Name Tanadol Chuntarasupt ID 6481259

Exercise 1 (22 points) – individual work

- The answers can be typed or handwritten (handwriting must be clear and readable), in this exercise sheet or your own sheet (put your name & ID at the top of the sheet). All answers must be saved to only 1 PDF file.
- Some questions also require the submission of processes/workflows (file.rmp or file.ipynb).
- In case of re-submission (after first grading) or submission after solution is given, your points will be weighted by 0.5.

----

1. (Total 15 points) Retrieve **toydata**. Perform the following tasks to handle missing values.

1.1 (3 points) Consider attribute age

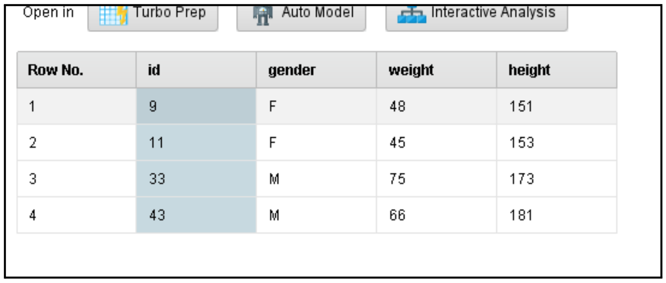
Instructions	Questions
Step 1. Find mean & SD of age, calculated from the whole dataset.	Answer in 2 decimal places Mean of age = 48.887 SD of age = 16.877
Step 2. Identify records with missing age. Impute missing age in these records by central tendency of the whole dataset.	Imputed value = 49
Step 3. Find mean & SD of age, calculated from the whole dataset, after imputation.	Answer in 2 decimal places Mean of age = 48.888 SD of age = 16.765

1.2 (5 points) Consider attribute gender

Instructions	Questions
Step 1. Find mean & SD of weight and height, calculated from only male records.	Answer in 2 decimal places Mean of weight = 70.083 SD of weight = 10.625 Mean of height = 171.774 SD of height = 7.543
Step 2. Identify records with missing gender. Impute missing gender in these records by central tendency of the whole dataset.	Imputed value = M
Step 3. Find mean & SD of weight and height, calculated from only male records, after imputation.	Answer in 2 decimal places Mean of weight = 69.557 SD of weight = 10.988 Mean of height = 171.443 SD of height = 8.011

1.3 (4 points) Select only attributes: gender, weight, height. Apply KNN imputation with K=5 (other parameters = default) instead of the imputation in step 2 of 1.2.

<table>
<tr><th>Instructions</th><th>Questions</th></tr>
<tr><td>Step 2. Identify records with <u>missing gender</u>. Impute missing gender in these records by KNN. This time these records may get different imputed values.</td><td>Show record ID, imputed gender, weight, height of these records (there are 4 records).

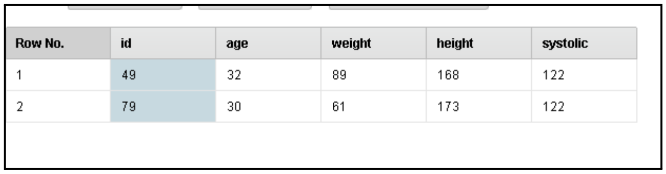


<table>
<tr><th>Row No.</th><th>id</th><th>gender</th><th>weight</th><th>height</th></tr>
<tr><td>1</td><td>9</td><td>F</td><td>48</td><td>151</td></tr>
<tr><td>2</td><td>11</td><td>F</td><td>45</td><td>153</td></tr>
<tr><td>3</td><td>33</td><td>M</td><td>75</td><td>173</td></tr>
<tr><td>4</td><td>43</td><td>M</td><td>66</td><td>181</td></tr>
</table>
</td></tr>
<tr><td>Step 3. Find mean & SD of weight and height, calculated from <u>only male records</u>, after imputation.</td><td>Answer in 2 decimal places
Mean of weight = 70.093
SD of weight = 10.522

Mean of height = 171.895
SD of height = 7.521</td></tr>
</table>

1.4 (3 points)

<table>
<tr><th>Instructions</th><th>Questions</th></tr>
<tr><td>Step 1. Perform age imputation as in 1.1, then select <u>only male records</u></td><td></td></tr>
<tr><td>Step 2. Identify male records with <u>missing systolic</u>. Apply linear regression imputation using only values from male records (parameters can be default).</td><td>Show record ID, age, weight, height, imputed systolic of these records (there are 2 records).

