PREDICTION MODEL CREATION AND EVALUATION

COS10022- Introduction to Data Science

M ANZOR YOUSUF
STUDENT ID: 102849043

Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Introduction to Data Science			Class day/time	Wednesday
Unit code	COS10022	Assignment no.	2	Due date	14th May 2021
Name of lecturer/teacher	Pei-Wei Tsai				
Tutor/marker's name	Pei-Wei Tsai				

Office use only

Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Yousuf	M Anzor	102849043
(2)			
(3)			
(4)			
(5)			
(6)			

DECLARATION AND STATEMENT OF AUTHORSHIP

1. I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment.
2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made.
3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned.
4. I/we have not previously submitted this work for this or any other course/unit.
5. I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.

I/we understand that:

6. Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.

Student signature/s

I/we declare that I/we have read and understood the declaration and statement of authorship.

(1)	M Anzor Yousuf	(4)	
(2)		(5)	
(3)		(6)	

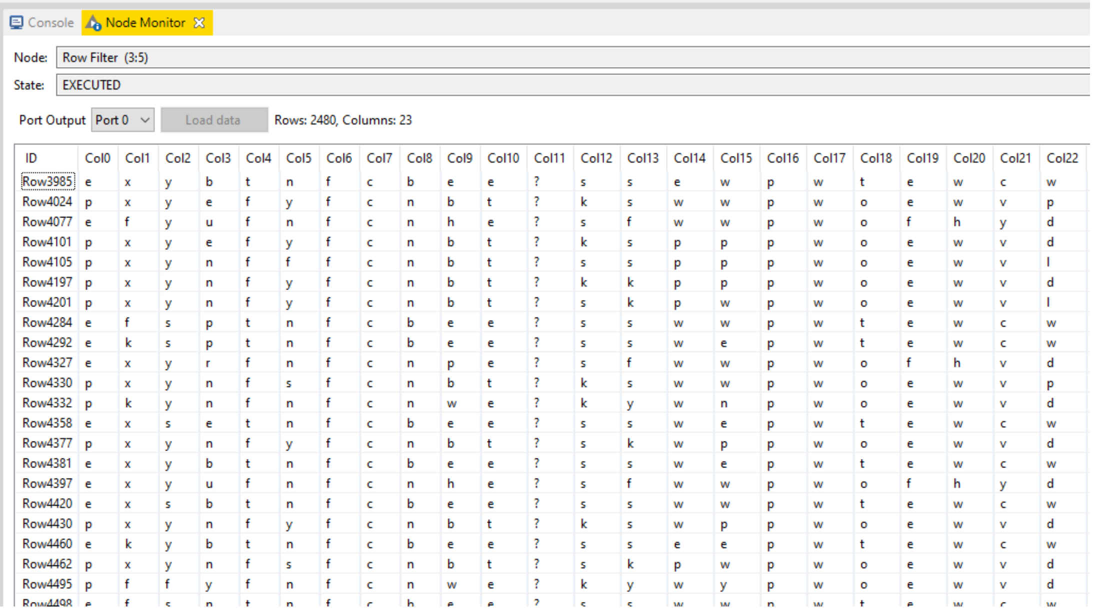
Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at **www.swin.edu.au/student/**

Copies of this form can be downloaded from the Student Forms web page at **www.swinburne.edu.au/studentforms/**

Task 1:

My strategy was to analyze the data before using tools to deep analyze it. By analyzing the data, I found that, some values of **stalk-root** are missing.

By using the Row Filter node in Knime, I found the number of missing values in the csv file. The number of missing values is 2480, which is also shown in the diagram below.



Node: Row Filter (3:5)

