AIT614_005

Project Proposal

Team – 5
Pravallika Avula
Saipriya Bethi
Sai Roopesh Diddi
Snehita Moturu
Sameeksha Muralidhar Gupta

Professor – Dr. Lindi Liao
AIT614 - 005: Big Data Essentials

George Mason University

19 March 2023

Project Proposal
Team 5

Introduction:

The economy of a country is dependent on the employment and well-being of its citizens. Many people use income to determine the well-being of the people. Income has traditionally been used to measure household living standards. However, income is not the only available metric. Other factors that influence income include education, gender, family, race, and age. Every day, a person's income can fluctuate for a variety of reasons. As a result, it is critical to assess an individual's financial situation and estimate their earnings.
As a result, we used census income data to build a model that predicts an individual's income. Census income data is frequently used in big data analytics because it provides valuable information about a population's demographic and economic characteristics. This data can be used to gain understanding of consumer behavior, market trends, and economic patterns. Census income data shows how income is distributed across different segments of the population. Income data from the census can be used to decipher economic patterns and trends. This data is critical for forecasting economic growth, identifying economic opportunities, and informing policy decisions.

Related Work:

Several scientists and analysts have used this dataset to predict people with incomes greater than $50,000. The dataset has been referenced in many academic papers. In many of these papers, researchers studied the performance of augmenting existing classification machine learning algorithms, such as SVMs and K-NNs by boosting, partitioning, squashing, etc. Our goal is to build a simple model to predict the income based on the provided data, so that we can evaluate the various factors like age, education, occupation, hours per week and try to advance them in order to improve the financial status of citizens.

Objectives:

- Goal of the project is to apply Visualization techniques to census data to gain insights into the distribution of various attributes such as age, education, and occupation across different income levels.
- We also want to detect the anomalies, such as individuals who have extremely high or low incomes compared to others with similar attributes.
- we would like to determine the relative importance of different attributes on income levels. The goal is to identify the most important factors that contribute to high or low incomes.
- To identify the similarities in the attributes and group them.
- To build a model to predict the income based on different attributes.

Proposed selected Dataset:

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was

extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). As users we are mainly focusing on the income and the factors affecting the income. The adult dataset is a fairly large set, consisting of 48,842 instances. There are 14 attributes prescribed to each person. The dataset contains missing values that are marked with a question mark character (?). There are a total of 48,842 rows of data, and 3,620 with missing values, leaving 45,222 complete rows. There are two class values '*>50K*' and '*<=50K*', meaning it is a binary classification task. The classes are imbalanced, with a skew toward the '*<=50K*' class label.

The dataset provides 14 input variables that are a mixture of categorical, ordinal, and numerical data types. The complete list of variables is as follows:

- Age.
- Workclass.
- Final Weight.
- Education.
- Education Number of Years.
- Marital-status.
- Occupation.
- Relationship.
- Race.
- Sex.
- Capital-gain.
- Capital-loss.
- Hours-per-week.
- Native-country.
- Income

Final weight refers to population totals derived from CPS (Current Population Survey) by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights.

The income of the individual is dependent on the above attributes. We have to figure out the following

- Analysing the features contributing to the highest income.
- Which age group, working class, occupation, and sex have the highest income?
- How are the attributes dependent on each other, and identify if there are any similarities.

Project Proposal
Team 5

Description of the proposed system:

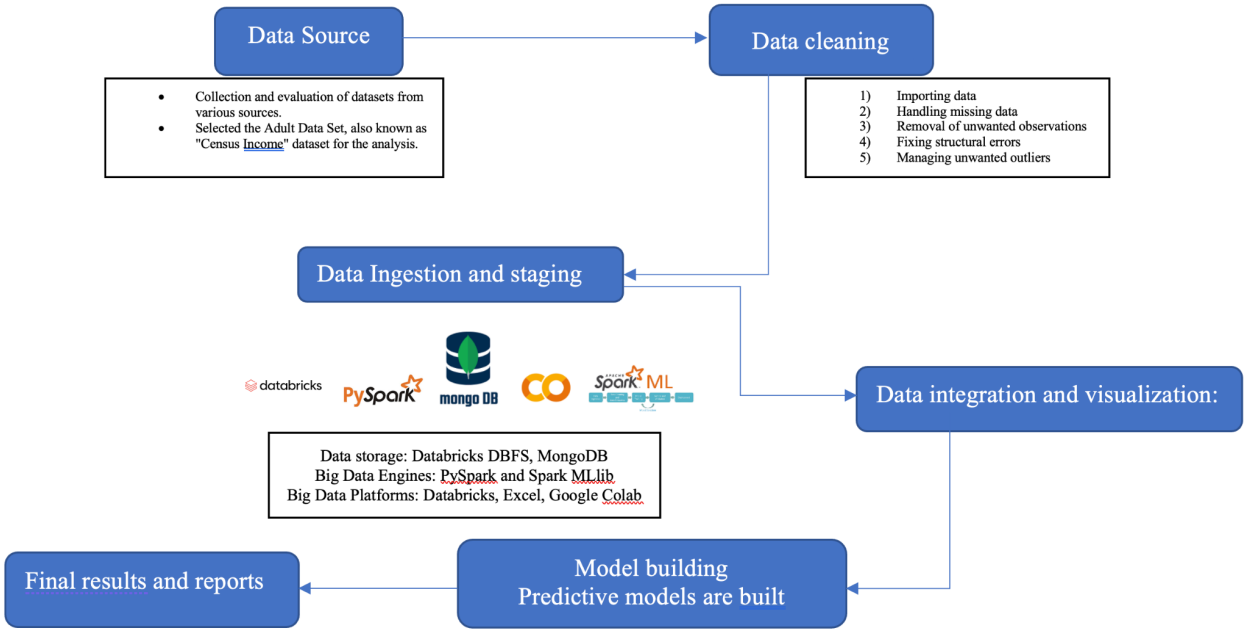


Figure 1: Adult Census Income System Architecture

Data Processing:

Data source and Data Cleaning: The data that we have chosen for our project is Census Income for which as a part of data processing we are performing the below in order to ensure that the data is accurate, complete, and consistent.

1. Checking for missing values
2. Removing irrelevant columns
3. Removing duplicates
4. Removing outliers
5. Scaling numerical data

- For the data storage the system being proposed uses MongoDB as a NoSQL database and a cloud-based platform called Databricks to perform the data analysis, processing, and visualizations using Apache Spark.

- The MLlib library, which is scalable and used for machine learning, will be utilized for model building in the proposed system.

- Python, specifically pySpark will be used as the programming language for interfacing with both MongoDB and Apache Spark.

Data integration and visualization:

Data integration refers to the process of merging data from diverse sources to form a consistent and unified dataset. This is frequently required because data may be stored in various formats or structures and may be dispersed across multiple databases or systems. To ensure the accuracy, coherence, and comprehensiveness of the final dataset, data integration activities such as data cleansing, data conversion, and data consolidation may be necessary.

On the other hand, visualization is putting data in a graphic or visual style that makes it simpler to grasp and comprehend. Creating charts, graphs, and other pictorial representations of data are examples of tasks that fall under this category. By using visualization, users can discover patterns, trends, and insights that might not be immediately obvious when viewing the raw data. Data integration and visualization work together to help the users gain a greater knowledge of large datasets so they may utilize that information to make better decisions. These are crucial tools for data scientists, analysts, and other experts who frequently work with huge and complex information.

Below mentioned are the different types of data visualization techniques which would be used for this proposal : [2]

1. **Bar charts:** To compare different categories or values.
2. **Line charts:** To show trends or changes over time.
3. **Scatter plots:** To show the relationship between two variables.
4. **Pie charts:** To show the proportion of different categories in a dataset.
5. **Box plots:** To show the distribution of a variable and identify outliers.

