Machine Learning for Data Analysis
Assignment 1

This assignment has been completed with SAS-studio and edited with MS word for explanations and analysis. The code is added as an appendix at the end of this document.

Response Variable: Life Expectancy,

1: High – Life expectancy greater or equal to 60 years

2: Low – Life expectancy less than 60 years

Explanatory Variables:

- country
- incomeperperson
- alcconsumption
- breastcancerper100th
- femaleemployrate
- hivrate
- suicideper100th
- employrate
- urbanrate

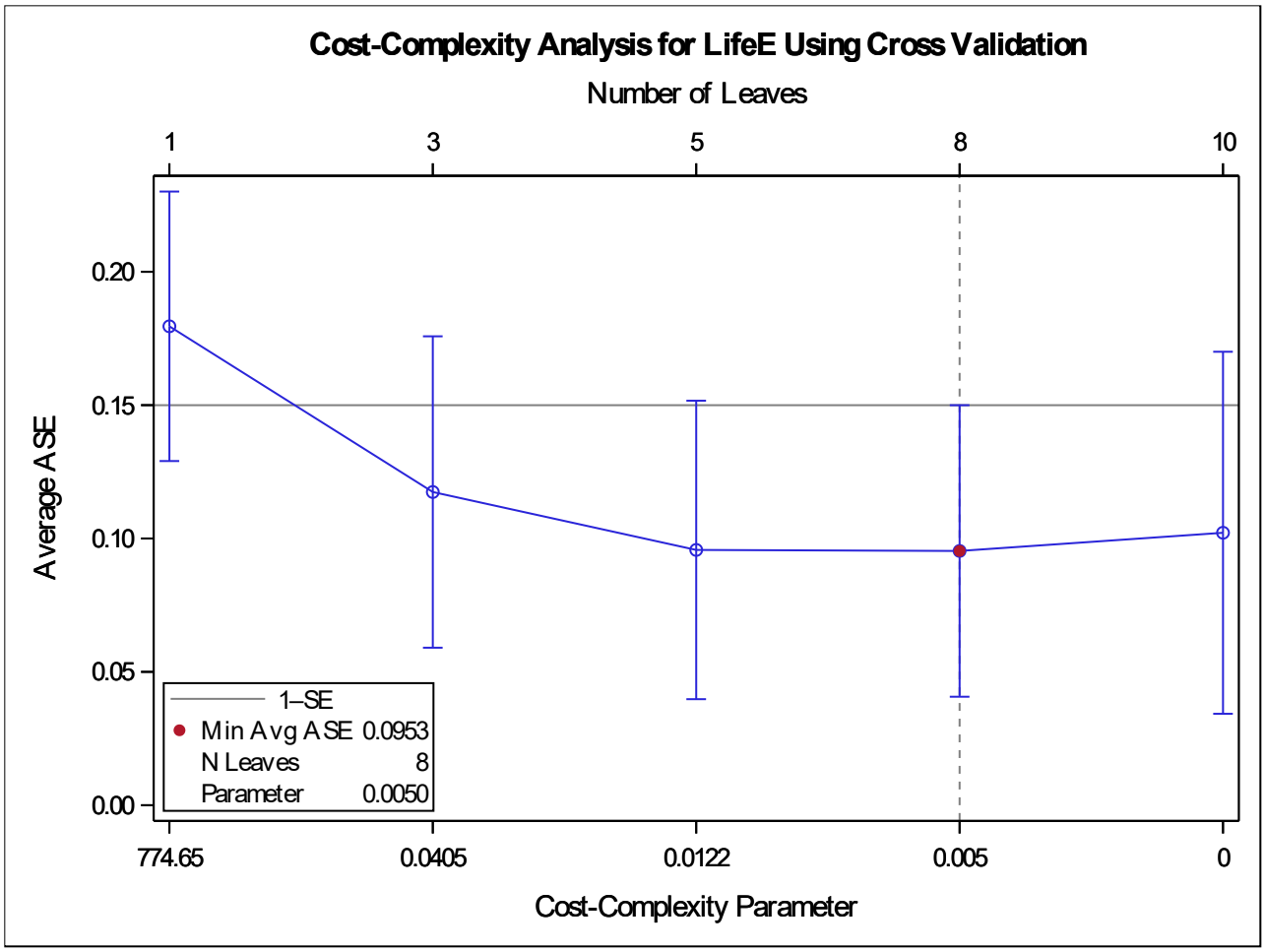
"LifeExpectancy" is converted to a categorized variable yielding values of 'high' and 'low' with a threshold at 60.
The decision tree analysis tests the non-linear relations among the specified variables.
Entropy and cost complexity criteria are included to obtain final subtree.

