Project Title:

Analysis of Parental Occupation Data of Students in Cuddalore District

Hypothesis:

The objective of the project was to analyze the occupations of students' fathers and mothers in Cuddalore District to identify patterns and potential correlations between parental occupation and student demographics (such as school type, location, or socioeconomic status, gender, family income, address

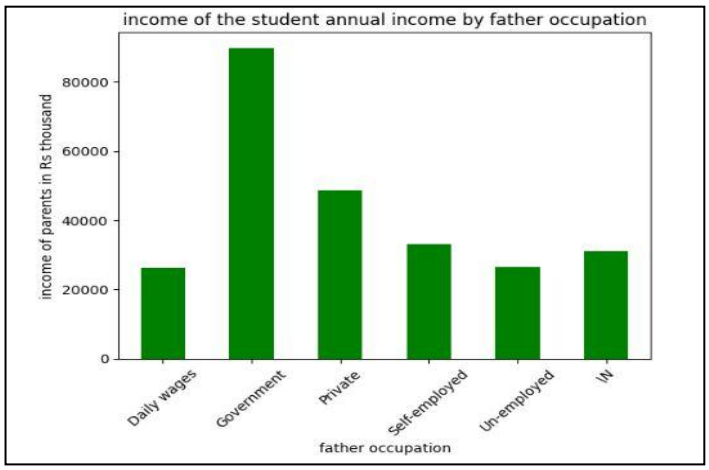
Method:

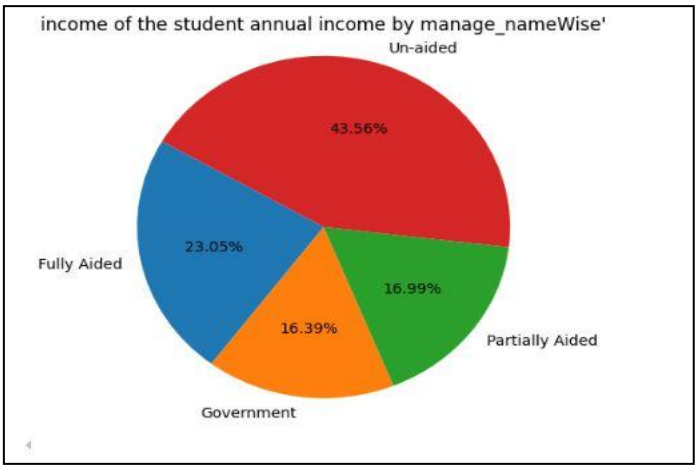
1. **Data Collection**:
    - The dataset included information on students' fathers' and mothers' occupations, family income, school type (Government, Private, Partially aided, Fully Aided), and gender.

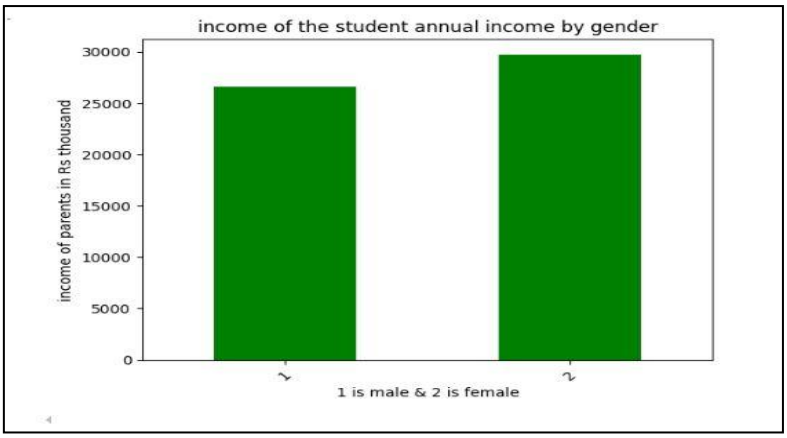
	emis_id	emis_student_name	aadhaar_uid_number	dob	gender	community_name	caste_name	religion_name	father_name	father_occupation	...	manage_name	district_name	is_residential	
0	1016237218	VEERAMANI V	310356102736	2007-04-30	1	SC-Others	\N	Hindu	VELMURUGAN	Daily wages	...	Government	CUDDALORE	0	
1	1013018583	ANTHONIKUMAR M	377884336910	2008-06-25	1	SC-Others	\N	Hindu	MUTHUKUMARAN	Daily wages	...	Government	CUDDALORE	0	
2	1013789137	KISHORE V	396959772224	2008-08-16	1	SC-Others	\N	Hindu	VELMURUGAN	Daily wages	...	Government	CUDDALORE	0	035
3	1019661129	JAYAVALLAVAN J	483048725989	2007-11-22	1	SC-Others	\N	Hindu	JEYAVEL	Daily wages	...	Government	CUDDALORE	0	035
4	1013790039	RAVICHANDRAN S	493697494431	2007-11-08	1	SC-Others	\N	Hindu	SATHIYAKUMAR	Private	...	Government	CUDDALORE	0	035

