UNIVERSITY OF STIRLING

CSCU9T6 – Data Mining Assignment 2020

Student ID: 2626628

Contents

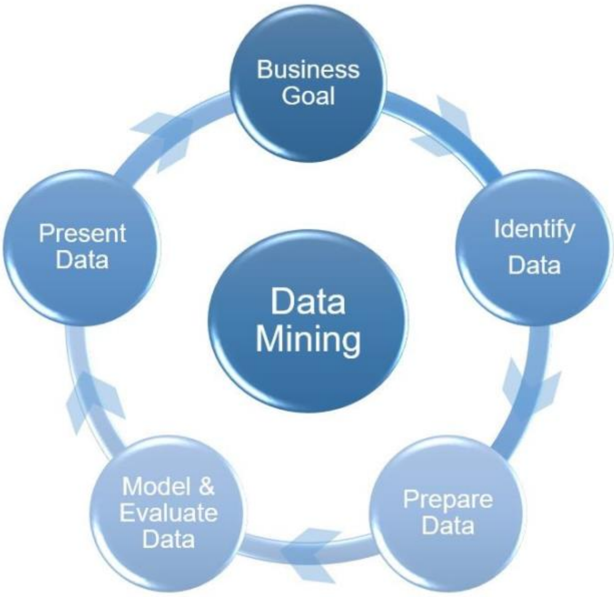
1.0 Introduction

The aim of this report is to produce a data analysis for world of bargains, a supply chain company with 100 shops in the UK. The objective is to identify why some shops are doing better than other shops and identify what factors are driving the profits. This report should also help Ivor Buquetlowd, the owner of world of bargains to build a computer program that takes input and predicts how much money each shop would make.

Using the Weka software, I have predicted the revenue by using Correlation Based Feature Selection with the 'CorrelationAttributeEval' technique. This technique requires Ranker search method. Correlation is more formally referred to as Pearson's correlation coefficient in statistics. To predict the revenue of the shops I have used 'Multilayer perceptron' technique. For classification task I have used 'decision trees – J48' technique. I also used 'cfsSubset' technique for profit prediction and this technique use 'GreedyStepwise' search method.

Data mining is used to discover the patterns in large data sets and database system. Data mining technique can be used to analyse the world bargains data and predict shop revenues. For this task, data mining helps to develop a computer program which helps to manage the choice of new shop location and provide an estimation through prediction or classification. So, data mining process is the best and well suitable for this task.

Data mining process



1.1 Data Summary

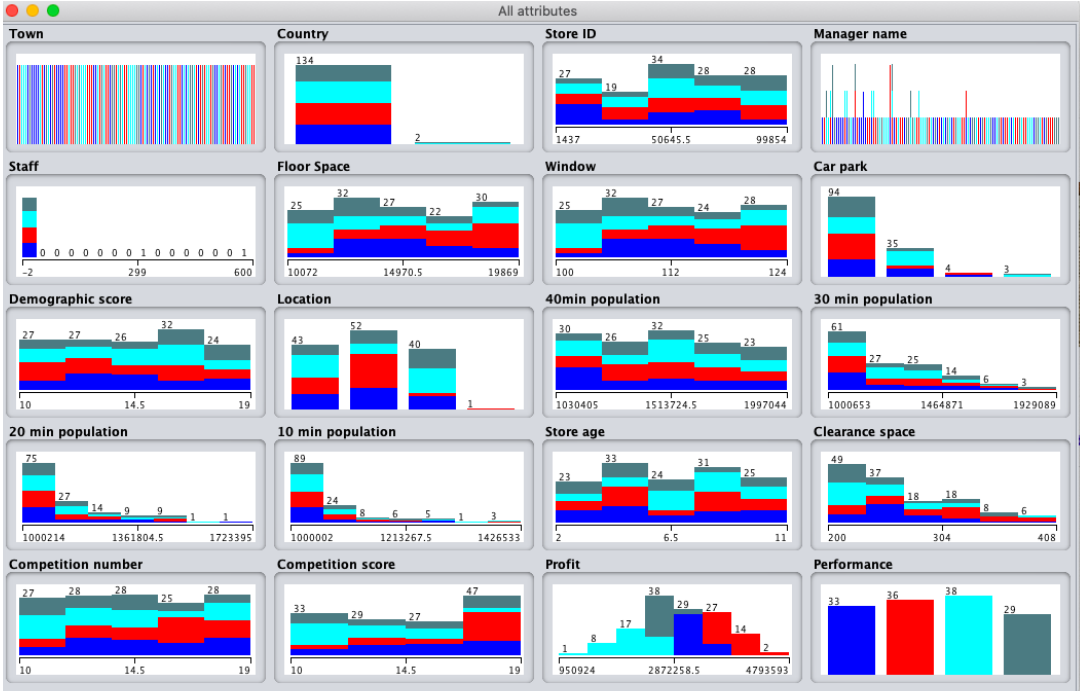
The given store data is in “comma-separated values”(csv) containing 136 observations and 20 attributes. So, in the data many attributes are self-explanatory, but some attributes are hard to understand like:

Attributes	Data Type	Description
Town	Nominal	It is a place, where stores are available.
Country	Nominal	It’s a geographic location
Store ID	Numeric	It is a unique number of your store account
Manger name	Nominal	Name of the manger for each store
Staff	Numeric	Number of the staff
Floor Space	Numeric	The space within the shop e.g. square meters or square foots of total shop area.
Window	Numeric	The showcasing space of the products.
Car park	Nominal	Weather a store has a car park or not.
Demographic	Numeric	It is the scale of the score e.g. 0 to 100 or 0 to 5 related to the structure of population.
Location	Nominal	It is a place of the shop
40min population	Numeric	From the shop this is the number of populations within the radius of particular shop location.
30min population	Numeric	From the shop this is the number of populations within the radius of particular shop location.
20min population	Numeric	From the shop this is the number of populations within the radius of particular shop location.
10min population	Numeric	From the shop this is the number of populations within the radius of particular shop location.
Store age	Numeric	Time since the store established
Clearance space	Numeric	It is a sale, which a larger number of items are discounted or reduced price e.g. square meters or square foot to place all sale items in that area.
Competition number	Numeric	Number of competitors
Competition score	Numeric	It’s the scale of the competitor’s e.g. o to 100 or 0 to 10
Profit	Numeric	Revenue of the stores
Performance	Nominal	Performance of the store

2.0 Data Preparation

The steps below describe the pre-processing of the data and Step 1 is done manually using Microsoft excel.

Pre-processing Variable Values

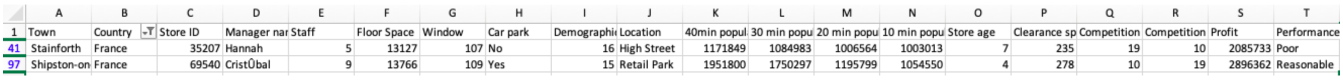


2.1 Step 1 – Problems Found and Fixed

I have found 4 problems in the “store.csv” file.

1. Country

As per the requirements it was mentioned that there are shops in the UK but in the given data there are 2 stores in the France, this might be a data entry issue. So, I assumed these 2 stores are also belongs to UK. So, to fix this I amended France with UK.

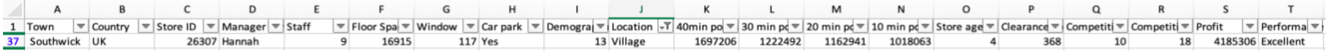


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Town	Country	Store ID	Manager nar	Staff	Floor Space	Window	Car park	Demographi	Location	40min popul	30 min popu	20 min popu	10 min popu	Store age	Clearance sp	Competition	Competition	Profit	Performance
41	Stainforth	France	35207	Hannah	5	13127	107	No	16	High Street	1171849	1084983	1006564	1003013	7	235	19	10	2085733	Poor
97	Shipston-on	France	69540	CristÛbal	9	13766	109	Yes	15	Retail Park	1951800	1750297	1195799	1054550	4	278	10	19	2896362	Reasonable

2. Location

There are 4 locations in the data.
1. High street
2. Retail park
3. Shopping centre
4. Village

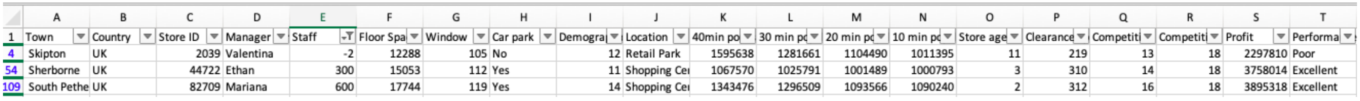
- There are 40 shops in High street, in the Retail park there are 43 shops, 52 shops in Shopping centre and 1 shop in village. The shop entry was confirmed geographically to be located on a High Street in a small fishing village. So, I amended the village to High Street.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Town	Country	Store ID	Manager	Staff	Floor Spa	Window	Car park	Demogra	Location	40min po	30 min pc	20 min pc	10 min pc	Store age	Clearance	Competiti	Competiti	Profit	Performa
37	Southwick	UK	26307	Hannah	9	16915	117	Yes	13	Village	1697206	1222492	1162941	1018063	4	368	10	18	4185306	Excellent

