CENG 484 – DATA MINING

Assignment 2 - Report

Instructor Name

Serap Şahin

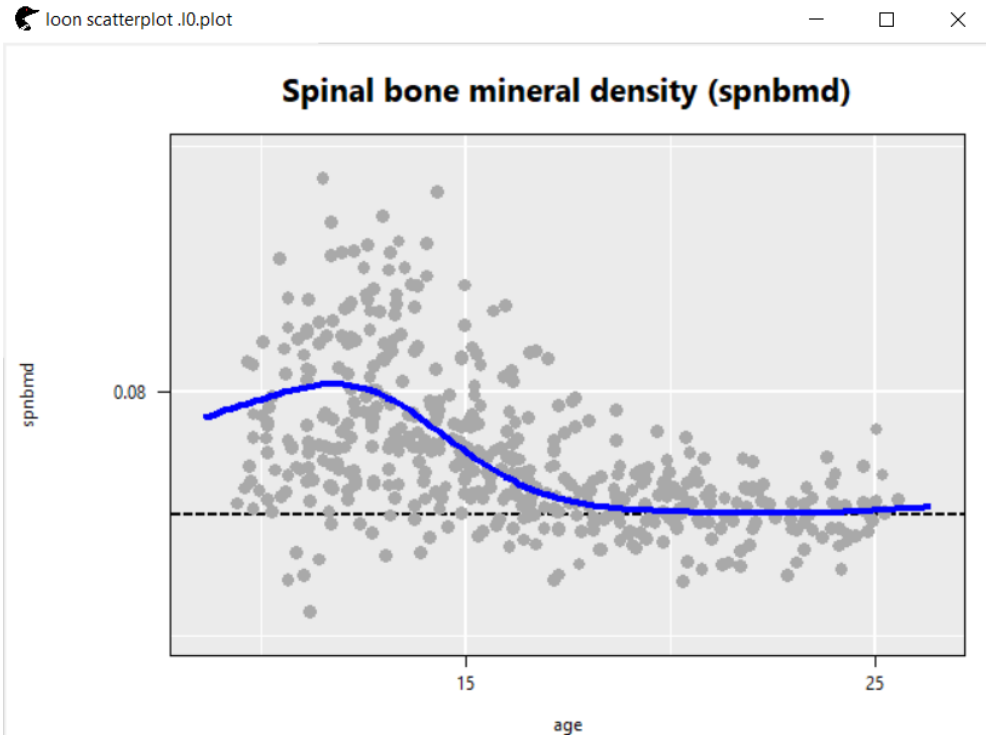
Student No – Name

220201029 – İlayda Cansın Koç

BONE.DATA FILE

After running the script ***spnbmd-age-gender.R*** you will see two graphics and one inspector screen like below.

- This scatter plot shows the relationship between features "***Spinal bone mineral density(spnbmd)***" and **"age".**



- The histogram below shows the "***Female***" and "***Male***" frequency in the given dataset bone.data.

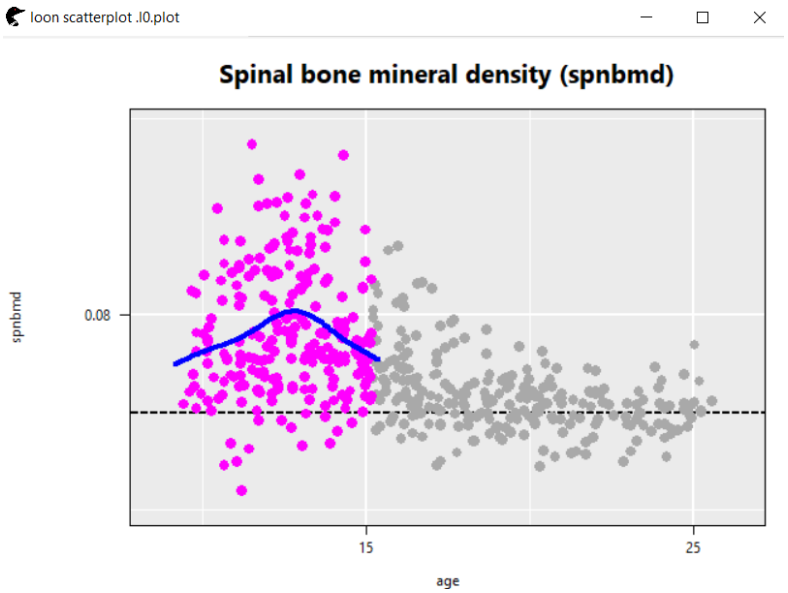
The other screen is the inspector screen in which you can modify graphs. I will not go into details about how it is used.

