**Group\_9\_Task 3**:

Measure Accuracy, Measure the accuracy of both predictions. We have used RMSE (Root Mean Square Error) as the metric.

**Approach:**

The Root Mean Square Error is a metric that indicates how well a regression line fits the data. The predictions were made by dividing the data into two parts: (0.8 for training and 0.2 for testing for linear regression and 0.7 for training and 0.3 for testing for random forest regression). So, following the predictions, we utilize the train and test data to determine the RMSE value for feature as input and output as ‘Edu’. The accuracy value for the predicted data characteristics is calculated using the RegressionEvaluator. We can tell which one fits the model well based on the RMSE numbers.

**Dataset:** Similar to task2 we have used the dataset after obtaining the corrected and cleaned data from Task1 by removing the missing and out of range values we have written the output to an output file. Which is taken as the input to the Task3, and we implemented Linear Regression and Random Forest Regression for the corrected data and then measured the accuracy of those predictions using RMSE as a metric

We then selected few columns based on our input dataset for prediction, we didn’t have any direct relation between any of the columns for accurate predictions, we have tried to identify the closest columns which would give good prediction results with a correlation. We have used RegressionEvaluator for evaluation we have rounding the model and coefficients and printed all the required RMSE values.

**Brief description of the dataset and variable predicting:**

The same data set which from the Task1 i.e., after filling the null values for measuring the accuracy of predicted model from task2. Since we are predicting the models from the task2 so we have used the some input columns like Kidhome, Teenhome, Income, Mar\_st, MntWines, MntMeatProducts to predict how accuracy the models are?.

The variable used here to predict the accuracy is the Edu for calculating the MeanAbsoluteError, MeanSquaredError, RootMeanSquaredError, R2 and AdjustedR2

The columns that we have used as input and output for predications are:

**Input Columns:** Kidhome, Teenhome, Income, Mar\_st, MntWines, MntMeatProducts

**Output Column:** Edu

* First we have imported all the required libraries. We have used the columns that have been all converted to a double type and cleaned from the task1 as input for task3.
* **Reading data:** We utilized data frame reader to read the data from the data source to load the path of the data specified for the specific question and to load the data frame in the same format which the tabular format.
* We have used VectorAssembler to define out input and output columns for predictions we have combined all the specified input column into a single feature, then fit() and transform() the model defined. We then split the dataset into train and test using randomSplit() function.
* Repeated the same process for both regressions and then applied the LinearRegression() and RandonForestRegression() for the input and output columns specified. Finally, we have obtained and printed out the predictions that were obtained from regression functions.
* Then on the predicted data for models for linear regression we have used linear regression formula used coefficients and used RegressionEvaluator to obtain the RMSE value.

**Instructions to compile and run the program:**

**Imports Used:**

from pyspark.sql import SparkSession

from pyspark.sql import SQLContext

from pyspark.ml.linalg import Vectors

from pyspark.sql.functions import col

from pyspark.sql import functions as F

from pyspark.ml.regression import LinearRegression, RandomForestRegressionModel, RandomForestRegressor #importing the regressions and models required

import pandas as pd

from pyspark.ml.feature import VectorAssembler

from pyspark.sql.functions import count, mean, when, lit, create\_map, regexp\_extract

from pyspark.ml.evaluation import RegressionEvaluator

from pyspark import SparkContext, SparkConf

**Input:** As mentioned, we have used the output of Task 1 as the input for Task 3. task\_1\_output.csv is the output file for Task1 which we use as input file using following line.

**Output:** Is printed on the console.

1. Open the cluster and login into it.

2. Upload the Group\_9\_Task\_3.py file into session storage(local or csh)

3. Command: spark-submit Group\_9\_Task\_3.py

**Discussion of Results:**

Used the predictions of all feature containing input columns and output column as Edu were used to calculate the RMSE values using RegressionEvaluator, we have obtained that the RMSE value as 1. The value is high as the dataset doesn’t contain proper relation between any of the columns, but we tried to use the best combination of input columns in feature to obtain better results.