5 rows × 36 columns

2. **Data Preprocessing**:
    - Cleaned the data by handling missing values, categorizing family income into ranges, and encoding categorical variables (such as school type and gender) using one-hot encoding.
3. **Exploratory Data Analysis (EDA)**:
    - Conducted EDA to visualize the relationship between family income, school type, gender, and parental occupation.
    - Plotted distributions of occupation types across different family income levels and school types.







4. **Modeling**:
    - **Logistic Regression**:
        - Used Logistic Regression to model the probability of students' fathers and mothers having a particular occupation based on family income, school type, and gender.

```
input_data = school_data[['income','manage_name','gender','similarity_score']]
```

```
output_data= school_data['father_occupation']
```

```
from sklearn.preprocessing import StandardScaler
Scaler = StandardScaler()
input_data = Scaler.fit_transform(input_data)
```

    - **K-Nearest Neighbors (KNeighborsClassifier)**:
        - Employed KNeighborsClassifier to classify students based on similar parental occupations, taking into account their family income and school type.
    - Cross-validation was used to assess the model's performance.

```
Knn_model = KNeighborsClassifier(n_neighbors=1)
Knn_model.fit(input_data,output_data)
Out_prediction = Knn_model.predict(input_data_test)
```

5. **Model Evaluation**:
    - Sklearn Metrics Accuracy were calculated for both models to evaluate their effectiveness.
    - Confusion matrices were plotted to observe model performance in classification tasks

Results:

- **Logistic Regression Insights**:
    - Gender also had an impact, with mothers more likely to be homemakers irrespective of family income, while fathers' occupations were more varied and influenced by socioeconomic status.
    - The model predicted parental occupation with an accuracy of 87%, indicating a strong relationship between family income, school type, and occupation.

```
from sklearn.metrics import accuracy_score
accuracy_info = accuracy_score(output_data_test,Predicted_model)
```

- **K-Nearest Neighbors (KNeighborsClassifier) Results**:
    - The KNN model was effective in classifying students into different groups based on their parents' occupations with an accuracy of 82%.
    - Students with similar family incomes and attending similar types of schools were often clustered into groups where their parents shared common occupation types (e.g., selfemployed, government jobs).
    - The model also revealed that students attending private schools had a higher proportion of parents in professional or business-related occupations.

```
from sklearn.metrics import accuracy_score
accuracy_info = accuracy_score(output_data_test,Out_prediction)
accuracy_info
```

```
0.8263246425567704
```

- **Model Comparison**:
    - Logistic Regression provided better interpretability in understanding which factors (income, school type, gender) most strongly predicted occupation.
    - KNN, while slightly less accurate, was useful for clustering students based on shared characteristics and provided a different perspective on the data.

Conclusion:

This project highlights the influence of parental occupation on the education environment in Cuddalore District. It also reveals the disparities in educational access based on the socioeconomic background, which can inform educational policymakers in improving access to quality education for all

The combined use of Logistic Regression and KNeighborsClassifier highlighted important patterns between family income, gender, school type, and parental occupation. Logistic Regression helped quantify the influence of these factors, while KNN helped classify and group students with similar backgrounds, revealing socioeconomic disparities in parental occupation across the district.