3. Staff

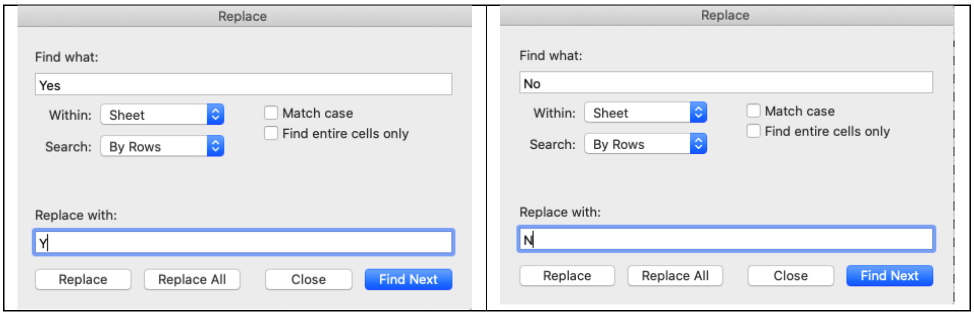
In the staff I found 3 errors, one with the value of -2 and I have corrected this with the positive value 2, one with the value of 300 and one with a value of 600 these are clear errors and I have deleted these rows from the data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Town	Country	Store ID	Manager	Staff	Floor Spa	Window	Car park	Demogra	Location	40min po	30 min pc	20 min pc	10 min pc	Store age	Clearance	Competiti	Competiti	Profit	Performa
4	Skipton	UK	2039	Valentina	-2	12288	105	No	12	Retail Park	1595638	1281661	1104490	1011395	11	219	13	18	2297810	Poor
54	Sherborne	UK	44722	Ethan	300	15053	112	Yes	11	Shopping Cei	1067570	1025791	1001489	1000793	3	310	14	18	3758014	Excellent
109	South Pethe	UK	82709	Mariana	600	17744	119	Yes	14	Shopping Cei	1343476	1296509	1093566	1090240	2	312	16	18	3895318	Excellent

4. Car park

The given data for car park is inconsistent, sometimes it says Yes/No and sometimes it is Y/N. So, to make it consistent across all the data fields I replaced all Yes to Y and all No to N.



Replace

Find what: Yes

Within: Sheet

Search: By Rows

☐ Match case

☐ Find entire cells only

Replace with: Y

Replace

Find what: No

Within: Sheet

Search: By Rows

☐ Match case

☐ Find entire cells only

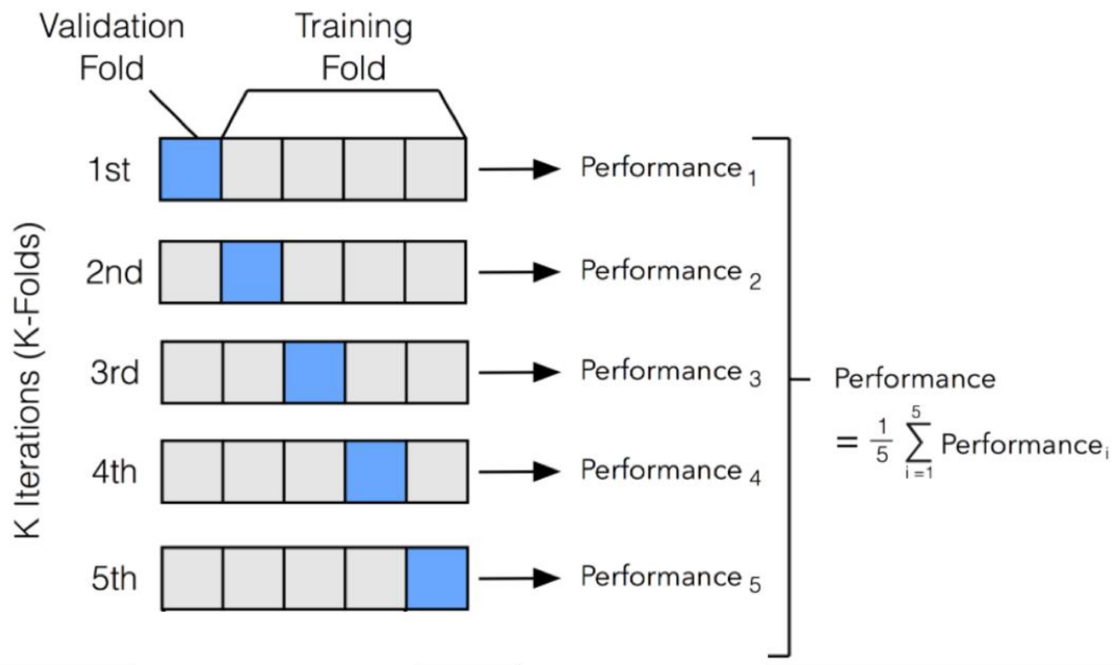
Replace with: N

2.2 Step 2 – Splitting the data into 2 parts Training and Test

Using Weka Cross-validation I have divided the data into 2 sets: training set and test set. This is to make sure the data built not only to execute well on training data but also on unseen data. Basically, training data is an initial set of data that is used to understand how a program apply technologies like neutral networks(NN) and produce some advanced results.

"K-Fold Cross Validation" is one of the most common techniques we use for model evaluation and model selection. For larger training data, k is commonly chosen as 10 and for smaller training data, k is chosen as 5. As you can see the figure below which illustrates the process of 5-fold cross validation.

K-fold cross-validation



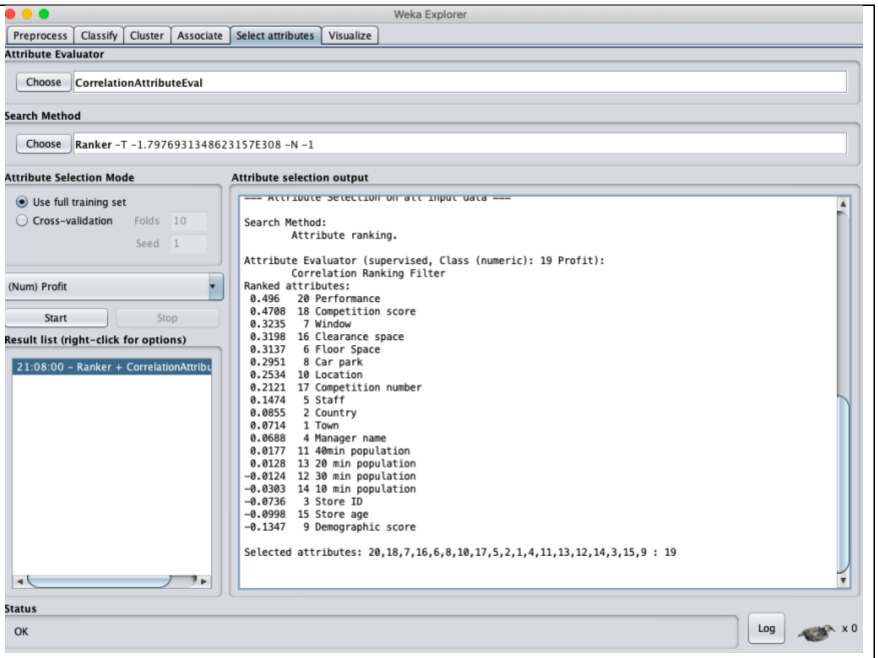
3.0 Model Training

Variable I chose to Investigate

I have chosen these variables using Weka software/Select attributes. The Select attributes are divided in two parts:

1. Attribute Evaluator
2. Search Method

So, I have chosen a technique called Correlation Based Feature Selection. This technique gives the ranked attributes according to profit.

1. Performance 2. Competition score 3. Window 4. Clearance space 5. Floor space 6. Car park 7. Location 8. Competition number 9. Staff	

Attribute Evaluator: CorrelationAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set
○ Cross-validation Folds 10 Seed 1

(Num) Profit

Result list (right-click for options)

21:08:00 - Ranker + CorrelationAttribu

Attribute selection output

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 19 Profit):
        Correlation Ranking Filter
Ranked attributes:
 0.496   20 Performance
 0.4708  18 Competition score
 0.3235   7 Window
 0.3198  16 Clearance space
 0.3137   6 Floor Space
 0.2951   8 Car park
 0.2534  10 Location
 0.2121  17 Competition number
 0.1474   5 Staff
 0.0855   2 Country
 0.0714   1 Town
 0.0688   4 Manager name
 0.0177  11 40min population
 0.0128  13 20 min population
-0.0124  12 30 min population
-0.0303  14 10 min population
-0.0736   3 Store ID
-0.0998  15 Store age
-0.1347   9 Demographic score