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

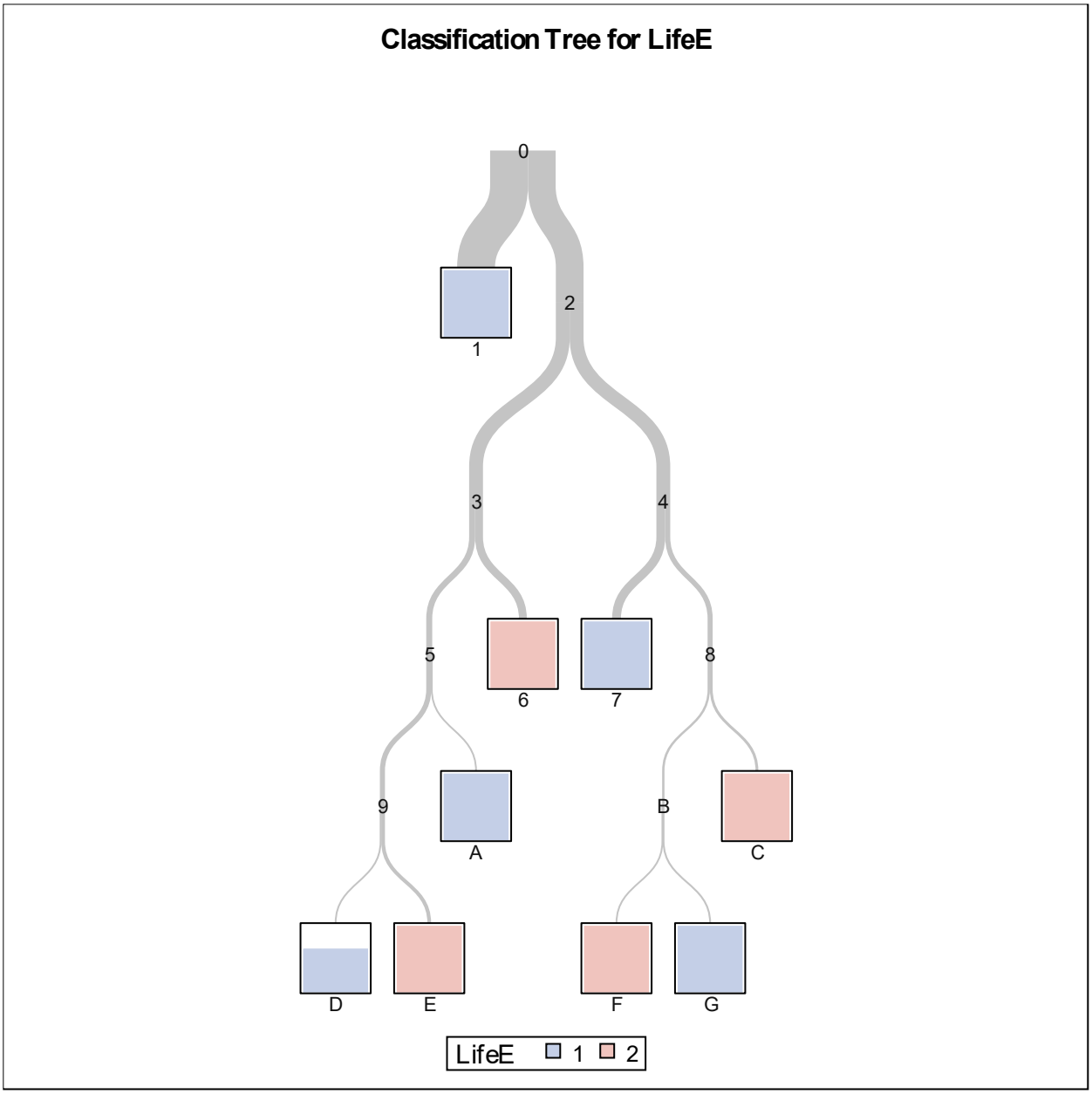
Data Access Information			
Data	**Engine**	**Role**	**Path**
WORK.NEW	V9	Input	On Client