- For the first graph that shows the "***Spinal bone mineral density(spnbmd)"*** and **"age"** relationship, it can be seen that for the age between 11 – 12, the spinal bone mineral density(***spnbmd***) is in its highest and after the age 15 it drastically decreases and reaches negative values and then increase a little bit again.

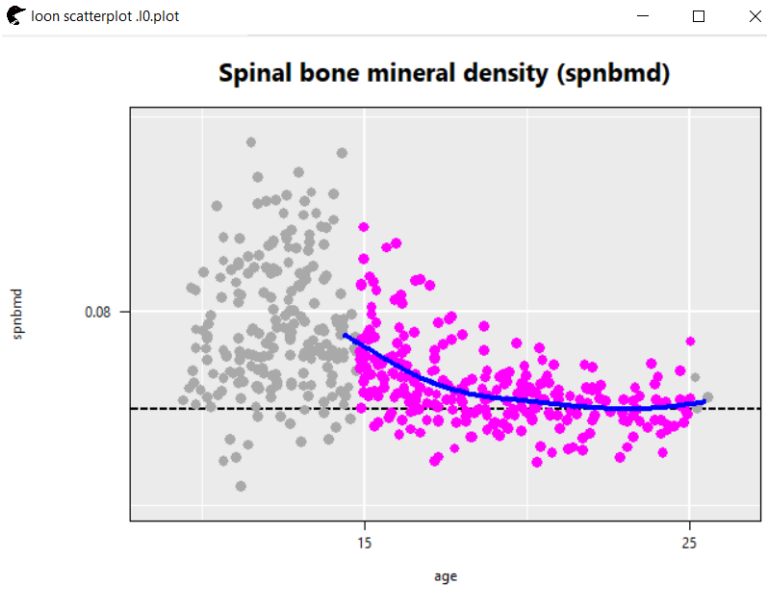
Also, the relationship of gender among this graph can be seen by applying following steps.

Select data by dragging your mouse when pressing right click of the mouse.

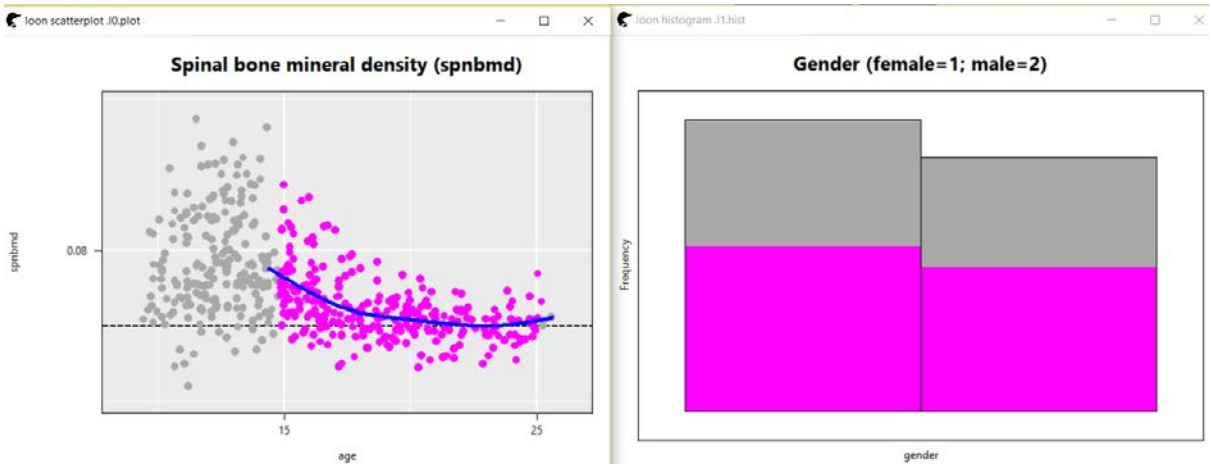
- By applying this method, for the ages between 0 – 15, graph will look like below.



- For the ages between 15 – 25, graph will look like below.

