## شعار جامعة بنها الجديدLogo Description automatically generated

**Benha University**

**Faculty of Computers &**

**Artificial Intelligence**

**Benha University**

**Faculty of Computers &   
 Artificial Intelligence**

**Data Science**

**“**Adult Census Income**”**

**Computer Science Departement,**

***Project Team***

1- Mona Mohammed Hamdy Gomaa.

2- Manar Hamada Elsayed Mahrous.

3- Menna Allah Hesham Eraky.

4- Ahmed Mohammed Saeed Mohammed.

5- Ahmed Mohammed Kassem Mohammed.

6- Ahmed Mahmoud Mohammed Mohammed.

7- Adham Ashraf Ebrahim Elganzoury.

8- Khaled Abdelmohsen Sayed Mohammed.

***Under Supervision of***

**Dr. Amr Nagy.**

**Eng. Amal.**

**Problem Definition:**

The problem domain is binary income classification. The goal is to build a model that predicts whether an individual makes over $50,000 per year based on anonymized census data. This can be valuable for understanding factors influencing income inequality and potentially informing targeted social programs.

**About the Dataset**

* **Age**: Describes the age of individuals.
* **Workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* **fnlwgt**: Discrete.
* **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* **education-num**: Number of years spent in education.
* **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
* **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* **sex**: Female, Male.
* **capital-gain**: Discrete.
* **capital-loss**: Discrete.
* **hours-per-week**: Discrete.
* **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
* **salary**: >50K,<=50K.

**Project Brief:**

**Data Exploration:**

* Analyze the features and their distributions.
* Identify missing values and handle them using appropriate techniques (e.g., imputation, deletion).
* Explore potential relationships between features and income levels using visualizations and statistical analysis.
* Remove Outliers.
* Handle imbalanced class distributions.

**Data Preprocessing:**

* Encode categorical features into numerical representations suitable for machine learning algorithms.
* Consider feature scaling or normalization if necessary.
* Split the data into training and testing sets for model development and evaluation.

**Visualizations:** allow us to easily understand any patterns, trends, or outliers in a data set. The task performing is the basic visualization methods like:

* Line plots.
* Area plots.
* Histogram.
* Bar charts.
* Pie charts.
* Box plots.
* Scatter plots.
* Bubble plots.

**Model Building**:

* Train and evaluate various machine learning models (Decision trees, logistic regression, random forests, support vector machines, KNN, Naïve Bayes, K-means).
* Fine-tune hyperparameters of the chosen models to optimize performance.
* A screenshot of a graph

  Description automatically generatedConsider techniques like ensemble methods to potentially improve classification accuracy.

**Model Evaluation**:

* Use metrics like accuracy, precision, recall, and ROC to assess the performance of your models on the held-out testing set.
* Compare the performance of different models and choose the one with the best overall outcome.

A graph of a graph

Description automatically generated with medium confidence**Visualization:**



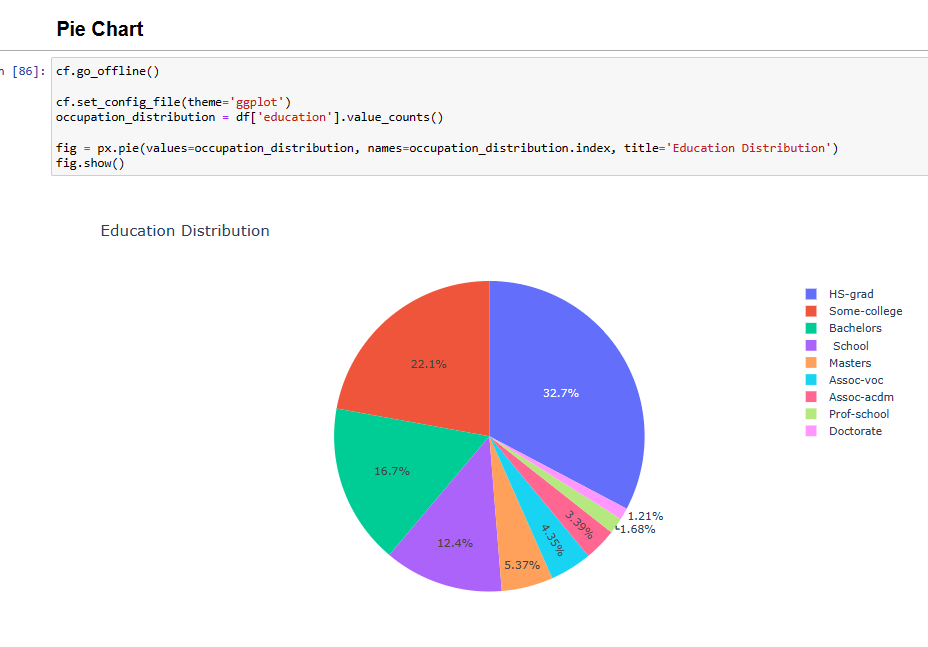
A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

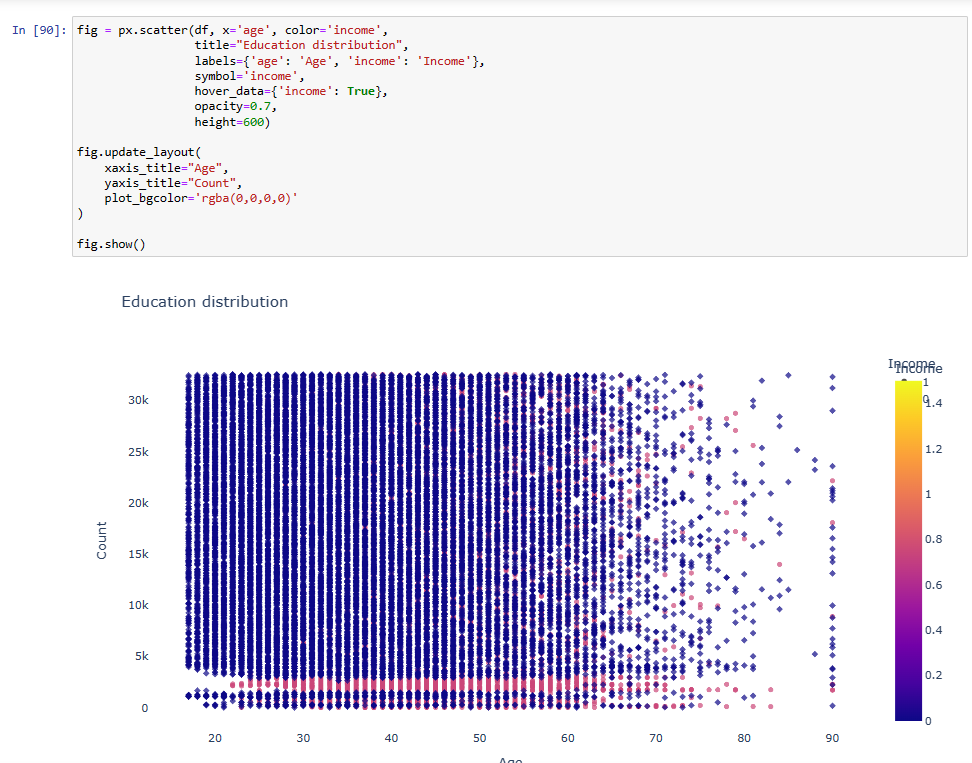


A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated



A screen shot of a computer

Description automatically generated