Data Analytics Methods:[3]

There are various data analytics methods that can be used for the chosen dataset. Here are some methods that are being considered:

1. **Descriptive analysis:** This method could be used to gain insights into the characteristics of the data such as the distribution of the variables, summary statistics, and correlation between the variables. This can be useful to identify trends and patterns in the data.
2. **Predictive analysis:** This method would be used to develop models that can predict the outcome of interest such as the income of a person based on their demographic and socioeconomic variables. This can be useful in making informed decisions and developing strategies based on predictions.

3. **Classification analysis:** In order to classify the individuals into different categories based on their income such as whether they earn more than 50K or not this method would be an appropriate one. This can be useful to understand the characteristics of different groups and develop targeted policies accordingly.
4. **Clustering analysis:** To group the individuals based on similar characteristics such as age, education, and occupation. This can be useful to identify homogeneous groups and develop policies based on their specific needs.

Overall, based on our analysis we will be considering the other analytical methods for our proposal.

Proposed development platforms:

Below mentioned are the software platforms that would be required for this project:

- **MS Excel:** Excel can be used for minor data cleaning tasks as CSV files can be opened in Excel. Additionally, Excel will be used to organize and view data during the project.

- **Databricks (Community Edition):** Databricks Community Edition is a free cloud platform for experimenting with Apache Spark, offering an interactive workspace and tools for data exploration, visualization, and collaboration, as well as machine learning libraries for building predictive models and advanced analytics.

- **MongoDB (Community Edition):** The free and open-source edition of MongoDB, called MongoDB Community Edition, enables users to save and retrieve data in a flexible JSON-like format and offers robust functionalities for data processing and analysis.

- **Apache Spark:** Apache Spark is an open-source, distributed computing system that enables fast and flexible processing of large-scale data sets. It offers a wide range of libraries and tools for data processing, machine learning, and graph processing, making it a popular choice for big data analytics.

- **Google Colab:** Google Colab is a web-based platform that provides users with free access to a Jupyter notebook environment, allowing them to write, run, and share Python code, as well as to access various pre-installed packages and libraries.

Apart from that, Python is a popular programming language for data analysis and machine learning. It has numerous libraries and frameworks such as NumPy, pandas, and Scikit-Learn that can be used to manipulate and analyze data.

Project Proposal
Team 5

Project Tasks and Timeline:

Tasks Proposed: Our project is divided into two parts:

- Data Analysis
- Model Development

Data Analysis: This section primarily consists of bridging data from MongoDB to data bricks, followed by data cleaning, processing, and visualizing the presented data in graphs for better and more practical understanding.
Model Development: This stage entails selecting the best model for classification with the lowest error rate, highest accuracy, and highest precision by dividing the data into train and test sets.

Timelines proposed:

Part 1: Data Analysis (Timeline – March 25th, 2023 to April 15th, 2023)
Sameeksha Gupta and Saipriya Bethi
Part 2: Model Construction: (Timeline – April 16th, 2023 – April 24th, 2023) Snehita Moturu, Pravallika Avula, Roopesh Diddi

References:

[1] Dataset link: UCI Machine Learning Repository: Adult Data Set. (n.d.). Retrieved March 18, 2023, from https://archive.ics.uci.edu/ml/datasets/Adult

[2] Great Learning Team. (2022, September 2). *Data Visualization Techniques, Tools and Concepts*. Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/understanding-data-visualization-techniques/

[3] Stevens, E. (2023, January 4). *The 7 Most Useful Data Analysis Methods and Techniques*. CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/