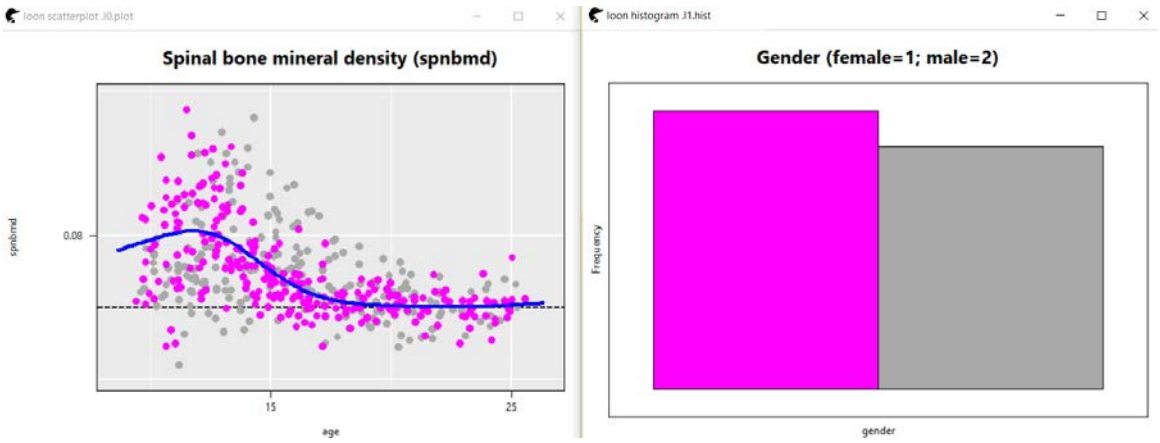


- Also, for the ages between 15 – 25, the histogram graph will look like below.

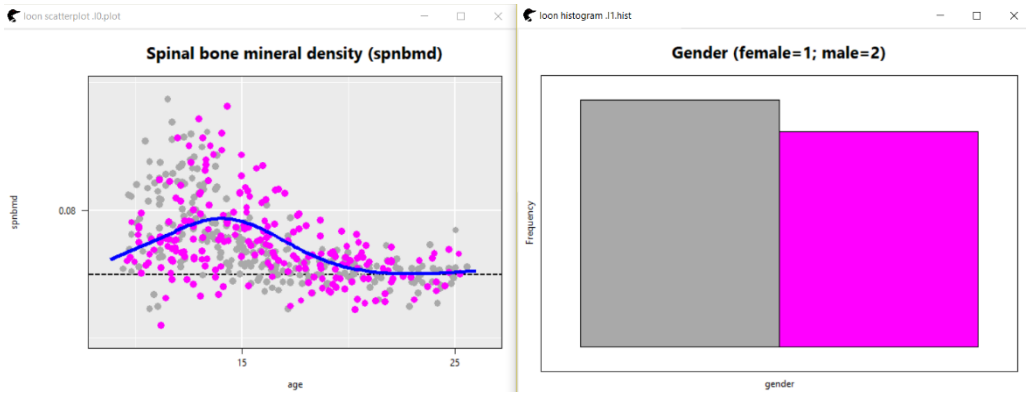


It can be resulted that for the ages between 15 – 25, the number of females is a little bit higher than the males.

- If you select gender as female, then the graph's smooth will be updated as below.



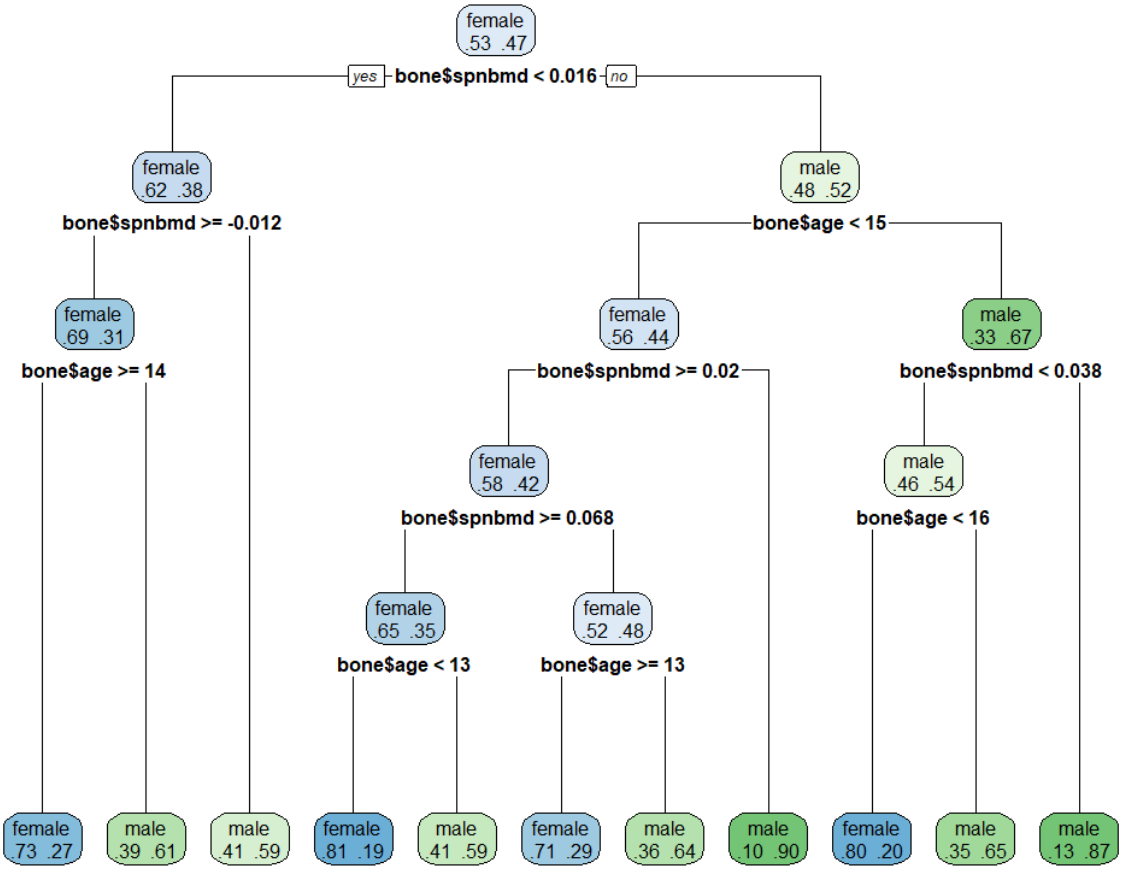
- If you select gender as male, then the graph's smooth will be updated as below.