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	7
Tree Depth	5
Number of Leaves Before Pruning	11
Number of Leaves After Pruning	9
Model Event Level	1

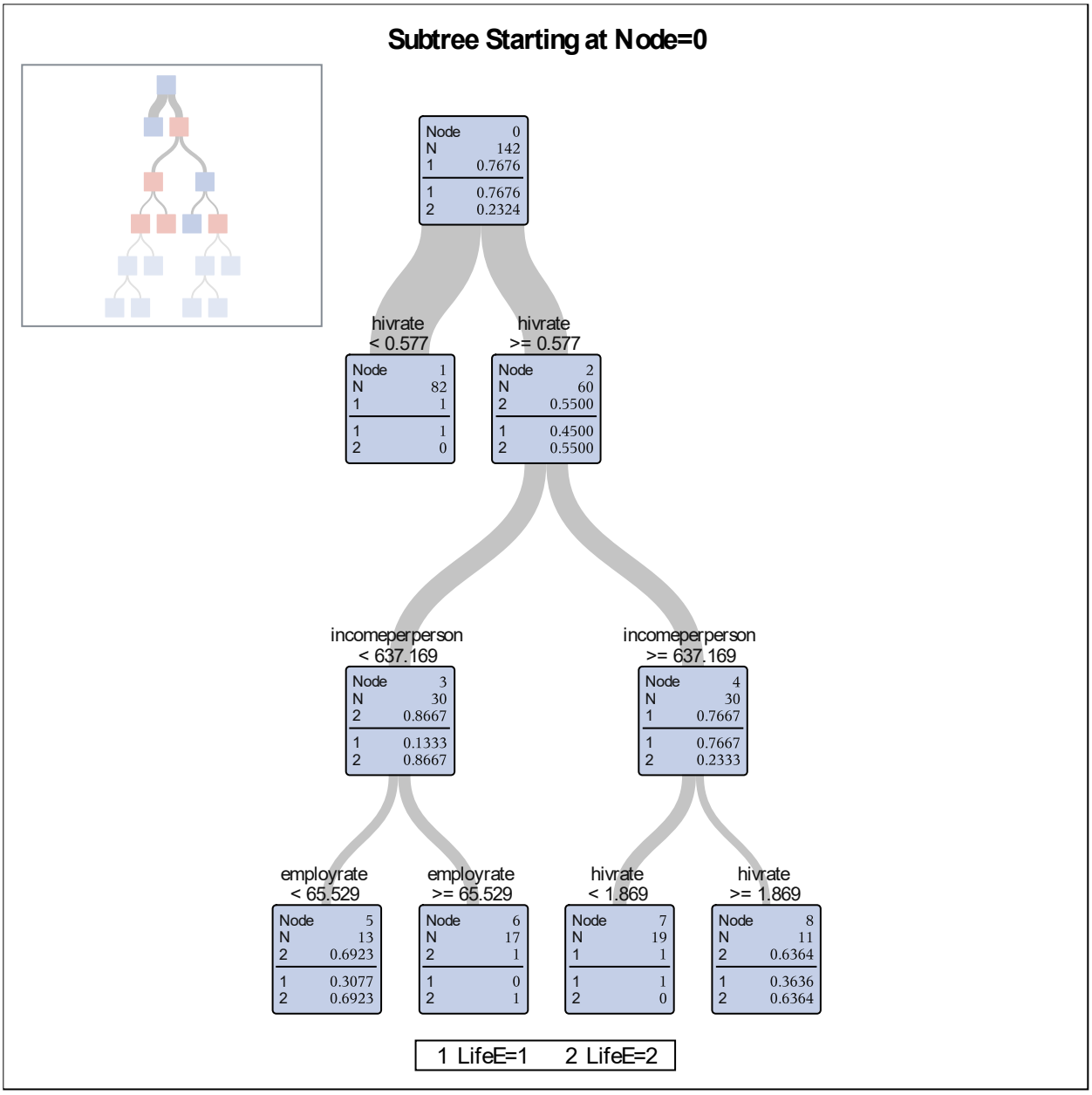
Number of Observations Read	213
Number of Observations Used	142


