count is 3


school_type == ['Non-public' 'Public']  and the count is 2


classroom == ['6OL' 'ZNS' '2B1' 'EPS' 'IQN' 'PGK' 'UHU' 'UWK' 'A33' 'EID' 'HUJ' 'PC6'
 '1Q1' 'BFY' 'OMI' 'X6Z' '2AP' 'PW5' 'ROP' 'ST7' 'XXJ' '197' '5LQ' 'JGD'
 'HCB' 'NOR' 'X78' 'YUC' 'ZDT' 'ENO' 'TSA' 'VA6' '18K' 'CXC' 'HKF' 'PBA'
 'U6J' 'W8A' '05H' '98D' 'G2L' 'P2A' 'XZM' '1VD' '21Q' '2BR' '3D0' '5JK'
 'O6A' 'QTU' 'AJ1' 'J8J' 'RA5' '5SZ' '6U9' 'FS3' 'XJ8' '0N7' '3XJ' 'RK7'
 'SUR' 'X2O' 'XZ4' '1SZ' '62L' 'NWZ' 'S98' '08N' '9AW' 'IPU' 'KXB' 'PGH'
 'XXE' '6C1' 'AE1' 'H7S' 'P8I' 'SSP' 'CD8' 'J6X' 'KR1' '341' 'D33' 'DFQ'
 'GYM' 'IEM' '7BL' 'A93' 'TB5' 'YTB' '1UU' '4NN' 'V77' 'CII' 'Q0E' 'QA2'
 'ZBH']  and the count is 97


teaching_method == ['Standard' 'Experimental']  and the count is 2


n_student == [20. 21. 18. 15. 16. 19. 17. 28. 27. 24. 14. 22. 23. 31. 25. 26. 29. 30.]  and the count is 18


student_id == ['2FHT3' '3JIVH' '3XOWE' ... 'YDR1Z' 'YUEIH' 'ZVCQ8']  and the count is 2133


gender == ['Female' 'Male']  and the count is 2


lunch == ['Does not qualify' 'Qualifies for reduced/free lunch']  and the count is 2


pretest == [62. 66. 64. 61. 63. 60. 67. 57. 56. 58. 54. 59. 65. 55. 68. 73. 70. 74.
 76. 69. 75. 78. 72. 71. 49. 53. 48. 52. 50. 46. 44. 51. 47. 43. 37. 40.
 39. 41. 38. 45. 36. 42. 31. 35. 33. 27. 30. 34. 32. 29. 28. 23. 26. 77.
 79. 82. 80. 85. 83. 84. 86. 89. 93. 88. 81. 87. 91. 22. 25.]  and the count is 69