So, it can be said that the spinal bone mineral density is at its highest between the ages 11 – 12 for females and between 13 – 14 for males.

DECISION TREE 1 – BONE.DATA FILE

- **Gender – SPNBMD & Age Relationship**

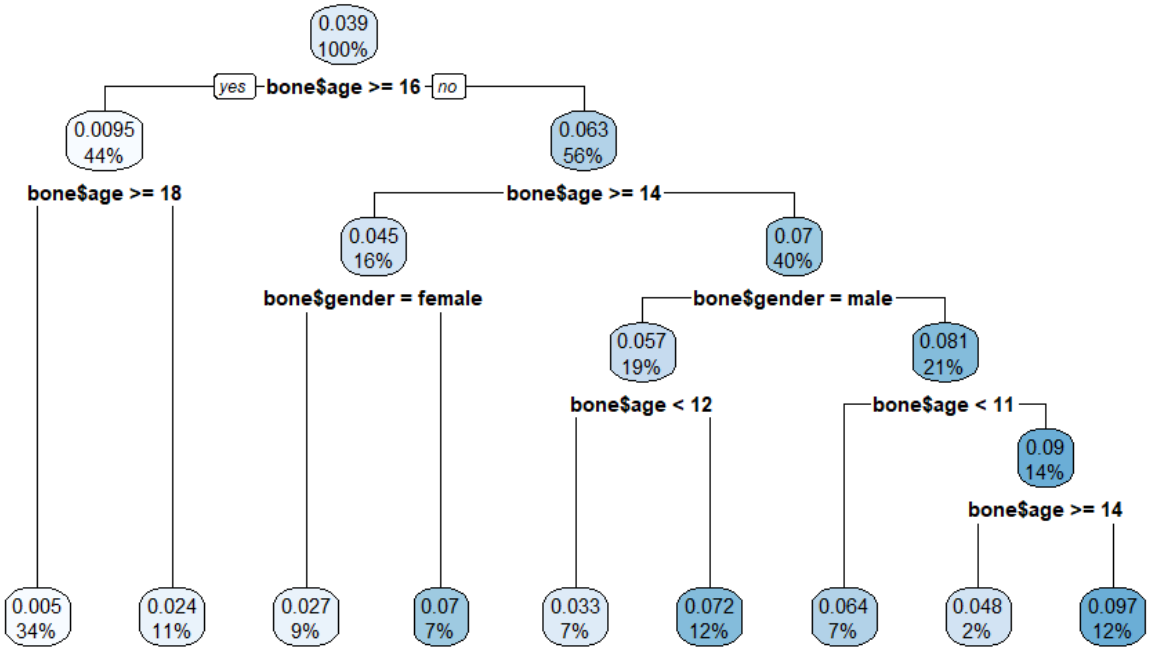


Predictions about Decision Tree

According to the decision tree, a person with age 15, spnbmd < 0.02 will be classified as ***female***. You can see the code in "***deciontree1.R"*** file and also the decision tree in ***DecisionTree1.pdf*** file.