Selected attributes: 20,18,7,16,6,8,10,17,5,2,1,4,11,13,12,14,3,15,9 : 19
```

Status

OK

4.0 Technique Description

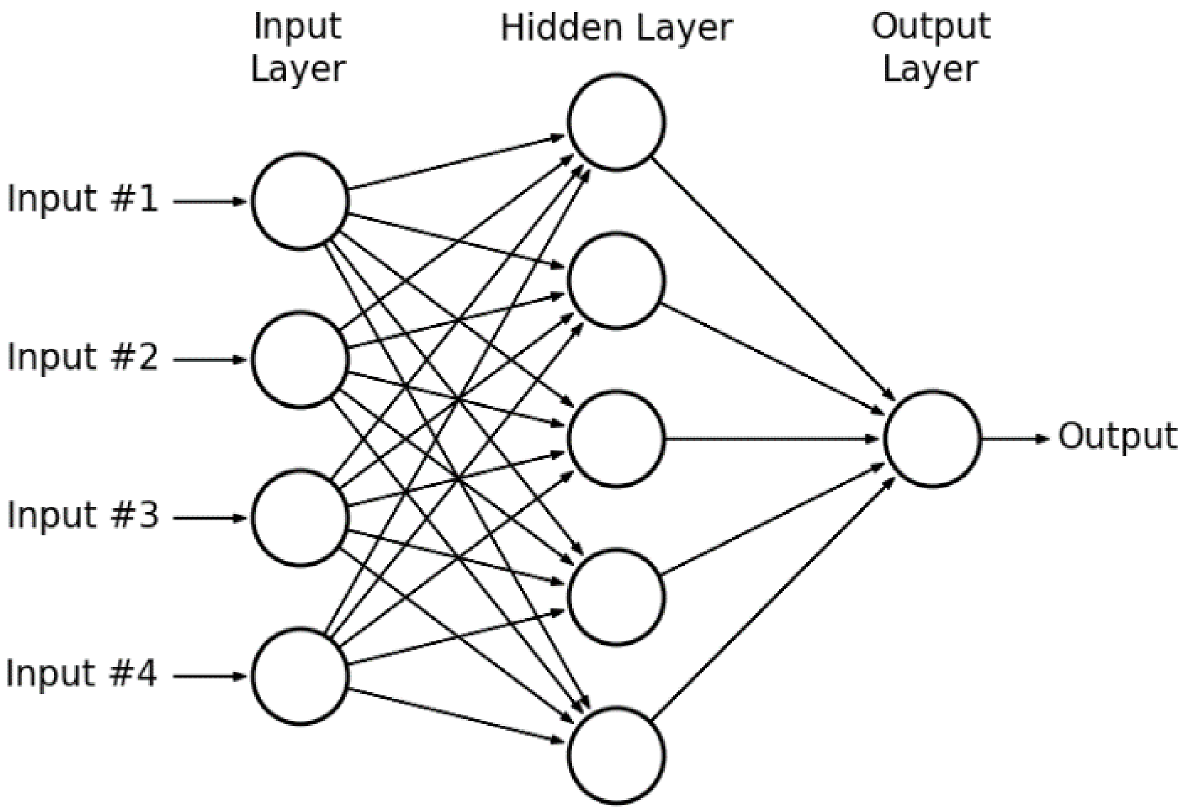
Multi-layer perceptron(MLP)

Multi-layer perceptron (MLP) can be used for prediction and classification tasks, also MLP is called as neutral networks(NN). An MLP consist of 3 layers of nodes:

1. input layer
2. hidden layer
3. output layer

Each node is neutron which uses a non-linear function except the input node. Each node connects with a certain weight to every node. As you can see figure below, it's an example of MLP. There are 4 input layer, 5 hidden layers and 1 output layer. So, Input node receives a value for a given instance in the data and passes the value through the node as output.

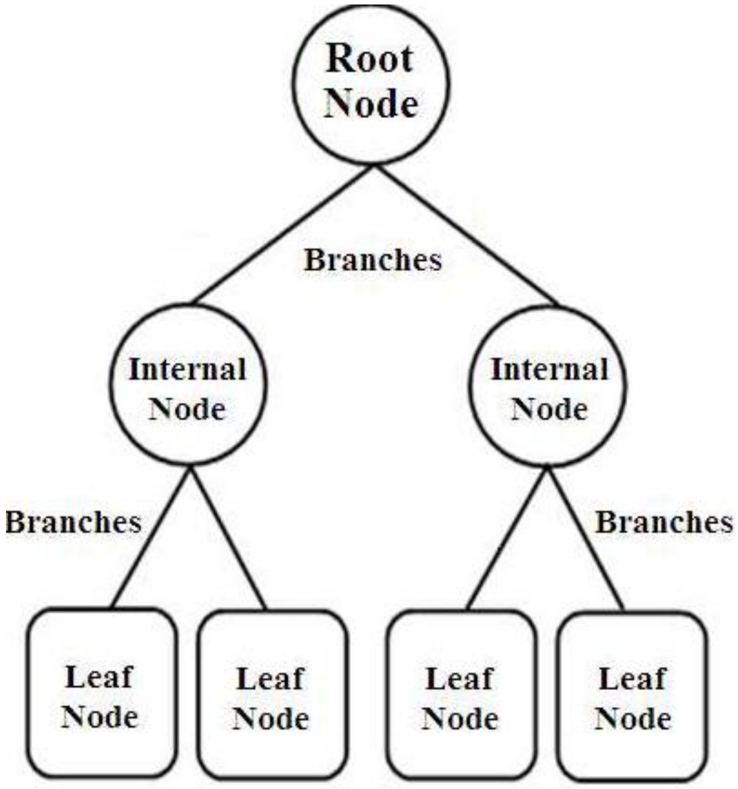
Multilayer Perceptron Network



Decision trees

A decision tree is a structure that includes a root node, branches and leaf nodes, where a node represents a single variable. Every Internal node denote a test on an attribute, branch node denotes the outcome of a test and leaf nodes represent the classified/predicted variable and the top most node is root node. As you can see figure below, is the structure of decision trees.

Structure of decision tree

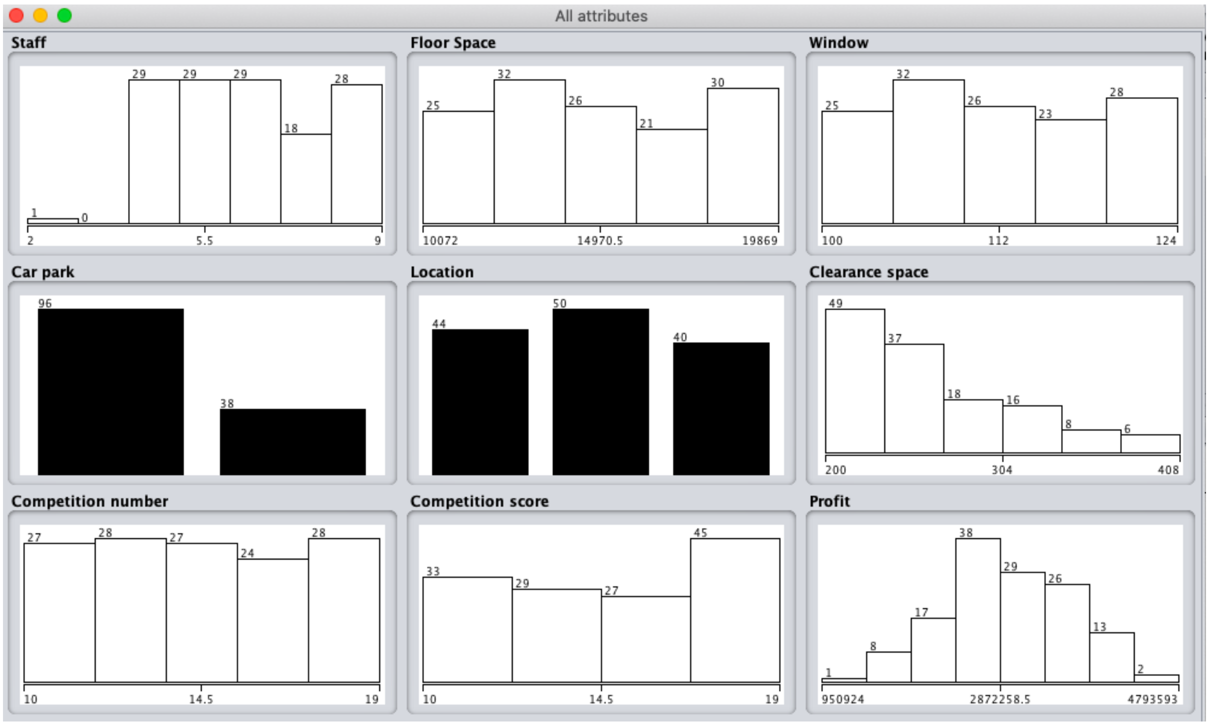


Many decision trees use ID3 algorithm which use entropy to calculate the homogeneity of a sample.

5.0 Predicted results(Profit)

In the figure below, the variable values distribution post-processing for the profit prediction.

Chosen Variable values for Prediction



Matrix X represents all the attributes and our targeted variable Y which is Profit.

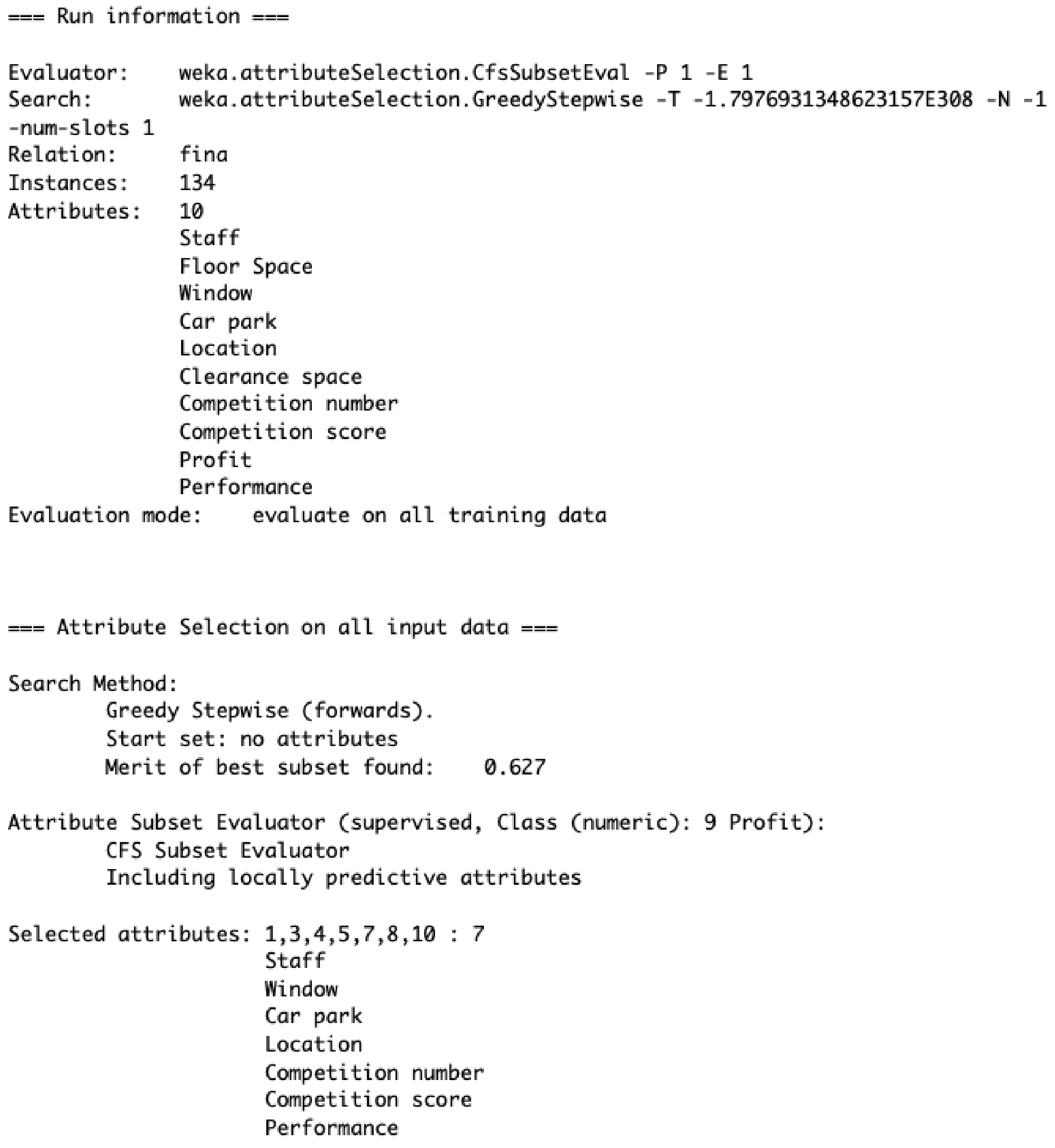
Example:

X = [Staff, Floor Space, Window, Car Park, Demographic Score, Location, 40min population, 30min population, 20min population, 10min population, Store age, Clearance Space, Competition Score]

y = [Profit]

For prediction task the training data was loaded via “Select Attributes” tab In the Attribute Evaluator of Weka software. I have chosen ‘cfcSubsetEval’ technique identifies the most effective attributes by considering the individual predictive and with the degree of redundancy between them. If we select ‘cfcSubsetEval’, Weka will give a pop-up alert window saying in the search method should use ‘Greedy Stepwise’ method. So, Weka detects automatically which is relevant to the attribute. I have used ‘Greedy Stepwise’ as full training set containing all attributes in it, results can be seen below:

Prediction results