Cost-Complexity Analysis for LifeE Using Cross Validation
Number of Leaves
1
3
5
8
10
Average ASE
0.20
0.15
0.10
0.05
0.00
1–SE
Min Avg ASE 0.0953
N Leaves 8
Parameter 0.0050
774.65
0.0405
0.0122
0.005
0
Cost-Complexity Parameter

The HPSPLIT Procedure

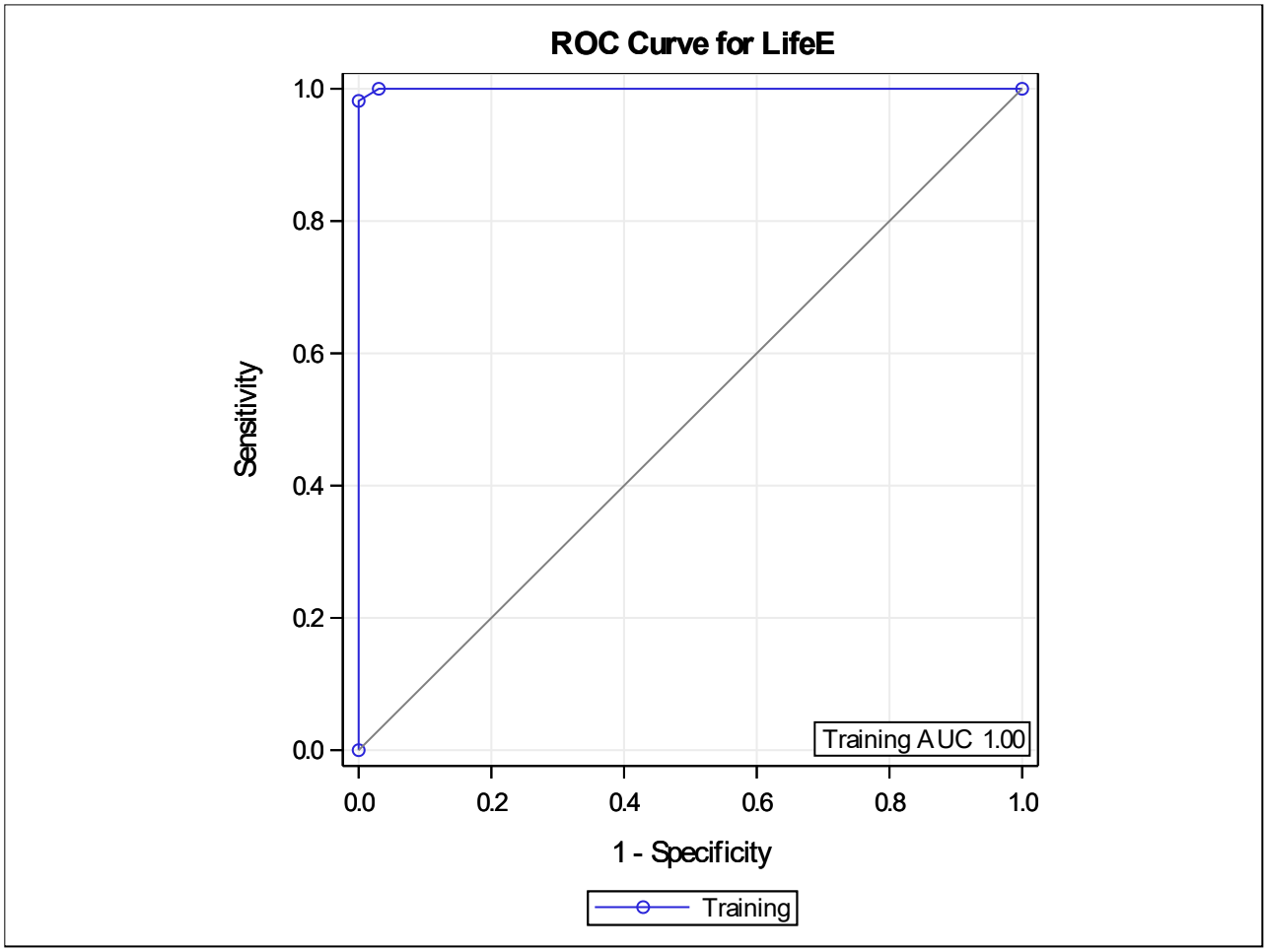


The HPSPLIT Procedure



Model-Based Confusion Matrix			
	Predicted		Error
Actual	1	2	Rate
1	109	0	0.0000
2	1	32	0.0303

Model-Based Fit Statistics for Selected Tree								
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
9	0.00469	0.0070	1.0000	0.9697	0.0194	0.00939	1.3333	0.9997



Variable Importance				
		Training		
Variable	Variable Label	Relative	Importance	Count
hivrate	HIVRATE	1.0000	5.1579	2
incomeperperson	INCOMEPERPERSON	0.7159	3.6923	2
employrate	EMPLOYRATE	0.3709	1.9133	2
suicideper100th	SUICIDEPER100TH	0.3622	1.8684	1
femaleemployrate	FEMALEEMPLOYRATE	0.2700	1.3926	1

Analysis

142 were considered for the analysis, out of 213 observations

The initial tree yields 11 nodes. When pruning is applied, subtree yields 9 leaves.

I - The first classification is obtained when the HIV rate branches into two groups:

a) 82 countries with HIV rate < 0.577 and 100% has a higher life expectancy

b) 60 countries with HIV rate (>= 0.577) where 45% have high life expectancy and 55% have low.

II - This subgroup of 60 countries is then divided by income per person with a threshold of 637.169.

III - The subgroup tree reveals that:

a) Countries with higher HIV rate, and

b) Low income per person (30 countries) and

c) Low employ rate (<65.529) (13 countries) results in:

Only 30.77% have high life expectancy and 69.23% has low life expectancy

IV - The tree also reveals that:

a) Countries with higher HIV rate, and

b) Higher income per person (30 countries), and

c) lower HIV rate (19 countries) results in:

100% higher life expectancy. This is opposed to counties with higher HIV rate where 35.36% has higher life expectancy and rest low higher life expectancy.

V - Confusion Matrix:

The total model correctly classifies those countries with high life expectancy 100%

The confusion matrix reveals the classification tree classifies It correctly classifies countries with low life expectancy 98% of the time (1 - 0.0303 = 0. 0.9697 or 96.97%).

Finally, the model variable importance table. Due to the fact that decision trees attempt to maximize correct classification with the simplest tree structure, it's possible for variables that do not necessarily represent primary splits in the model to be of notable importance in the prediction of the target variable.

Potential explanatory variables are highly correlated, or provide similar information, for example: HIVRATE, INCOMEPERPERSON. EMPLOYRATE, SUICIDEPER100TH, FEMALEEMPLOYRATE are likely to be selected for the model. The absence of the alternate variable from the model does not necessarily suggest that it's unimportant, but rather that it's masked by the others.

Appendix 1

3/25/2017 Code: W1.sas

```
1  /* Machine Learning for Data Analysis */
2  /* Running a Classification Tree*/
3  /* AB Lopez*/
4  LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
5
6  DATA new;
7       set mydata.gapminder;
8
9       /*life expectancy can be classified in Low, average and High */
10      if lifeexpectancy GE 60 THEN
11              LifeE=1;
12
13      /*above 60 high*/
14      else
15              LifeE=2;
16
17      /*low*/
18 PROC SORT ;
19      BY country;
20      ods graphics on;
21
22 proc hpsplit seed=155311;
23      class LifeE country;
24      model LifeE=country incomeperperson alcconsumption breastcancerper100th
25              femaleemployrate hivrate suicideper100th employrate urbanrate;
26      grow entropy;
27      prune costcomplexity;
28 RUN;
```