DECISION TREE 1.1 – BONE.DATA FILE

- **SPNBMD – Gender & Age Relationship**

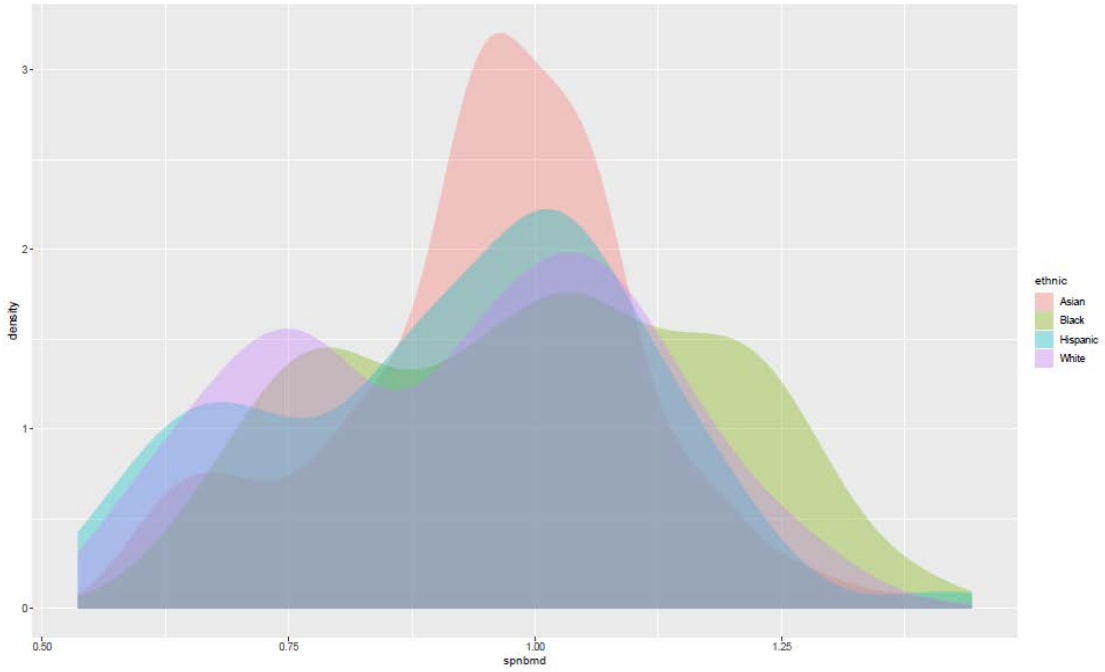


Predictions about Decision Tree

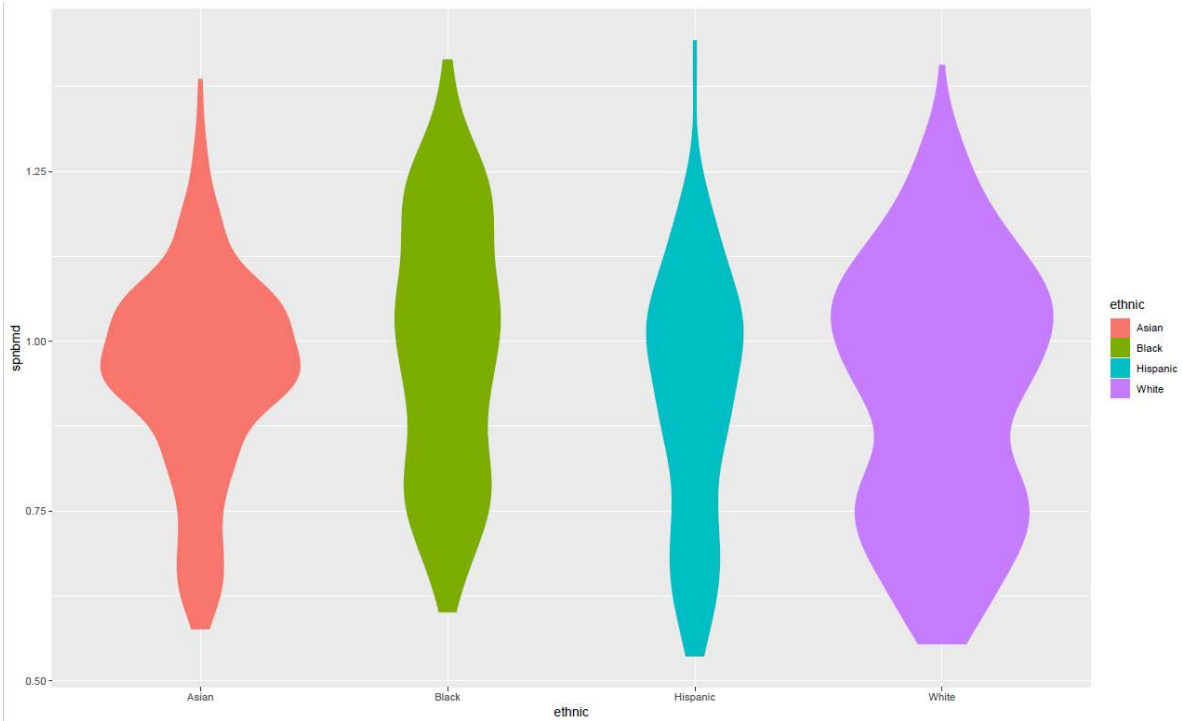
According to the decision tree, a person's spnbmd value, with age 11, gender male, will be predicted as 0.033 with the probability 7%. Also, it can be seen that females with age between 11 and 12 are likely to have spnbmd value 0.097 which confirms the scatter plot that is draw in first part of the report. You can see the code in "***deciontree1_1. R"*** file and also the decision tree in ***DecisionTree1_1.pdf*** file.