posttest == [ 72.  79.  76.  77.  74.  75.  73.  78.  71.  70.  68.  66.  65.  67.
  63.  69.  82.  87.  80.  83.  81.  84.  85.  91.  86.  64.  88.  61.
  62.  58.  57.  59.  56.  60.  55.  54.  49.  53.  52.  50.  51.  48.
  39.  43.  45.  47.  42.  44.  46.  41.  36.  40.  32.  38.  35.  34.
  93.  90.  92.  97.  95.  99.  89.  94.  96.  98. 100.  37.]  and the count is 68
```

Uniqueness of the features explained:

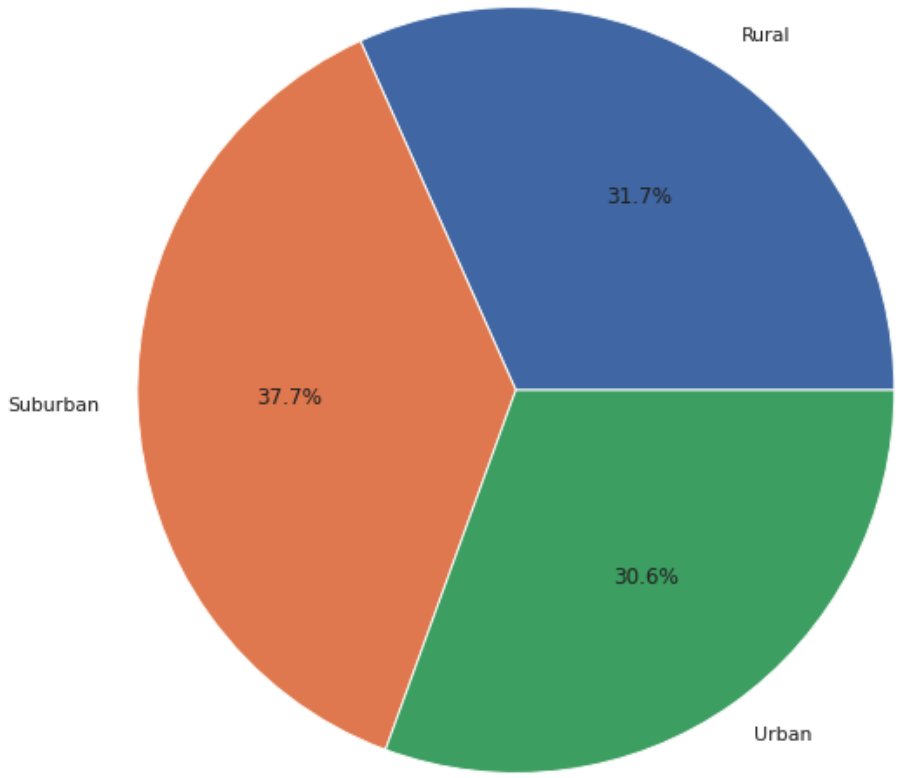
1. 23 Unique schools with 3 different settings
2. There are 97 classrooms with 2 different teaching methods
3. 2133 students with differences in qualified/ does not qualified for free lunch



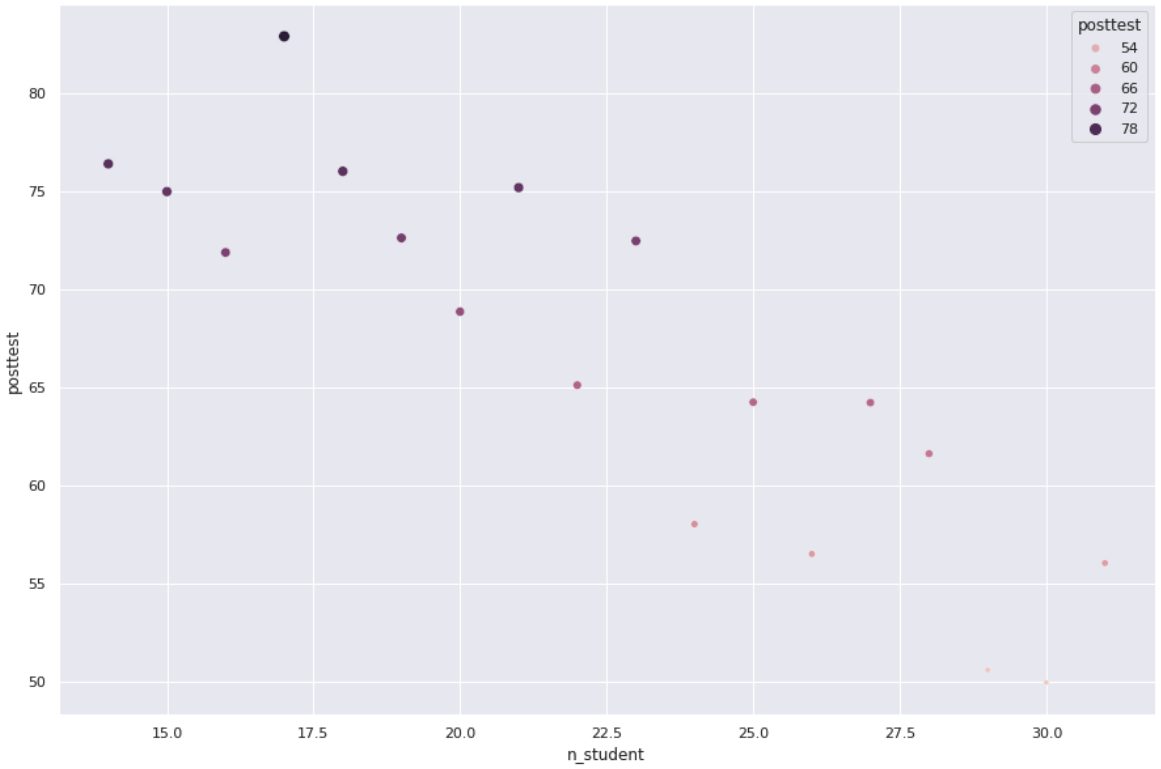
1. Pretest and posttest have a great relation
   (People who studied well, really performed well in their end exams)
2. School type has some impact on posttest and pretest
   (Private schools are really performing well)
3. Lunch and no of students in a class has some relation
   (Public schools had great no of students and they're eligible for free lunch)
4. Teaching method has little impact on posttest
   (Sometimes teaching method worked)
5. School type and school setting has decent negative correlation with no of student in a class
   (Public school had more students than private and there are 42% urban schools)
6. Lunch has strong negative relation with pretest and posttest
   (I think people who had proper food, they had good energy to score great marks)



Sometimes schools matter too. They produced good results compared to others.



Urban schools are performing poorly when Sub-urban was doing a great job and rural schools took runner-up place.



Number of students in the class is inversely proportional to the marks scored by the students in the class. While a class with approx 18 students scored better marks than others class with less or more students.

Model Results via Manual training:

1. Tried 4 models and ended up building an ensemble to check the performance. XGBoost is the winner out of other models trained for the problem with considering the explained variance score.
2. Tried Linear Regression, Logistic Regression and SVR.

Model Results via AutoML

1. Thought of building model via auto-sklearn and h2o automl but lack of time, so went with h2o automl automation
2. Out of 10 models automated for the model training, I found stacked ensemble model (GBM+RF) is really performing better for our problem

Thanks for reading the findings from my work.
Feedback is welcomed.