Data Science for Social Good

Jobin Shery Mathew
Anurag Sunil

Description

Project Title on Quality Education

These are the steps involved in

1. **Data Cleaning** - First we need to gather the data and after that we need to clean the data by removing the null values, arranging each according to their groups
2. **Exploratory Design Analysis** – We figure out the variability measures, central tendency, IQR values use them to figure out which performance is required
3. **Data Visualisation** - Data visualization is an essential part of data science that involves representing data and information in a visual form such as graphs, charts, and maps.
4. **Model Building** – The model we used is based on classification, using the KNN values and predicting its accuracy by visualizing the ROC curve of the fpr and tpr values
5. **Performance Analysis** – Using the above model , we check how accurate our code and how precise it is checking at the error received

Name - Jobin Shery Mathew, Anurag Sunil
Roll No. - URK21CS1033 , URK21CS5014
jobinshery@karunya.edu.in anuragsunil@karunya.edu.in
GitHub Repository Link - https://github.com/AnuragSunil/DataScience

TABLE OF CONTENTS

PROBLEM STATEMENT

- You are given the State-wise Gender parity index - primary school from 2001-2010. Make a model to analyze the data to get estimations.
- The UN SDG for the above problem statement is to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

SOLUTION

- **With our visualizaion we can predict how accurate is our result**
- **The output of the model would be a set of estimated GPI values for primary schools in different states over the given time period. This information could be used by policymakers and education professionals to understand the trends in gender equality in primary education and develop strategies to promote gender parity in education**
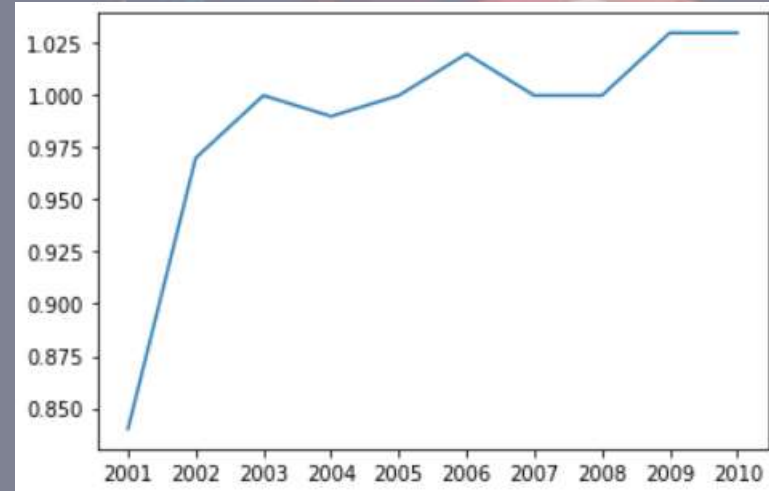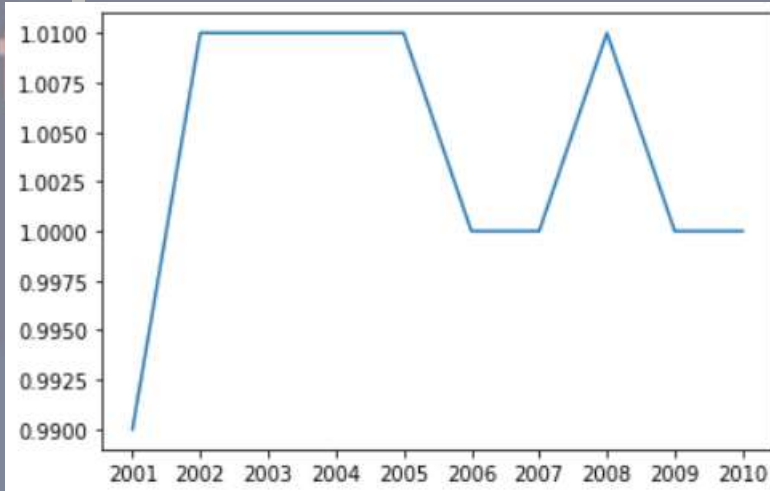- **Comparison with existing models**

DATASET USED

- **Data set used is 8.csv**
- **It contains the gender parity index of each state**

	School Type	India/ State/ UTs	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
0	Primary School	Andaman and Nicobar Islands	1.00	0.95	0.95	0.98	1.00	1.02	1.06	1.00	0.98	0.97
1	Primary School	Andhra Pradesh	0.99	1.01	1.01	1.01	1.01	1.00	1.00	1.01	1.00	1.00
2	Primary School	Arunachal Pradesh	0.83	0.88	0.90	0.89	0.90	0.90	0.92	0.95	0.96	0.96
3	Primary School	Assam	0.84	0.97	1.00	0.99	1.00	1.02	1.00	1.00	1.03	1.03
4	Primary School	Bihar	0.64	0.81	0.80	0.75	0.75	0.77	0.82	0.84	0.87	0.94
5	Primary School	Chandigarh	0.97	0.98	0.98	0.90	0.87	0.83	0.87	0.98	0.99	0.99
6	Primary School	Chhattisgarh	0.87	0.95	0.99	0.94	0.77	0.94	0.95	0.95	0.95	0.96
7	Primary School	Dadra and Nagar Haveli	0.72	0.86	0.88	0.93	0.96	0.98	1.01	1.08	1.05	1.03
8	Primary School	Daman and Diu	0.81	0.96	0.99	0.88	0.87	0.92	0.86	1.02	1.04	1.08
9	Primary School	Delhi	0.95	1.03	0.99	1.11	1.04	1.00	1.02	1.02	1.02	1.03