SPNBMD.CSV FILE

Instead of drawing l_plot (), I decided plotting density (*see density2.pdf*) and violin (see *violinPlot2.pdf*) graphs to show relationship between ethnic of the people and their spinal bone mineral density measurements. Furthermore, I created two decision trees as well.



As you can see, in the density graph above, Asian people have the most density and it is around the ***spnbmd*** value of 1.00.

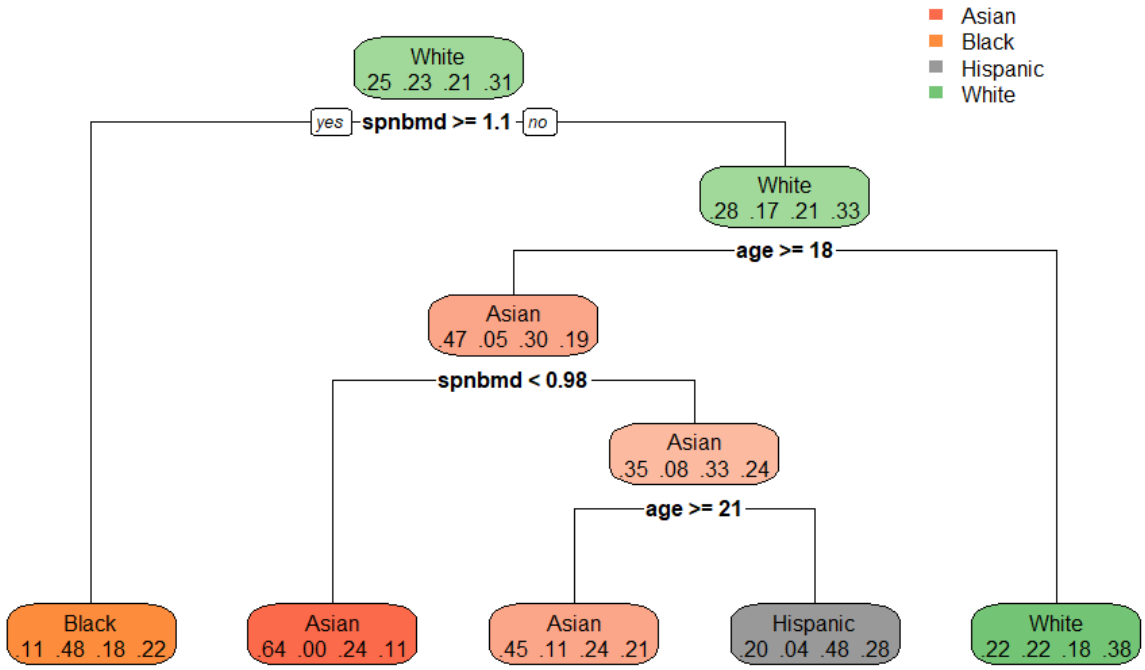


In the violin graph above, it can be confirmed that Asians accumulated around the spnbmd value of 1.00. Also, for people with White ethnicity there are two accumulations around 1.15 and 0.75. All ethnicities have their least number of people above the spnbmd value 1.25.