```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.GreedyStepwise -T -1.7976931348623157E308 -N -1
-num-slots 1
Relation:     fina
Instances:    134
Attributes:   10
              Staff
              Floor Space
              Window
              Car park
              Location
              Clearance space
              Competition number
              Competition score
              Profit
              Performance
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.627

Attribute Subset Evaluator (supervised, Class (numeric): 9 Profit):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,3,4,5,7,8,10 : 7
                     Staff
                     Window
                     Car park
                     Location
                     Competition number
                     Competition score
                     Performance
```

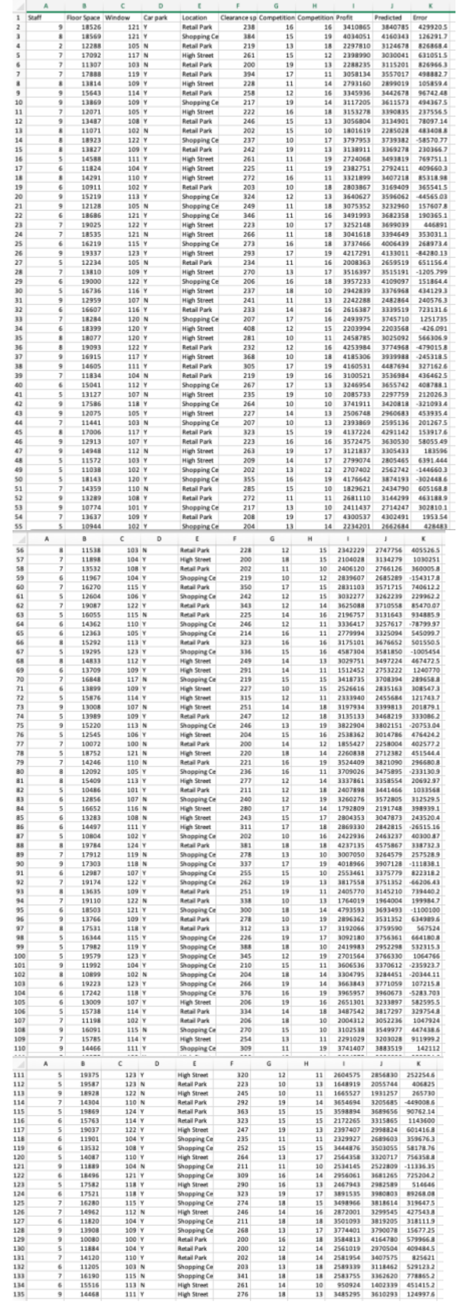
5.1 Multilayer Perceptron(MLP)

I have loaded the data in training set in the Weka classify. I have chosen Multilayer Perceptron technique to predict the profit and these are the results.

You can see in the 2 new columns in the figure saying ‘Predicted’ and ‘Error’

In the predicted column it shows the predicted profit.

In the Error column it shows the difference between the given profit and what are actually predicted.



	A	B	C	D	E	F	G	H	I	J	K
1	Staff	Floor Space	Window	Car park	Location	Clearance sp	Competition	Competition	Profit	Predicted	Error
2	9	18526	121	Y	Retail Park	238	16	16	3410865	3840785	429920.5
3	8	18569	121	Y	Shopping Ce	384	15	19	4034051	4160343	126291.7
4	2	12288	105	N	Retail Park	219	13	18	2297810	3124678	826868.4
5	7	17092	117	N	High Street	261	15	12	2398990	3030041	631051.5
6	7	11307	103	N	Retail Park	200	19	13	2288235	3115201	826966.3
7	7	17888	119	Y	Retail Park	394	17	11	3058134	3557017	498882.7
8	8	13814	109	Y	High Street	228	11	14	2793160	2899019	105859.4
9	9	15643	114	Y	Retail Park	258	12	16	3345936	3442678	96742.48
10	9	13869	109	Y	Shopping Ce	217	19	14	3117205	3611573	494367.5
11	7	12071	105	Y	High Street	222	16	18	3153278	3390835	237556.5
12	9	13487	108	Y	Retail Park	246	15	13	3056804	3134901	78097.14
13	8	11071	102	N	Retail Park	202	15	10	1801619	2285028	483408.8
14	8	18923	122	Y	Shopping Ce	237	10	17	3797953	3739382	-58570.77
15	8	13827	109	Y	Retail Park	242	19	13	3138911	3369278	230366.7
16	5	14588	111	Y	High Street	261	11	19	2724068	3493819	769751.1
17	6	11824	104	Y	High Street	225	11	19	2382751	2792411	409660.3
18	8	14291	110	Y	High Street	272	16	11	3321899	3407218	85318.98
19	6	10911	102	Y	Retail Park	203	10	18	2803867	3169409	365541.5
20	9	15219	113	Y	Shopping Ce	324	12	13	3640627	3596062	-44565.03
21	9	12128	105	N	Shopping Ce	249	11	18	3075352	3232960	157607.8
22	6	18686	121	Y	Shopping Ce	346	11	16	3491993	3682358	190365.1
23	7	19025	122	Y	High Street	223	10	17	3252148	3699039	446891
24	7	18535	121	N	High Street	266	11	18	3041618	3394649	353031.1
25	6	16219	115	Y	Shopping Ce	273	16	18	3737466	4006439	268973.4
26	9	19337	123	Y	High Street	293	17	19	4217291	4133011	-84280.13
27	5	12234	105	N	Retail Park	234	11	16	2008363	2659519	651156.4
28	7	13810	109	Y	High Street	270	13	17	3516397	3515191	-1205.799
29	6	19000	122	Y	Shopping Ce	206	16	18	3957233	4109097	151864.4
30	5	16736	116	Y	High Street	237	18	10	2942839	3376968	434129.3
31	9	12959	107	N	High Street	241	11	13	2242288	2482864	240576.3
32	6	16607	116	Y	Retail Park	233	14	16	2616387	3339519	723131.6
33	7	18284	120	N	Shopping Ce	207	17	16	2493975	3745710	1251735
34	6	18399	120	Y	High Street	408	12	15	2203994	2203568	-426.091
35	8	18077	120	Y	High Street	281	10	11	2458785	3025092	566306.9
36	8	19093	122	Y	Retail Park	232	12	16	4253984	3774968	-479015.8
37	9	16915	117	Y	High Street	368	10	18	4185306	3939988	-245318.5
38	9	14605	111	Y	Retail Park	305	17	19	4160531	4487694	327162.6
39	7	11834	104	N	Retail Park	219	19	16	3100521	3536984	436462.5
40	6	15041	112	Y	Shopping Ce	267	17	13	3246954	3655742	408788.1
41	5	13127	107	N	High Street	235	19	10	2085733	2297759	212026.3
42	9	17586	118	Y	Shopping Ce	264	10	10	3741911	3420818	-321093.4
43	9	12075	105	Y	High Street	227	14	13	2506748	2960683	453935.4
44	7	11441	103	N	Shopping Ce	207	10	13	2393869	2595136	201267.5
45	8	17006	117	Y	Retail Park	323	15	19	4137224	4291142	153917.6
46	9	12913	107	Y	Retail Park	223	16	16	3572475	3630530	58055.49
47	9	14948	112	N	High Street	263	19	17	3121837	3305433	183596
48	5	11572	103	Y	High Street	209	14	17	2799074	2805465	6391.444
49	5	11038	102	Y	Shopping Ce	202	13	12	2707402	2562742	-144660.3
50	5	18143	120	Y	Shopping Ce	355	16	19	4176642	3874193	-302448.6
51	7	14359	110	N	Retail Park	285	15	10	1829621	2434790	605168.8
52	9	13289	108	Y	Retail Park	272	11	11	2681110	3144299	463188.9
53	9	10774	101	Y	Shopping Ce	217	13	10	2411437	2714247	302810.1
54	7	13637	109	Y	Retail Park	208	19	17	4300537	4302491	1953.54
55	5	10944	102	Y	Shopping Ce	204	13	14	2234201	2662684	428483
56	8	11538	103	N	Retail Park	228	12	15	2342229	2747756	405526.5
57	7	11898	104	Y	High Street	200	18	15	2104028	3134279	1030251
58	7	13532	108	Y	Retail Park	202	11	10	2406120	2766126	360005.8
59	6	11967	104	Y	Shopping Ce	219	10	12	2839607	2685289	-154317.8
60	7	16270	115	Y	Retail Park	350	17	15	2831103	3571715	740612.2
61	5	12604	106	Y	Shopping Ce	242	12	15	3032277	3262239	229962.2
62	7	19087	122	Y	Retail Park	343	12	14	3625088	3710558	85470.07
63	5	16055	115	N	Retail Park	225	14	16	2196757	3131643	934885.9
64	6	14362	110	Y	Shopping Ce	246	12	11	3336417	3257617	-78799.97
65	6	12363	105	Y	Shopping Ce	214	16	11	2779994	3325094	545099.7
66	8	15292	113	Y	Retail Park	323	16	16	3175101	3676652	501550.5
