ANALYSIS DEDUCTIONS

CLASSIFICATION

	DECISION TREE (DT) ALGORITHM	K-NEAREST NEIGHBOR (KNN) ALGORITHM	NAIVE BAYES (NB) ALGORITHM
Correctly classified	64,559	55,268	37,897
Wrongly classified	14,992	24,283	41,654
Accuracy	81.154%	69.475%	47.639%
Error	18.846%	30.525%	52.361%
Cohen's Kappa(k)	0.743	0.581	0.328

Table 1: Summary of Classification Algorithms

From table 1, the **D*ecision Tree Algorithm*** has the highest accuracy of about 81% and the cohen's kappa which is a better metric for evaluating classifier performance and which shows the level of agreement between the predicted and actual values has its highest score in the DT (0.743)followed by KNN. Going by the cohen's kappa, the **D*ecision Tree Algorithm*** is the best algorithm for this model.

CLASS	RECALL			PRECISION			F-MEASURE		
	DT	KNN	NB	DT	KNN	NB	DT	KNN	NB
AB	0.783	0.602	0.768	0.783	0.636	0.371	0.783	0.618	0.5
C1	0.828	0.754	0.109	0.832	0.685	1	0.83	0.718	0.197
C2	0.812	0.608	0.599	0.815	0.683	0.43	0.813	0.643	0.501
DE	0.813	0.772	0.671	0.805	0.779	0.659	0.809	0.775	0.665

Table 2: Comparison between Recall, Precision and F-measure of the Algorithms

ANALYSIS OF RESULTS

From the table 2 above, the following observations were made:

1. Looking at the value of the f-measure which is the mean of recall and precision, all three algorithms predicted efficiently the C2 and DE Approximated Social Grade classes, most especially the DE_Approximated Social Grade class
2. The algorithms - DT and KNN produced very great performances when compared to NB.
3. Lastly, not all the algorithms efficiently predicted the C1_Approximated Social Grade. It is good to note that DT and KNN recorded higher recall, precision and f-measure whereas the NB recorded the lowest recall and f-measure for the C1 class.

REGRESSION

Two regression algorithms were applied to predict no of hours:

- Simple Regression Tree
- Linear Regression

Comparing the two algorithms, the linear regression (LR) algorithm gives better results because it has the higher coefficient of determination - R-squared (0.877) and the lower error values (4.065). R-squared being closer to 1 shows the actual values are close to the predicted values.

The error values showed by the Mean Absolute Error shows the average error magnitude produced in the model and as seen from the diagrams the error in LR is about 4 hours and that of SRT is a bit over 5 hours. The metric, Mean Signed Difference which tells the disparity between the actual and predicted values is lower in LR (0.03) than in SRT (0.251) proving that LR produces a better performance.

METRIC	LINEAR REGRESSION (LR)	SIMPLE REGRESSION TREE (SRT)
R^2	0.877	0.771
Mean Absolute Error	4.065	5.425
Mean Squared Error	22.405	41.904
Root Mean Squared Error	4.733	6.473
Mean Signed Difference	0.03	0.251
Mean Absolute Percentage Error	0.206	0.261
Adjusted R^2	0.877	0.771

Table 3: Showing Evaluation Statistics for the Algorithms

ASSOCIATION RULE MINING

Association rules come in form of If/Then statements (the antecedent and the consequent) showing the relationships between attributes in the data.

Several rules were generated and the rules with the highest rule lift were selected.

The highest rule lift is 6.269 and the least is 0.229.

	CONSEQUENT	ANTECEDENT	RULE LIFT	INTERPRETATION
1	Skilled Trades Occupation_Occupation	C2_Approximated Social Grade, Male_Sex, White_Ethnic Group, No_Student, Non-communal resident_Residence Type, Usual resident_Population Base	6.166	If an individual's occupation is in skilled trades then the individual belongs to C2 approximated social grade, is a white male who is not a student,

				and is both a non-communal resident or usual resident.
2	EA: Full-time student_Economic Activity	C1_Approximated Social Grade, Single_Marital Status, Non-communal resident_Residence Type, Usual resident_Population Base	5.749	If an individual is a economically active full-time student, then the individual is single, belongs to the C1 social grade and is both a non-communal resident and usual resident
3	Process and machine operatives_Occupation,	DE_Approximated Social Grade, Male_Sex, No_Student, Non-communal resident_Residence Type, Usual resident_Population Base	5.059	If an individual's occupation is in the process and machine operatives, then the individual is a male belonging to DE approximated social grade, is not a student, is a usual resident and a non-communal resident
4	Construction_Industry	Skilled Trades Occupation_Occupation, Male_Sex, No_Student, Usual resident_Population Base,	5.053	If an individual works in the construction industry, then the individual is a male who works in a skilled trades,

				not a student and is a usual resident.
5	FT:49 or more_Hours worked per week	(45-60]_No of hours[Binned], Male_Sex, UK_Country of Birth, Usual resident_Population Base, White_Ethnic Group, No_Student	4.853	If an individual works full-time for 49 or more hours, then the individual is a white male from UK who works 45-60 hours, is not a student and is a usual resident

Table 4: The Different Association Rules and Their Interpretation

CLUSTERING

Clustering involves assigning data points to cluster centers. The data preparation included filtering out some columns like person id, using category to number node to assign seven(7) variables to integers, removing missing values, randomly selecting some 2,700 rows (random seed was used) and normalizing them.
The two clustering algorithms used:
1. **K-means** - Three(3) clusters were formed with coverage as:
Cluster 0 = 567
Cluster 1 = 850
Cluster 2 = 1283
2. **K-medoids -** Three(3) clusters were formed with coverage and the center of the data points;
A. Row302658 - 118
B. Row380519 - 1388
C. Row412488 - 1194
From the analysis done, the overall mean silhouette coefficient for k-means is 0.175 showing that the clusters are overlapping - a sample belongs to more than one cluster. Also, the mean silhouette coefficient for k-medoids is 0.226 showing that the clusters are overlapping - a sample belongs to more than one cluster.

COMPARISON
Both clustering algorithms have overlapping clusters and they also share similarities such as:
1. They all belong to people with very good health.
2. All cluster points have single individuals.
3. The cluster points include individuals who work either part-time or full-time.
4. The cluster points include more females than males.

The major difference between the clustering algorithms is the fact that;

- In K-means, the cluster data points cover people from the C2 approximated social grade whereas K-medoids cluster points are people from C1 approximated social grade.
- The k-means has better coverage when compared to k-medoids.