DECISION TREE 2 – SPNBMD.CSV FILE

- **Ethnic – SPNBMD & Age & Sex Relationship**

The decision tree below takes "***age***", "***sex***" and "***spnbmd***" values and then tries to classify which ethnicity the person belongs to. (*see DecisionTree2.pdf and decisionTree2.R*)



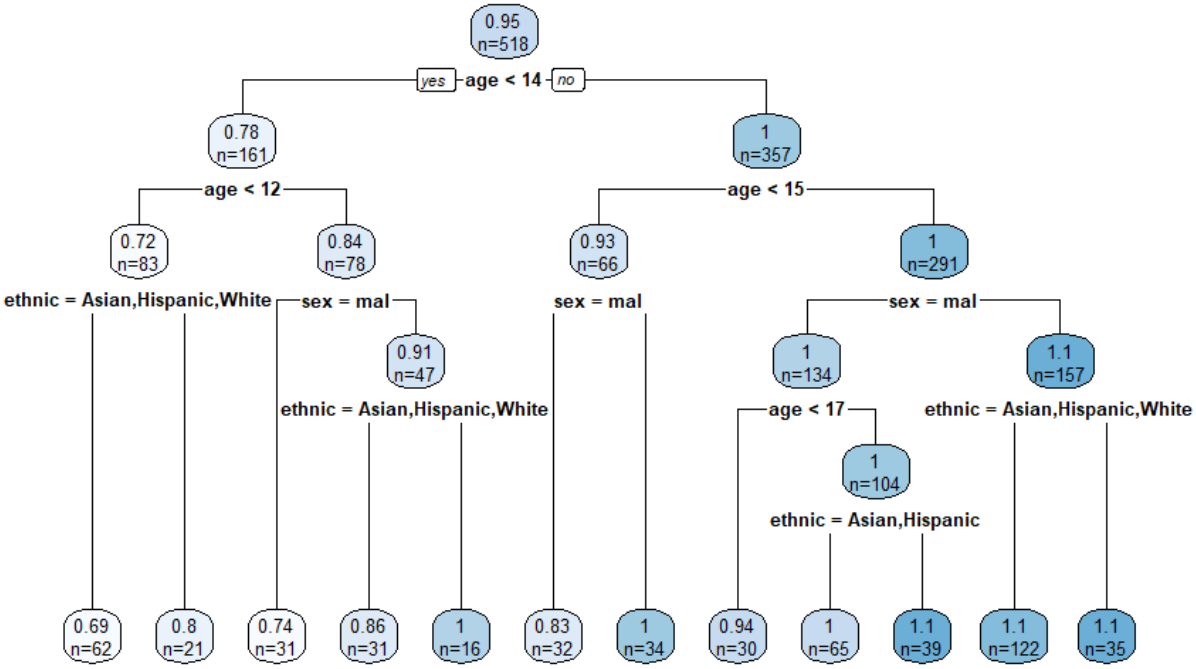
Predictions about Decision Tree

For person with age 22 and spnbmd > 0.02, according to the tree h/she will belong to Asian ethnicity.

DECISION TREE 3 – SPNBMD.CSV FILE

- **SPNBMD – Ethnic & Age & Sex Relationship**

The decision tree below takes "***age***", "***sex***" and "***ethnic***" values and then tries to classify which ethnicity the person belongs to. (*see DecisionTree3.pdf and decisionTree3.R*)

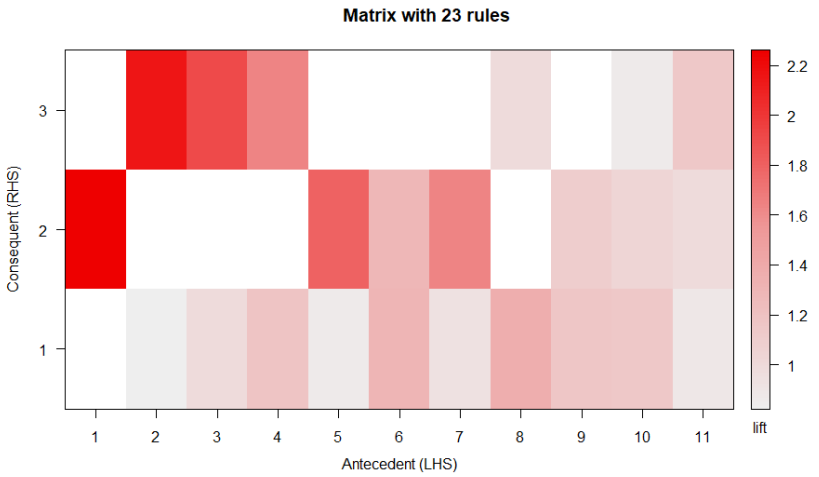


Predictions about Decision Tree

For person with age 22, ethnic Asian and sex Female is likely to have spnbmd value 0.86.