67	5	19295	123	Y	Shopping Ce	336	15	16	4587304	3581850	-1005454
68	8	14833	112	Y	High Street	249	14	13	3029751	3497224	467472.5
69	6	13709	109	Y	High Street	291	14	11	1512452	2753222	1240770
70	7	16848	117	N	Shopping Ce	219	15	15	3418735	3708394	289658.8
71	6	13899	109	Y	High Street	227	10	15	2526616	2835163	308547.3
72	5	15876	114	Y	High Street	315	12	11	2333940	2455684	121743.7
73	9	13008	107	N	High Street	251	14	18	3197934	3399813	201879.1
74	5	13989	109	Y	Retail Park	247	12	18	3135133	3468219	333086.2
75	9	15220	113	N	Shopping Ce	246	13	19	3822904	3802151	-20753.04
76	5	12545	106	Y	High Street	204	15	16	2538362	3014786	476424.2
77	7	10072	100	N	Retail Park	200	14	12	1855427	2258004	402577.2
78	5	18752	121	N	High Street	220	18	14	2260838	2712382	451544.4
79	7	14246	110	N	Retail Park	221	16	19	3524409	3821090	296680.8
80	8	12092	105	Y	Shopping Ce	236	16	11	3709026	3475895	-233130.9
81	8	15409	113	Y	High Street	277	12	14	3337861	3358554	20692.97
82	5	10486	101	Y	Retail Park	211	12	18	2407898	3441466	1033568
83	6	12856	107	N	Shopping Ce	240	12	19	3260276	3572805	312529.5
84	5	16652	116	N	High Street	280	17	14	1792809	2191748	398939.1
85	6	13283	108	N	High Street	243	15	17	2804353	3047873	243520.4
86	6	14497	111	Y	High Street	311	17	18	2869330	2842815	-26515.16
87	5	10804	102	Y	Shopping Ce	202	10	16	2422936	2463237	40300.87
88	8	19784	124	Y	Retail Park	381	18	18	4237135	4575867	338732.3
89	7	17912	119	N	Shopping Ce	278	13	10	3007050	3264579	257528.9
90	9	17303	118	N	Shopping Ce	337	17	19	4018966	3907128	-111838.1
91	6	12987	107	Y	Shopping Ce	255	15	10	2553461	3375779	822318.2
92	7	19174	122	Y	Shopping Ce	262	19	13	3817558	3751352	-66206.43
93	8	13635	109	Y	Retail Park	251	19	11	2405770	3145210	739440.2
94	7	19110	122	N	Retail Park	338	10	13	1764019	1964004	199984.7
95	6	18503	121	Y	Shopping Ce	300	18	14	4793593	3693493	-1100100
96	9	13766	109	Y	Retail Park	278	10	19	2896362	3531352	634989.6
97	8	17531	118	Y	Retail Park	312	13	17	3192066	3759590	567524
98	5	16344	115	Y	Shopping Ce	226	19	17	3092180	3756361	664180.8
99	5	17982	119	Y	Shopping Ce	388	18	10	2419983	2952298	532315.3
100	5	19579	123	Y	Shopping Ce	345	12	19	2701564	3766330	1064766
101	9	11992	104	Y	Shopping Ce	210	15	11	3606536	3370612	-235923.7
102	8	10899	102	N	Shopping Ce	204	18	14	3304795	3284451	-20344.11
103	6	19223	123	Y	Shopping Ce	266	19	14	3663843	3771059	107215.8
104	6	17242	118	Y	Shopping Ce	376	16	19	3965957	3960673	-5283.703
105	6	13009	107	Y	High Street	206	19	16	2651301	3233897	582595.5
106	5	15738	114	Y	Retail Park	334	14	18	3487542	3817297	329754.8
107	7	11198	102	Y	Retail Park	206	18	10	2004312	3052236	1047924
108	9	16091	115	N	Shopping Ce	270	15	10	3102538	3549977	447438.6
109	7	15785	114	Y	High Street	254	13	11	2291029	3203028	911999.2
110	9	14466	111	Y	Shopping Ce	309	11	19	3741407	3883519	142112
111	5	19375	123	Y	High Street	320	12	11	2604575	2856830	252254.6
112	5	19587	123	N	Retail Park	223	10	13	1648919	2055744	406825
113	9	18928	122	N	High Street	245	10	11	1665527	1931257	265730
114	7	14304	110	N	Retail Park	292	19	14	3654694	3205685	-449008.6
115	5	19869	124	Y	Retail Park	363	15	15	3598894	3689656	90762.14
116	6	15763	114	Y	Retail Park	323	15	15	2172265	3315865	1143600
117	5	19037	122	Y	High Street	247	19	13	2397407	2998824	601416.8
118	6	11901	104	Y	Shopping Ce	235	11	11	2329927	2689603	359676.3
119	6	13532	108	Y	Shopping Ce	252	15	15	3444876	3503055	58178.76
120	5	14087	110	Y	High Street	264	13	17	2564358	3320717	756358.8
121	9	11889	104	N	Shopping Ce	211	11	10	2534145	2522809	-11336.35
122	6	18496	121	Y	Shopping Ce	309	16	14	2956061	3681265	725204.2
123	5	17582	118	Y	High Street	290	16	13	2467943	2982589	514646
124	6	17521	118	Y	Shopping Ce	323	19	17	3891535	3980803	89268.08
125	7	16280	115	Y	Shopping Ce	274	18	15	3498966	3818614	319647.5
126	7	14962	112	N	High Street	246	14	16	2872001	3299545	427543.8
127	6	11820	104	Y	Shopping Ce	211	18	18	3501093	3819205	318111.9
128	9	13908	109	Y	Shopping Ce	268	13	17	3774401	3790078	15677.25
129	9	10080	100	Y	Retail Park	200	16	18	3584813	4164780	579966.8
130	5	11884	104	Y	Retail Park	200	12	14	2561019	2970504	409484.5
131	7	14120	110	Y	Retail Park	202	18	14	2581954	3407575	825621
132	6	11205	103	N	Shopping Ce	203	13	18	2589339	3118462	529123.2
133	7	16190	115	N	Shopping Ce	341	18	18	2583755	3362620	778865.2
134	6	15516	113	N	High Street	261	14	10	950924	1402339	451415.2
135	9	14468	111	Y	High Street	276	18	13	3485295	3610293	124997.6

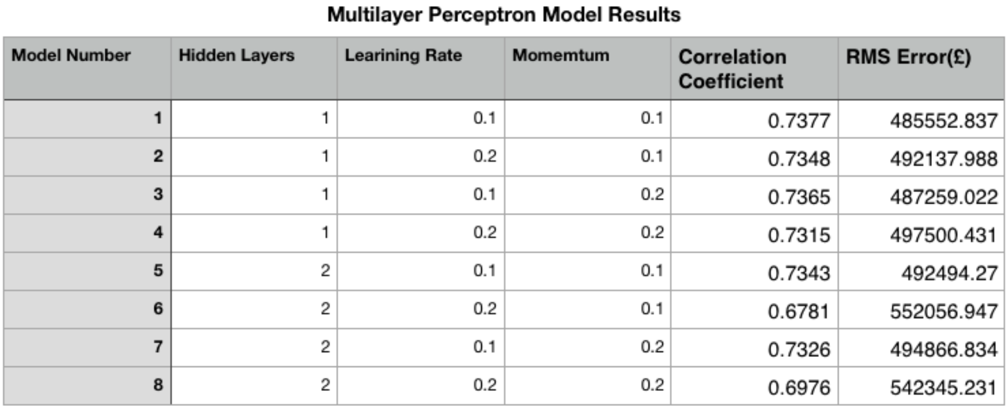
The hyperparameters evaluated the MLP are ‘Number of hidden layers ’, ‘Learning rate’ and ‘Momentum’

Number of hidden layers: This is a process; how many hidden layers are used to find relationships in data. A single hidden layer is more a simple relationship in the data, it’s better to have multiple hidden layers to have more complex relationships in the data.

Learning rate: This value impact how quickly the network learns, so a low learning rate will result slow learning of weights. A high learning rate makes the weight and results in the progressive learning.

Momentum: This influence, the model which can escape local minima during backpropagation.

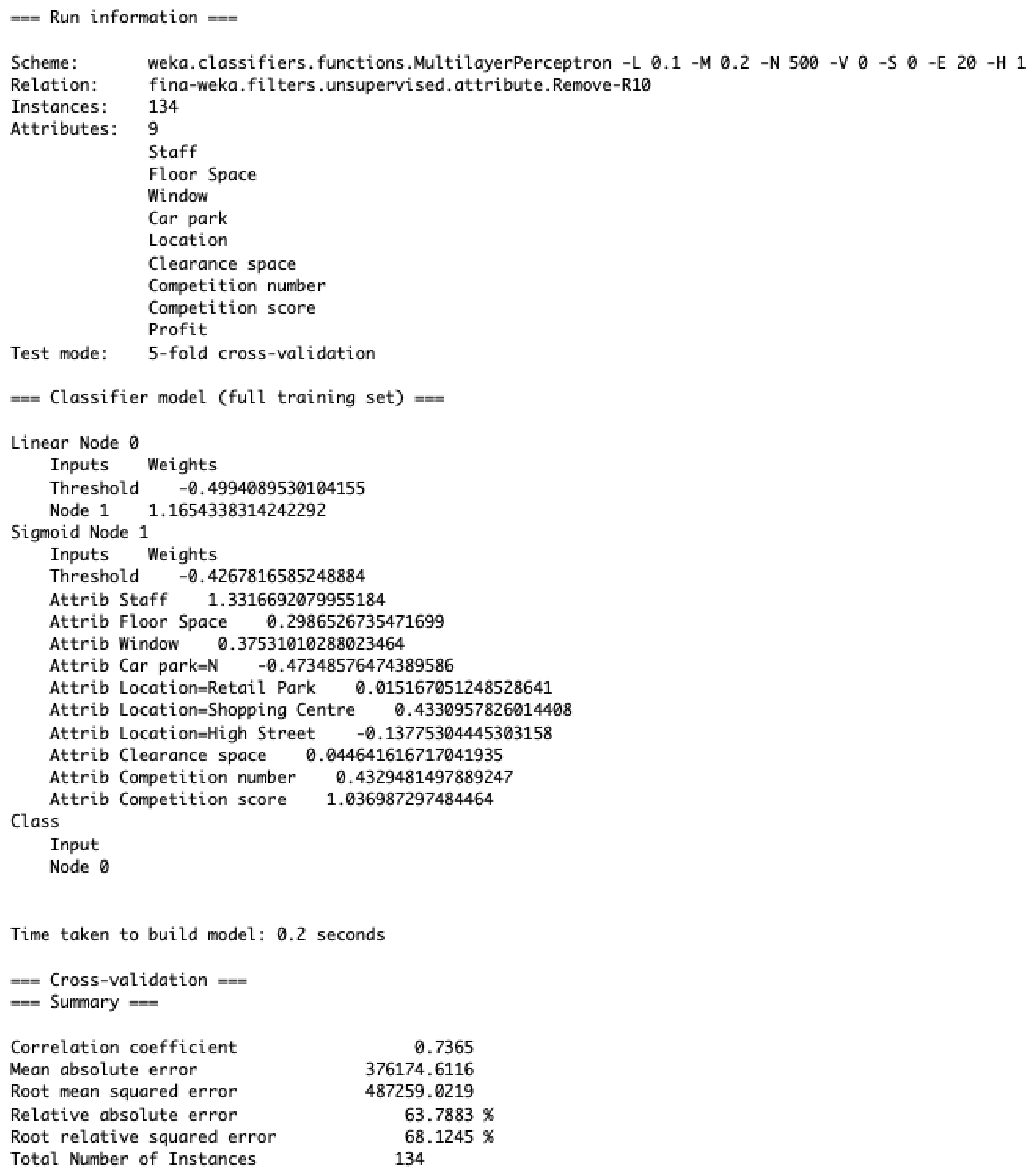
Below are the results MLP model.



Multilayer Perceptron Model Results

Model Number	Hidden Layers	Learining Rate	Momemtum	Correlation Coefficient	RMS Error(£)
1	1	0.1	0.1	0.7377	485552.837
2	1	0.2	0.1	0.7348	492137.988
3	1	0.1	0.2	0.7365	487259.022
4	1	0.2	0.2	0.7315	497500.431
5	2	0.1	0.1	0.7343	492494.27
6	2	0.2	0.1	0.6781	552056.947
7	2	0.1	0.2	0.7326	494866.834
8	2	0.2	0.2	0.6976	542345.231

Prediction model