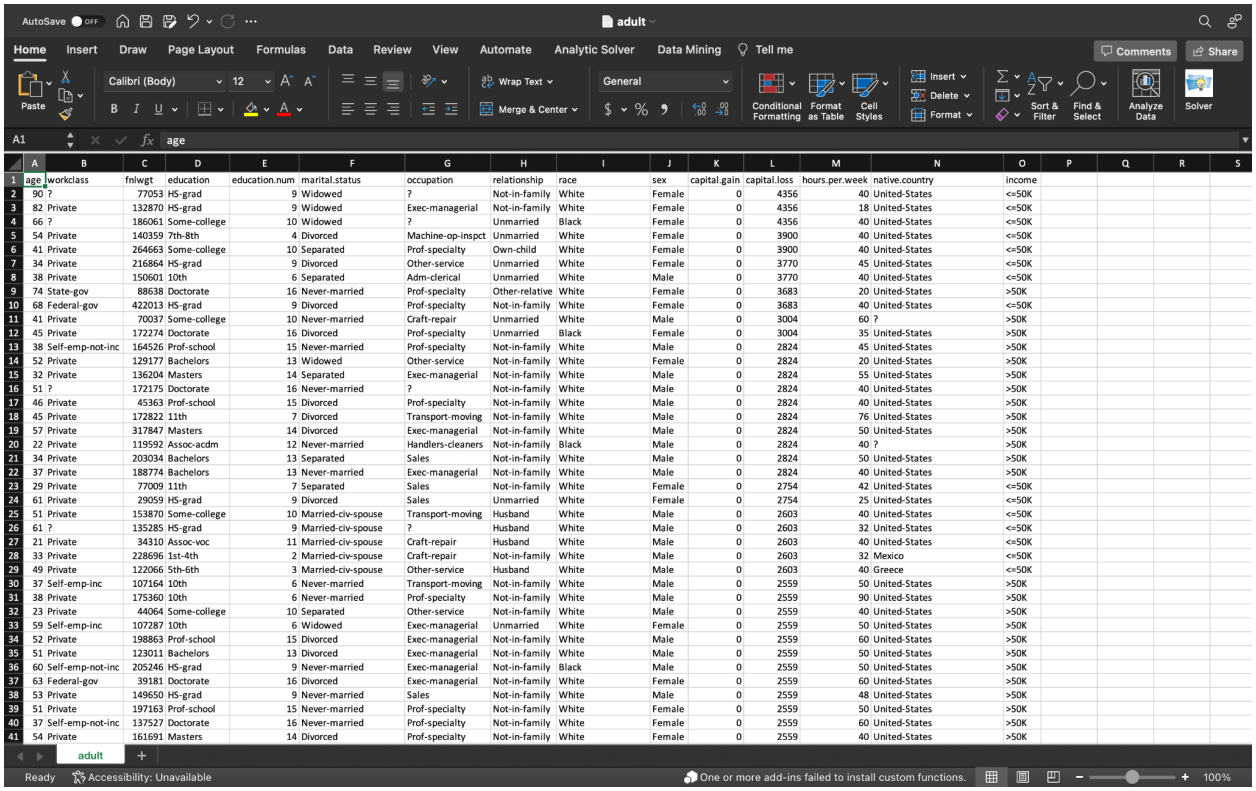
[4] Kohavi, R. (n.d.). *Data Mining and visualization group Silicon Graphics, inc.*. nbtree.dvi. Retrieved March 19, 2023, from http://robotics.stanford.edu/people/ronnyk/nbtree-talk.pdf

[5] Bureau, U. S. C. (2023, March 7). Census.gov. Retrieved March 18, 2023, from https://www.census.gov/

Project Proposal
Team 5

Appendix:

Below is the screenshot of the dataset that contains 48,842 rows and 14 attributes.