Apriori Algorithm

I also applied Apriori algorithm to this data set to see rule set which you can see the implementation in the script ***apriori.R*** **.** I visualised them using Matrix based visualisation as you can see below.



And rules generated as follows:

	lhs		rhs	support	confidence	lift	count
[1]	{age=[13.6,17.7)}	=>	{spnbmd=[0.0119,0.0475)}	0.12783505	0.3850932	1.1600633	62
[2]	{age=[9.4,13.6)}	=>	{spnbmd=[0.0119,0.0475)}	0.09484536	0.2839506	0.8553792	46
[3]	{age=[17.7,25.6]}	=>	{spnbmd=[0.0119,0.0475)}	0.10927835	0.3271605	0.9855456	53
[4]	{gender=male}	=>	{spnbmd=[0.0119,0.0475)}	0.17525773	0.3761062	1.1329907	85
[5]	{gender=female}	=>	{spnbmd=[0.0119,0.0475)}	0.15670103	0.2934363	0.8839541	76
[6]	{age=[13.6,17.7)}	=>	{spnbmd=[0.0475,0.22]}	0.12164948	0.3664596	1.0971168	59
[7]	{age=[9.4,13.6)}	=>	{spnbmd=[0.0475,0.22]}	0.20000000	0.5987654	1.7926002	97
[8]	{gender=male}	=>	{spnbmd=[0.0475,0.22]}	0.15876289	0.3407080	1.0200208	77
[9]	{gender=female}	=>	{spnbmd=[0.0475,0.22]}	0.17525773	0.3281853	0.9825301	85
[10]	{age=[17.7,25.6]}	=>	{spnbmd=[-0.0641,0.0119)}	0.21237113	0.6358025	1.9034827	103
[11]	{gender=male}	=>	{spnbmd=[-0.0641,0.0119)}	0.13195876	0.2831858	0.8478095	64
[12]	{gender=female}	=>	{spnbmd=[-0.0641,0.0119)}	0.20206186	0.3783784	1.1327995	98
[13]	{age=[13.6,17.7),gender=male}	=>	{spnbmd=[0.0119,0.0475)}	0.04536082	0.3055556	0.9204624	22
[14]	{age=[13.6,17.7),gender=female}	=>	{spnbmd=[0.0119,0.0475)}	0.08247423	0.4494382	1.3538977	40
[15]	{age=[9.4,13.6),gender=male}	=>	{spnbmd=[0.0119,0.0475)}	0.06804124	0.4285714	1.2910382	33
[16]	{age=[17.7,25.6],gender=male}	=>	{spnbmd=[0.0119,0.0475)}	0.06185567	0.3896104	1.1736710	30
[17]	{age=[17.7,25.6],gender=female}	=>	{spnbmd=[0.0119,0.0475)}	0.04742268	0.2705882	0.8151261	23
[18]	{age=[13.6,17.7),gender=male}	=>	{spnbmd=[0.0475,0.22]}	0.08041237	0.5416667	1.6216564	39
[19]	{age=[13.6,17.7),gender=female}	=>	{spnbmd=[-0.0641,0.0119)}	0.05979381	0.3258427	0.9755167	29
[20]	{age=[9.4,13.6),gender=male}	=>	{spnbmd=[0.0475,0.22]}	0.06804124	0.4285714	1.2830688	33
[21]	{age=[9.4,13.6),gender=female}	=>	{spnbmd=[0.0475,0.22]}	0.13195876	0.7529412	2.2541757	64
[22]	{age=[17.7,25.6],gender=male}	=>	{spnbmd=[-0.0641,0.0119)}	0.08659794	0.5454545	1.6329966	42
[23]	{age=[17.7,25.6],gender=female}	=>	{spnbmd=[-0.0641,0.0119)}	0.12577320	0.7176471	2.1485113	61

- **Why Apriori Algorithm?**

I used apriori algorithm in order to mine frequent item sets and relevant association rules to the data set given.

Why I Used Decision Tree in My Analysis?

I choose Decision Trees because of its key advantages:

- Decision Trees implicitly perform feature selection. So, when we fit a decision tree to a training dataset, it is easy to select which features that you want to work on and analyse.
- Decision Trees saves data preparation time.
- Decision Trees are so intuitive and also easy to explain.

Using decision trees in order to find a relationship among features of dataset was simple and easy to understand and implement. Also, I applied decision tree methodology because it is a model that is used frequently for establishing classification systems for developing prediction algorithms for a target variable. The algorithm's being non-parametric and since it can efficiently deal with large dataset were also the reason of choosing this method. By applying the decision tree methodology, it will be easier to predict and understand where the new entry will belong to.