```
=== Run information ===

Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.1 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 1
Relation:     fina-weka.filters.unsupervised.attribute.Remove-R10
Instances:    134
Attributes:   9
              Staff
              Floor Space
              Window
              Car park
              Location
              Clearance space
              Competition number
              Competition score
              Profit
Test mode:    5-fold cross-validation

=== Classifier model (full training set) ===

Linear Node 0
    Inputs    Weights
    Threshold    -0.4994089530104155
    Node 1    1.1654338314242292
Sigmoid Node 1
    Inputs    Weights
    Threshold    -0.4267816585248884
    Attrib Staff    1.3316692079955184
    Attrib Floor Space    0.2986526735471699
    Attrib Window    0.37531010288023464
    Attrib Car park=N    -0.47348576474389586
    Attrib Location=Retail Park    0.015167051248528641
    Attrib Location=Shopping Centre    0.4330957826014408
    Attrib Location=High Street    -0.13775304445303158
    Attrib Clearance space    0.044641616717041935
    Attrib Competition number    0.4329481497889247
    Attrib Competition score    1.036987297484464
Class
    Input
    Node 0


Time taken to build model: 0.2 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7365
Mean absolute error                 376174.6116
Root mean squared error             487259.0219
Relative absolute error                 63.7883 %
Root relative squared error             68.1245 %
Total Number of Instances              134
```

6.0 Performance Classification

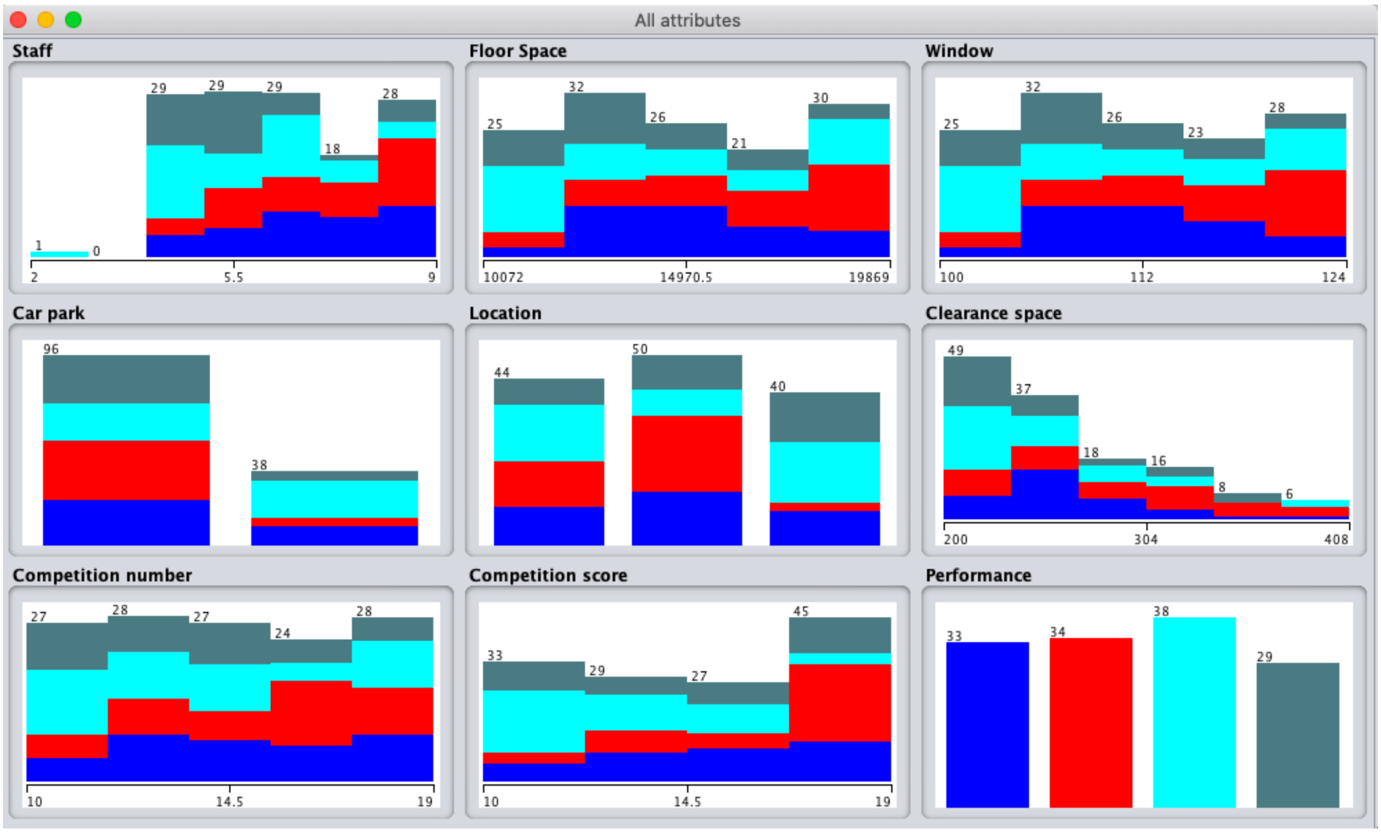
In the figure below, the variable values distribution for the performance classification data.

Example:

X = [Staff, Floor Space, Window, Car Park, Demographic Score, Location, 40min population, 30min population, 20min population, 10min population, Store age, Clearance Space, Competition Score]

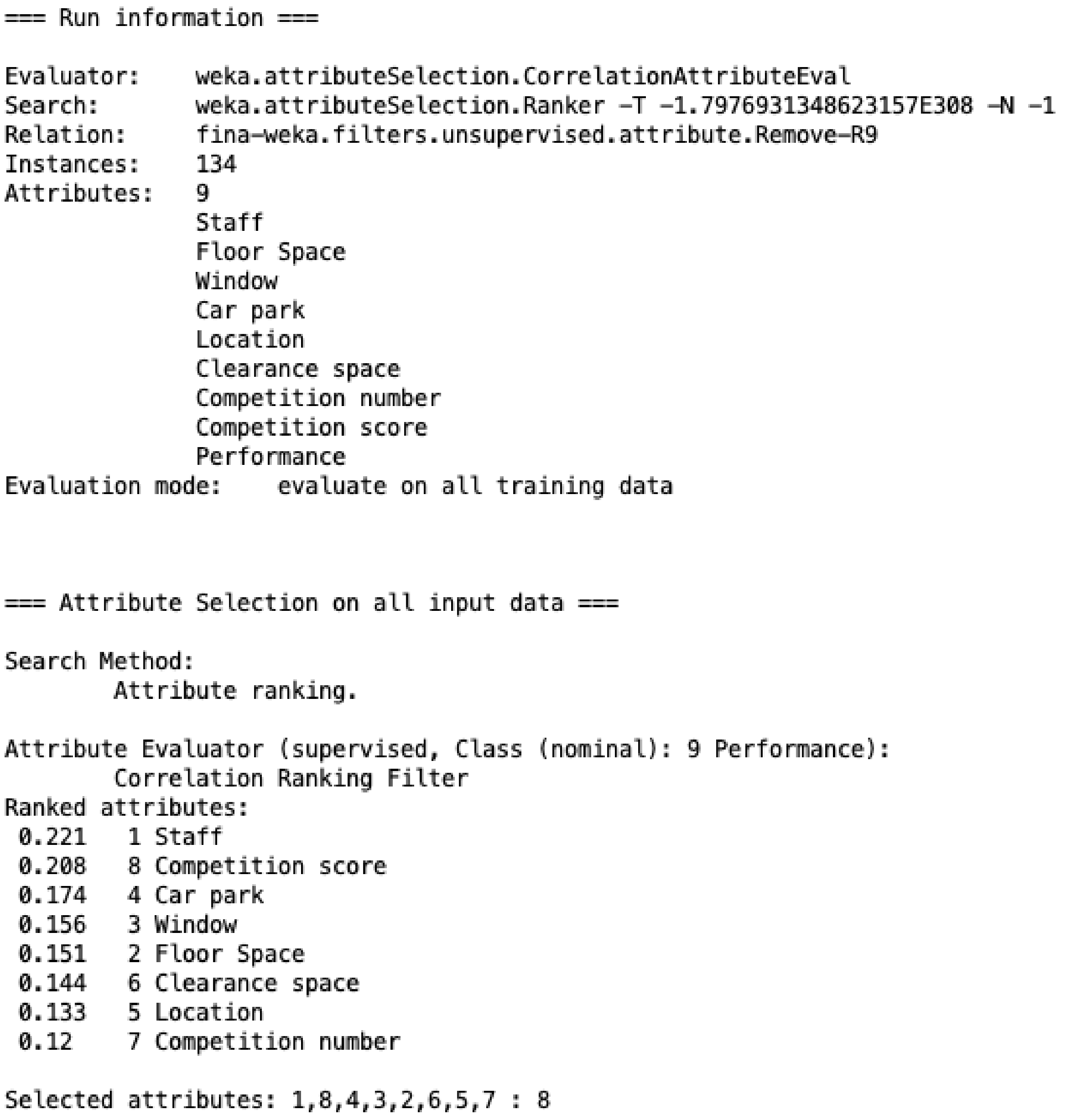
y = [Performance] ∈ {'Poor', 'Reasonable', 'Good', 'Excellent'}

Variable values for classification



In the prediction task above, I used ‘cfcSubsetEval’ and ‘Greedy Stepwise’ method. To perform a selection for classification task I used ‘CorrealtionAttributeEval’ and ‘Ranker’ method, below are the results of the model performance using all the attributes available to the model.

Selection algorithm results of classification



```
=== Run information ===

Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     fina-weka.filters.unsupervised.attribute.Remove-R9
Instances:    134
Attributes:   9
              Staff
              Floor Space
              Window
              Car park
              Location
              Clearance space
              Competition number
              Competition score
              Performance
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 Performance):
        Correlation Ranking Filter
Ranked attributes:
 0.221    1 Staff
 0.208    8 Competition score
 0.174    4 Car park
 0.156    3 Window
 0.151    2 Floor Space
 0.144    6 Clearance space
 0.133    5 Location
 0.12     7 Competition number

