Data Collection and Preprocessing Phase

GARMENT WORKER PRODUCTIVITY PREDICTION

Team ID: 1720673861

Date: 05/07/2024

TEAM:

NIKHIL PULAGAM

MULE VIGNESH

SHREYAS REDDY GUVVALA

1. Data Collection and Preprocessing Phase

3.2Data Collection Plan and Raw Data Sources Identified

There are many popular open sources for collecting the data. E.g., kaggle.com, UCI repository, etc.

In this project we have used .csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset. Download the dataset from the above link. Let us understand and analyze the dataset properly using visualization techniques.

Note: There are several approaches for understanding the data. But we have applied some of it here. You can also employ a variety of techniques.

Activity 1.1: Importing the libraries

Import the libraries required for this machine learning project, as shown in the image below.

Activity 1.2: Read the Dataset

Our dataset format could be in .csv, excel files,.txt, .json, and so on. With the help of pandas, we can read the dataset.

Since our dataset is a csv file, we use read_csv() which is pandas function to read the dataset. As a parameter we have to give the directory of the csv file.

df.head() will display first 5 rows of the dataset.

a. Data Exploration and Preprocessing

Data preparation, also known as data preprocessing, is the process of cleaning, transforming, and organizing raw data before it can be used in a data analysis or machine learning model.

The activity include following steps:

- removing missing values
- handling outliers
- encoding categorical variables
- normalizing data

Note: These are general steps to take in pre-processing before feeding data to machine learning for training. The pre-processing steps differ depending on the dataset. Depending on the condition of your dataset, you may or may not have to go through all these steps.

Activity 2.1: Handling Missing Values

Let's first figure out what kind of data is in our columns by using df.info(). We may deduce from this that our columns include data of the types "object", "float64" and “int64”.

This line of code is used to count the number of missing or null values in a pandas DataFrame. It returns a list of the total number of missing values in each column of the DataFrame.

This line of code fills the missing values in the “unfinished_items” column with the column mean.

Activity 2.2: Handling Independent Columns

This line of code drops the columns date and targeted_productivity from the dataframe.

The rename() method is being used to change the names of some columns. The method takes a dictionary as its argument, where the keys are the original column names and the values are the new names.

The first line of code selects a specific column named "quarter" from a table of data. The unique()method is then used to list all the different values in that column.
The second line of code changes some of the values in the "quarter" column. Specifically, it replaces any instances of the text "Quarter5" with the text "Quarter1" using the str.replace()method.

The first line of code is using the str.extract() method on the "quarter" column to extract the numerical part of each value. It uses a regular expression pattern r'(\d+)' to match any sequence of one or more digits in each value.
The second line of code is assigning the modified "quarter" column back to the same "quarter" column, effectively replacing the original column

with the modified one.

The first line of code is using the str.replace() method on the "department" column to replace any instances of the misspelled word "sweing" with the correct spelling "sewing".
The second line of code is also using the str.replace() method on the "department" column to replace any instances of the word "finishing " (with a space at the end) with the word "finishing" (without a space at the end). This will remove the extra space and standardize the spelling of the word.

Each line of code is using the astype() method to convert the data type of a specific column to an integer type. The column name is specified on the left-hand side of the equation, and the new integer type is specified as an argument to the astype() method.
Activity 2.3: Handling Categorical Values

This code is encoding the values in a column named "department" by using a LabelEncoder() object to convert the original values into numerical encoded values. The original and encoded values are printed before and after the encoding is performed.

This code is encoding the values in a column named "day" by using a LabelEncoder() object to convert the original values into numerical encoded values. The original and encoded values are printed before and after the encoding is performed.

3.3 Data Exploration and Preprocessing

Visual analysis is the process of examining and understanding data via the use of visual representations such as charts, plots, and graphs. It is a method for quickly identifying patterns, trends, and outliers in data, which can aid in gaining insights and making sound decisions.

Activity 2.1: Univariate analysis

Univariate analysis is a statistical method used to analyse a single variable in a dataset. This analysis focuses on understanding the distribution, central tendency, and dispersion of a single variable.

This code creates a bar chart using the Seaborn library to show the number of occurrences of each unique value in the "department" column of a Pandas DataFrame. The x-axis represents the unique values of the "department" column, and the y-axis represents the number of times each unique value appears in the column. The plt.xlabel(), plt.ylabel(), and plt.title() functions are used to add labels and a title to the plot. Finally, plt.show() is used to display the plot.

This code generates a pie chart using the Seaborn library to show the distribution of different quarters in a dataset. The value_counts() function counts the number of occurrences of each quarter in the dataset, and the resulting counts are used to create the pie chart. The pie chart shows the proportion of the total number of entries in the dataset that corresponds to each quarter. The title of the pie chart is "Quarter Distribution".

Activity 2.2: Bivariate analysis

Bivariate analysis is a statistical method used to analyse the relationship between two variables in a dataset. This analysis focuses on examining how changes in one variable are related to changes in another variable.

This code generates a line plot using the Seaborn library to show the relationship between team number and the number of unfinished items. The line plot shows how the number of unfinished items varies across different teams. The x-axis represents the team number, and the y-axis represents the number of unfinished items. The title of the line plot is "Line Plot of Unfinished Items by Team Number".