DATA VISUALIZATION

- **Data visualization is an essential part of data science that involves representing data and information in a visual form such as graphs, charts, and maps**
- **Data is visualized for every state in all the years. Here we are given with 10 years , visualizing each state with the subsequent year and its representations in graphs,**
- **For every state the graph changes depending on the rates**

DATA VISUALIZATION

- The snippet shows the HeatMap of the dataset given
- In a Heat Map – The darker region shows the higher range of rates in that particular year.
- Lighter region describes the lower concentration of rates in those years.
- In the X axis we take the years in consideration and in the Y axis we take the States in consideration.
- By viewing the map we can see the various regions where the rates are higher and lower in the particular year.

Primary School Enrollment Rates by State/ UT in India (2001-2010)
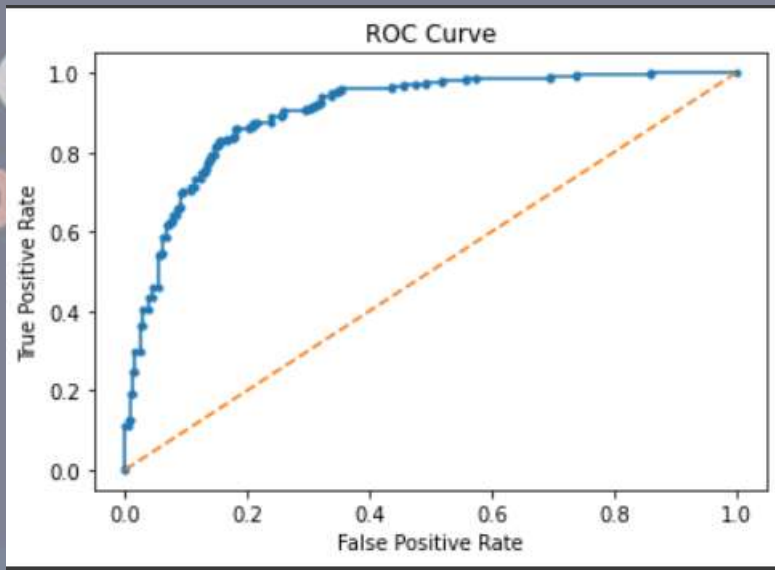
MODEL BUILDING

- **We use classification model to get classify the states in each of the years. Using this we use the KNN values predict the result on the basis of accuracy and precision**
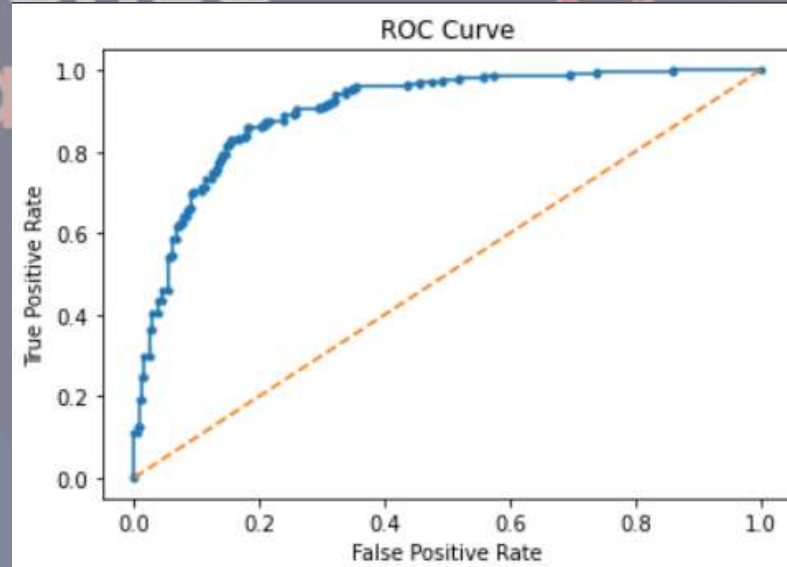
PERFORMANCE METRICS

- **Evaluation of the performance and accuracy of your model**


```
Confusion Matrix:
 [[75 14]
  [20 91]]
Accuracy: 0.83
Recall: 0.82
Precision: 0.867
Specificity: 0.843
F-score: 0.843
AUC score: 0.913
```

The accuracy of our model is 0.83
The precision of our model is 0.843

INSIGHTS GAINED AND REFERENCES USED

- **This project help us to understand the various implementations in data Science. Using classification we predicted the dataset to get how accurate the data is using KNN values and ROC curve**

- **References used – Google collabs from the trainers**