Selected attributes: 1,8,4,3,2,6,5,7 : 8
```

6.1 Decision Trees

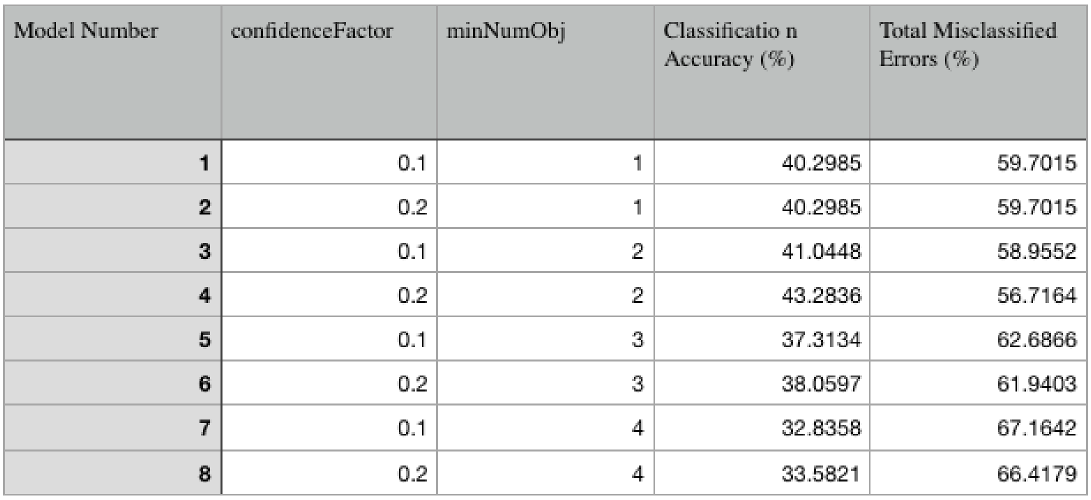
Below table are the results from the decision tree algorithm. I used Weka's J48 decision trees, it is used to build a classification model.

There are 2 main hyperparameters which helped to get the results:

1. confidenceFactor – a small confidence factors will force a high degree of pruning. It helps to reduce the size of the decision tree y removing parts of the tree.

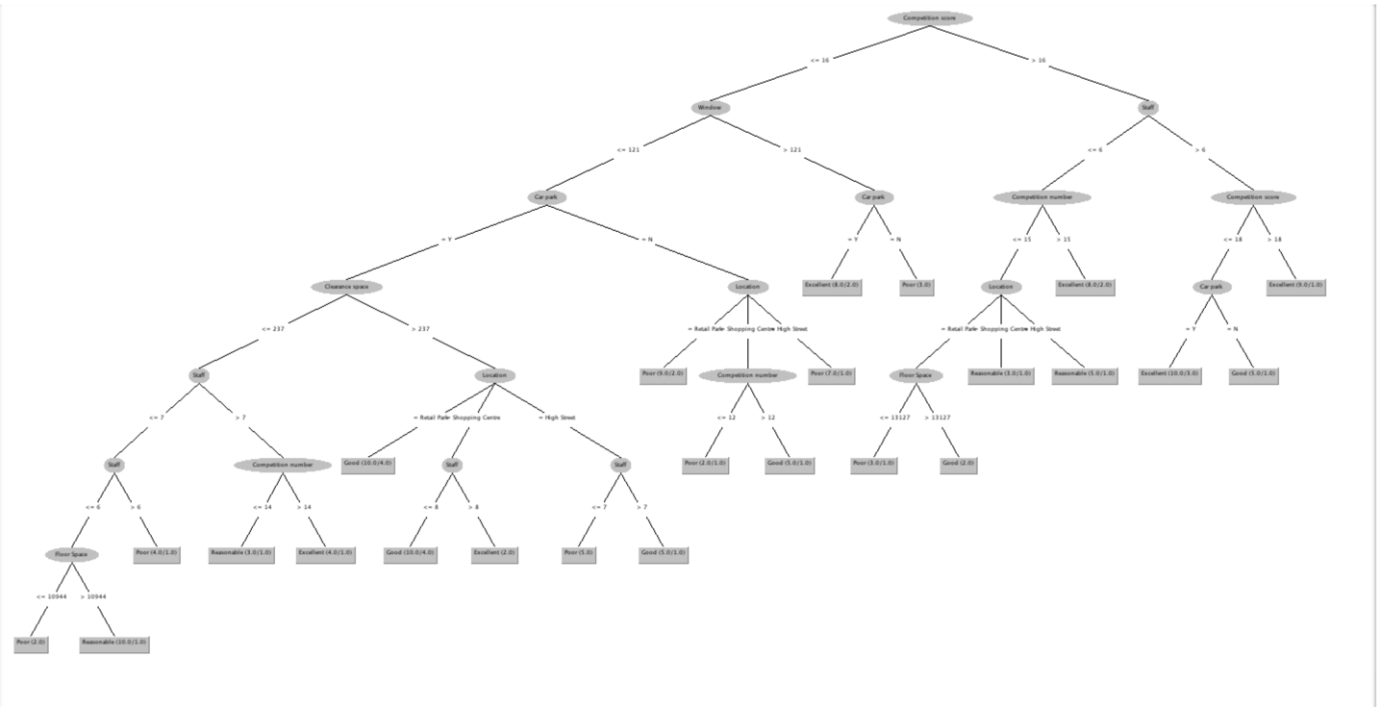
2. minNumObj – The minimum number of instances per leaf.

Decision Tree Model Results

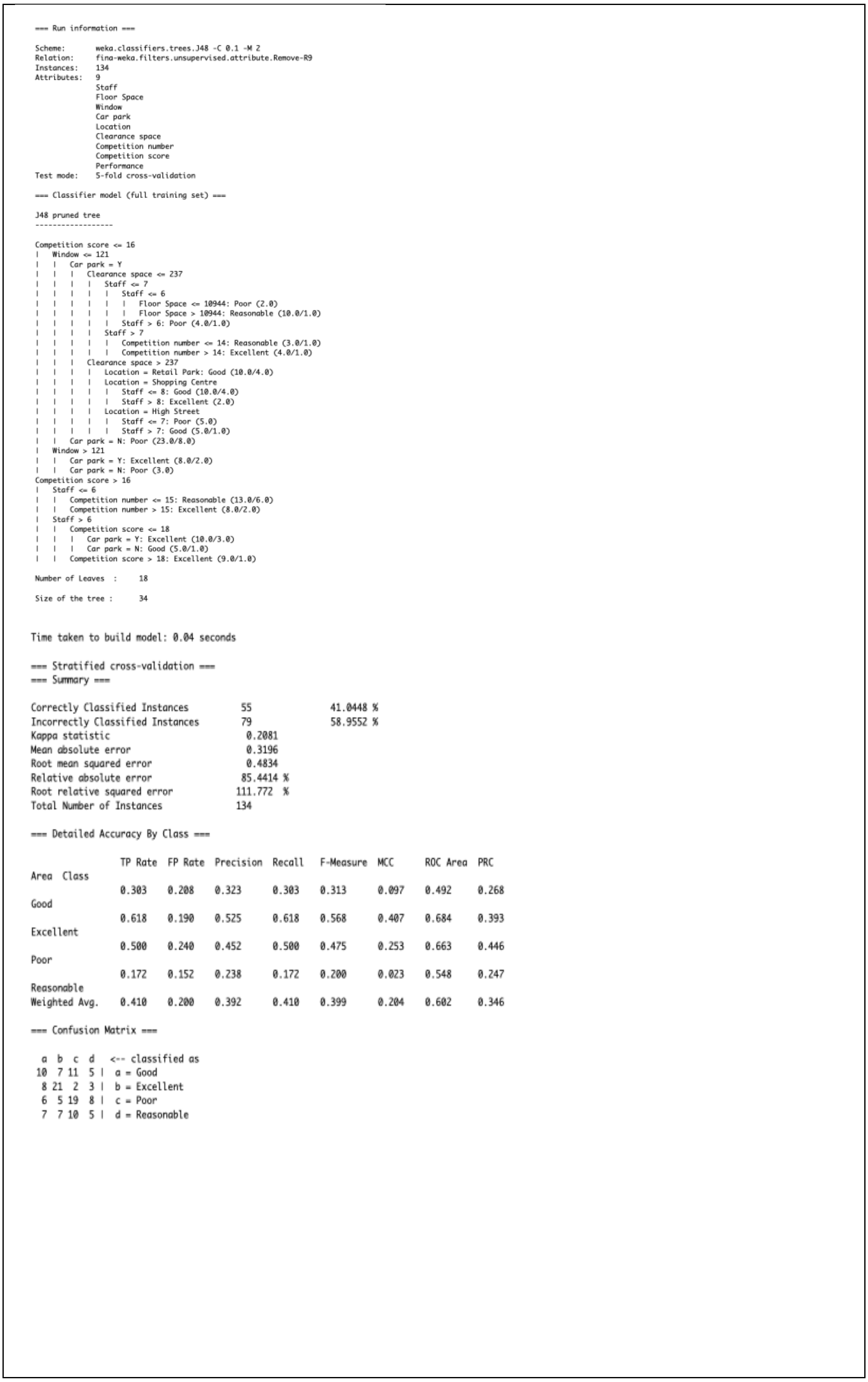


Model Number	confidenceFactor	minNumObj	Classificatio n Accuracy (%)	Total Misclassified Errors (%)
1	0.1	1	40.2985	59.7015
2	0.2	1	40.2985	59.7015
3	0.1	2	41.0448	58.9552
4	0.2	2	43.2836	56.7164
5	0.1	3	37.3134	62.6866
6	0.2	3	38.0597	61.9403
7	0.1	4	32.8358	67.1642
8	0.2	4	33.5821	66.4179

Decision tree visualise



Classification model



```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.1 -M 2
Relation:     fina-weka.filters.unsupervised.attribute.Remove-R9
Instances:    134
Attributes:   9
              Staff
              Floor Space
              Window
              Car park
              Location
              Clearance space
              Competition number
              Competition score
              Performance
Test mode:    5-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Competition score <= 16
|   Window <= 121
|   |   Car park = Y
|   |   |   Clearance space <= 237
|   |   |   |   Staff <= 7
|   |   |   |   |   Staff <= 6
|   |   |   |   |   |   Floor Space <= 10944: Poor (2.0)
|   |   |   |   |   |   Floor Space > 10944: Reasonable (10.0/1.0)
|   |   |   |   |   Staff > 6: Poor (4.0/1.0)
|   |   |   |   Staff > 7
|   |   |   |   |   Competition number <= 14: Reasonable (3.0/1.0)
|   |   |   |   |   Competition number > 14: Excellent (4.0/1.0)
|   |   |   Clearance space > 237
|   |   |   |   Location = Retail Park: Good (10.0/4.0)
|   |   |   |   Location = Shopping Centre
|   |   |   |   |   Staff <= 8: Good (10.0/4.0)
|   |   |   |   |   Staff > 8: Excellent (2.0)
|   |   |   |   Location = High Street
|   |   |   |   |   Staff <= 7: Poor (5.0)
|   |   |   |   |   Staff > 7: Good (5.0/1.0)
|   |   Car park = N: Poor (23.0/8.0)
|   Window > 121
|   |   Car park = Y: Excellent (8.0/2.0)
|   |   Car park = N: Poor (3.0)
Competition score > 16
|   Staff <= 6
|   |   Competition number <= 15: Reasonable (13.0/6.0)
|   |   Competition number > 15: Excellent (8.0/2.0)
|   Staff > 6
|   |   Competition score <= 18
|   |   |   Car park = Y: Excellent (10.0/3.0)
|   |   |   Car park = N: Good (5.0/1.0)
|   |   Competition score > 18: Excellent (9.0/1.0)

Number of Leaves  :     18

Size of the tree :      34


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          55               41.0448 %