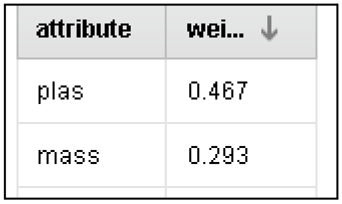
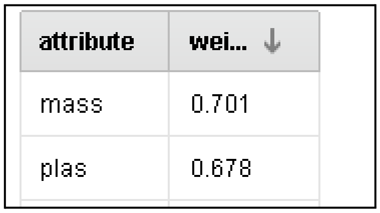
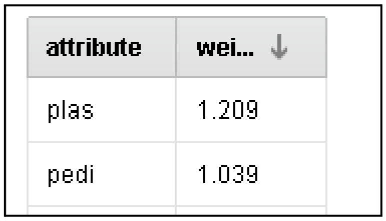


<table>
<tr><th>Row No.</th><th>id</th><th>age</th><th>weight</th><th>height</th><th>systolic</th></tr>
<tr><td>1</td><td>49</td><td>32</td><td>89</td><td>168</td><td>122</td></tr>
<tr><td>2</td><td>79</td><td>30</td><td>61</td><td>173</td><td>122</td></tr>
</table>
</td></tr>
<tr><td>Also submit your workflow that performs both steps. Name the workflow <b>question1_4.rmp</b></td><td></td></tr>
</table>

2. (Total 7 points) Retrieve **diabetes**. Perform the following tasks for attribute selection.

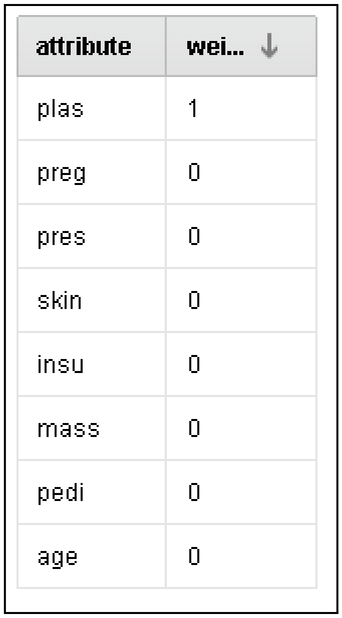
Attribute	Short description
preg	Number of times pregnant
plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
pres	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
insu	2-Hour serum insulin (mu U/ml)
mass	Body mass index (weight in kg/(height in m)^2)
pedi	Diabetes pedigree function, i.e. likelihood based on family history
age	Age (years)
Class	tested_positive, tested_negative

2.1 (3 points) Apply 3 attribute ranking methods. List 2 most important attributes given by each method in the table below. Note that importance is determined from the magnitude of weight, not just weight (e.g. attribute with weight -0.9 is more important than attribute with weight 0.1)

Method	2 most important attributes
1. Correlation	 attribute · wei... ↓ plas · 0.467 mass · 0.293
2. Rule	 attribute · wei... ↓ mass · 0.701 plas · 0.678
3. Random Forest	 attribute · wei... ↓ plas · 1.209 pedi · 1.039

2.2 (2 points) Use Optimize Selection method to obtain an optimal subset of attributes. You can set the workflow in the same way as Forward Selection in the chapter example.

Instructions	Questions
Step 1. Run Optimize Selection	List all selected attributes

<table>
<tr><td></td><td>
<table>
<tr><th>attribute</th><th>wei... ↓</th></tr>
<tr><td>plas</td><td>1</td></tr>
<tr><td>preg</td><td>0</td></tr>
<tr><td>pres</td><td>0</td></tr>
<tr><td>skin</td><td>0</td></tr>
<tr><td>insu</td><td>0</td></tr>
<tr><td>mass</td><td>0</td></tr>
<tr><td>pedi</td><td>0</td></tr>
<tr><td>age</td><td>0</td></tr>
</table></td></tr>
<tr><td>Step 2. Compare result from step 1 with results from your attribute ranking methods in 2.1.</td><td>The most important attribute for class prediction = plas

Reason: since three out of four rankings that I did gave this answer (correlation, rule, rf, optimize selection), then we can conclude that this attribute is important</td></tr>
</table>

2.3 (2 points) Also submit your workflow that performs 2.1 and 2.2 (they can be put in separate subprocesses). Name the workflow **question2.rmp**.