age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K
34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	Female	0	3770	45	United-States	<=50K
38	Private	150601	10th	6	Separated	Adm-clerical	Unmarried	White	Male	0	3770	40	United-States	<=50K
74	State-gov	88638	Doctorate	16	Never-married	Prof-specialty	Other-relative	White	Female	0	3683	20	United-States	>50K
68	Federal-gov	422013	HS-grad	9	Divorced	Prof-specialty	Not-in-family	White	Female	0	3683	40	United-States	<=50K
41	Private	70037	Some-college	10	Never-married	Craft-repair	Unmarried	White	Male	0	3004	60	?	>50K
45	Private	172274	Doctorate	16	Divorced	Prof-specialty	Unmarried	Black	Female	0	3004	35	United-States	>50K
38	Self-emp-not-inc	164526	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Male	0	2824	45	United-States	>50K
52	Private	129177	Bachelors	13	Widowed	Other-service	Not-in-family	White	Female	0	2824	20	United-States	>50K
32	Private	136204	Masters	14	Separated	Exec-managerial	Not-in-family	White	Male	0	2824	55	United-States	>50K
51	?	172175	Doctorate	16	Never-married	?	Not-in-family	White	Male	0	2824	40	United-States	>50K
46	Private	45363	Prof-school	15	Divorced	Prof-specialty	Not-in-family	White	Male	0	2824	40	United-States	>50K
45	Private	172822	11th	7	Divorced	Transport-moving	Not-in-family	White	Male	0	2824	76	United-States	>50K
57	Private	317847	Masters	14	Divorced	Exec-managerial	Not-in-family	White	Male	0	2824	50	United-States	>50K
22	Private	119592	Assoc-acdm	12	Never-married	Handlers-cleaners	Not-in-family	Black	Male	0	2824	40	?	>50K
34	Private	203034	Bachelors	13	Separated	Sales	Not-in-family	White	Male	0	2824	50	United-States	>50K
37	Private	188774	Bachelors	13	Never-married	Exec-managerial	Not-in-family	White	Male	0	2824	40	United-States	>50K
29	Private	77009	11th	7	Separated	Sales	Not-in-family	White	Female	0	2754	42	United-States	<=50K
61	Private	29059	HS-grad	9	Divorced	Sales	Unmarried	White	Female	0	2754	25	United-States	<=50K
51	Private	153870	Some-college	10	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2603	40	United-States	<=50K
61	?	135285	HS-grad	9	Married-civ-spouse	?	Husband	White	Male	0	2603	32	United-States	<=50K
21	Private	34310	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	White	Male	0	2603	40	United-States	<=50K
33	Private	228696	1st-4th	2	Married-civ-spouse	Craft-repair	Not-in-family	White	Male	0	2603	32	Mexico	<=50K
49	Private	122066	5th-6th	3	Married-civ-spouse	Other-service	Husband	White	Male	0	2603	40	Greece	<=50K
37	Self-emp-inc	107164	10th	6	Never-married	Transport-moving	Not-in-family	White	Male	0	2559	50	United-States	>50K
38	Private	175360	10th	6	Never-married	Prof-specialty	Not-in-family	White	Male	0	2559	90	United-States	>50K
23	Private	44064	Some-college	10	Separated	Other-service	Not-in-family	White	Male	0	2559	40	United-States	>50K
59	Self-emp-inc	107287	10th	6	Widowed	Exec-managerial	Unmarried	White	Female	0	2559	50	United-States	>50K
52	Private	198863	Prof-school	15	Divorced	Exec-managerial	Not-in-family	White	Male	0	2559	60	United-States	>50K
51	Private	123011	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Male	0	2559	50	United-States	>50K
60	Self-emp-not-inc	205246	HS-grad	9	Never-married	Exec-managerial	Not-in-family	Black	Male	0	2559	50	United-States	>50K
63	Federal-gov	39181	Doctorate	16	Divorced	Exec-managerial	Not-in-family	White	Female	0	2559	60	United-States	>50K
53	Private	149650	HS-grad	9	Never-married	Sales	Not-in-family	White	Male	0	2559	48	United-States	>50K
51	Private	197163	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Female	0	2559	50	United-States	>50K
37	Self-emp-not-inc	137527	Doctorate	16	Never-married	Prof-specialty	Not-in-family	White	Female	0	2559	60	United-States	>50K
54	Private	161691	Masters	14	Divorced	Prof-specialty	Not-in-family	White	Female	0	2559	40	United-States	>50K