Incorrectly Classified Instances        79               58.9552 %
Kappa statistic                          0.2081
Mean absolute error                      0.3196
Root mean squared error                  0.4834
Relative absolute error                 85.4414 %
Root relative squared error            111.772  %
Total Number of Instances              134

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 0.303    0.208    0.323      0.303    0.313      0.097    0.492     0.268
Good
                 0.618    0.190    0.525      0.618    0.568      0.407    0.684     0.393
Excellent
                 0.500    0.240    0.452      0.500    0.475      0.253    0.663     0.446
Poor
                 0.172    0.152    0.238      0.172    0.200      0.023    0.548     0.247
Reasonable
Weighted Avg.    0.410    0.200    0.392      0.410    0.399      0.204    0.602     0.346

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 10  7 11  5 |  a = Good
  8 21  2  3 |  b = Excellent
  6  5 19  8 |  c = Poor
  7  7 10  5 |  d = Reasonable
```

Recommendation

Based on the results produced by the Multilayer perceptron(MLP) and Decision tree(J48). It is strongly recommended for Ivor Buquetlowd that there is no necessity to collect the population data.

From the results produced by the MLP predictive model it is highly recommended that these variables Staff, Window, Car park, Location, Competition number, Competition score and Performance are sufficient to achieve Ivor Buquetlowd objectives.

Predictive modelling is more of an approach than a process. It is typically a machine learning algorithm. These algorithms perform the data mining statistical analysis, determining trends and patterns in the data. Weka software have built in algorithms that can be used to make predictive models. One of the predictive model that had been used in this report is **Multilayer Perceptron(MLP).**

As per given scenario, I strongly recommend MLP for prediction task. It fulfils the Ivor Buquetlowd's requirements such as:

1. Why some shops are doing better than others?
2. How much money that shop should make?

In the predictive model the data is classified in a simple but powerful form of multiple variable analysis called **Decision trees**. These are produced by algorithms that are identified various ways of splitting data into branch like segments. Decision trees partition data into subset based on the categories of input variables like:

1. Performance
2. Competition score
3. Window
4. Clearance space
5. Floor space
6. Car park
7. Location
8. Competition number
9. Staff

It helps Ivor Buquetlowd to make decisions for improving better profits and forecasting the revenue of each shop in the supply chain.

I conclude that predictive analytics solutions help organisations to turn their data into timely insides for better and faster decision making.