State: EXECUTED

Port Output Port 0 Rows: 2480, Columns: 23

ID	Col0	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Col13	Col14	Col15	Col16	Col17	Col18	Col19	Col20	Col21	Col22
Row3985	e	x	y	b	t	n	f	c	b	e	e	?	s	s	e	w	p	w	t	e	w	c	w
Row4024	p	x	y	e	f	y	f	c	n	b	t	?	k	s	w	w	p	w	o	e	w	v	p
Row4077	e	f	y	u	f	n	f	c	n	h	e	?	s	f	w	w	p	w	o	f	h	y	d
Row4101	p	x	y	e	f	y	f	c	n	b	t	?	k	s	p	p	p	w	o	e	w	v	d
Row4105	p	x	y	n	f	f	f	c	n	b	t	?	s	s	p	p	p	w	o	e	w	v	l
Row4197	p	x	y	n	f	y	f	c	n	b	t	?	k	k	p	p	p	w	o	e	w	v	d
Row4201	p	x	y	n	f	y	f	c	n	b	t	?	s	k	p	w	p	w	o	e	w	v	l
Row4284	e	f	s	p	t	n	f	c	b	e	e	?	s	s	w	w	p	w	t	e	w	c	w
Row4292	e	k	s	p	t	n	f	c	b	e	e	?	s	s	w	e	p	w	t	e	w	c	w
Row4327	e	x	y	r	f	n	f	c	n	p	e	?	s	f	w	w	p	w	o	f	h	v	d
Row4330	p	x	y	n	f	s	f	c	n	b	t	?	k	s	w	w	p	w	o	e	w	v	p
Row4332	p	k	y	n	f	n	f	c	n	w	e	?	k	y	w	n	p	w	o	e	w	v	d
Row4358	e	x	s	e	t	n	f	c	b	e	e	?	s	s	w	e	p	w	t	e	w	c	w
Row4377	p	x	y	n	f	y	f	c	n	b	t	?	s	k	w	p	p	w	o	e	w	v	d
Row4381	e	x	y	b	t	n	f	c	b	e	e	?	s	s	w	e	p	w	t	e	w	c	w
Row4397	e	x	y	u	f	n	f	c	n	h	e	?	s	f	w	w	p	w	o	f	h	y	d
Row4420	e	x	s	b	t	n	f	c	b	e	e	?	s	s	w	w	p	w	t	e	w	c	w
Row4430	p	x	y	n	f	y	f	c	n	b	t	?	k	s	w	p	p	w	o	e	w	v	d
Row4460	e	k	y	b	t	n	f	c	b	e	e	?	s	s	e	e	p	w	t	e	w	c	w
Row4462	p	x	y	n	f	s	f	c	n	b	t	?	s	k	p	w	p	w	o	e	w	v	d
Row4495	p	f	f	y	f	n	f	c	n	w	e	?	k	y	w	y	p	w	o	e	w	v	d
Row4498	e	f	s	p	t	n	f	c	b	e	e	?	s	s	w	w	p	w	t	e	w	c	w

Figure 1: The number of missing values

To fix the missing values in the dataset, I used **'N/A'** which is using a global constant to fill out the missing values. As, it is a very large dataset, it will be very hard and time consuming to fill out the whole dataset using the side information and also lack of numerical values makes it impossible to use the mean and median technique.

The attribute type from the given dataset is **string attribute** as there are no numerical value in the whole dataset. I have chosen to do the **regression model**; thus, I do not have to convert any data types.

The attributes I'm using are **bruises, odor, gill-size, gill-color, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, ring-type** and **spore-print-color.** I'm using these attributes because I think I can find a good result based on these attributes. The following data visualizations also support my idea.

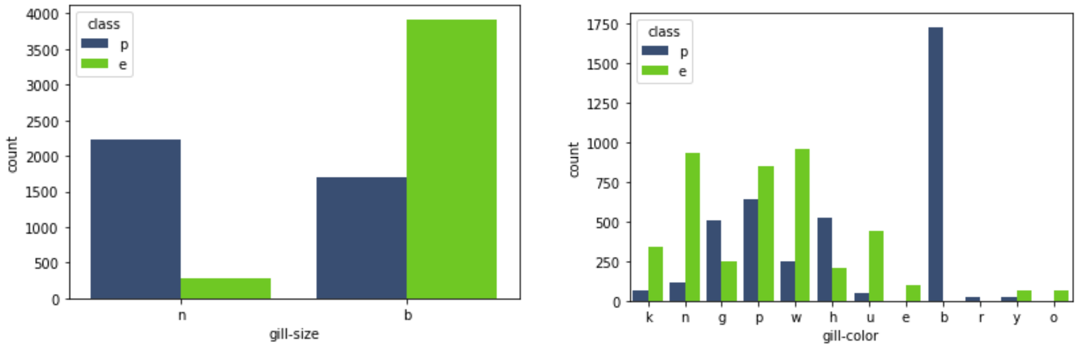


Figure 2: Data Visualization of gill-size and gill-color attributes

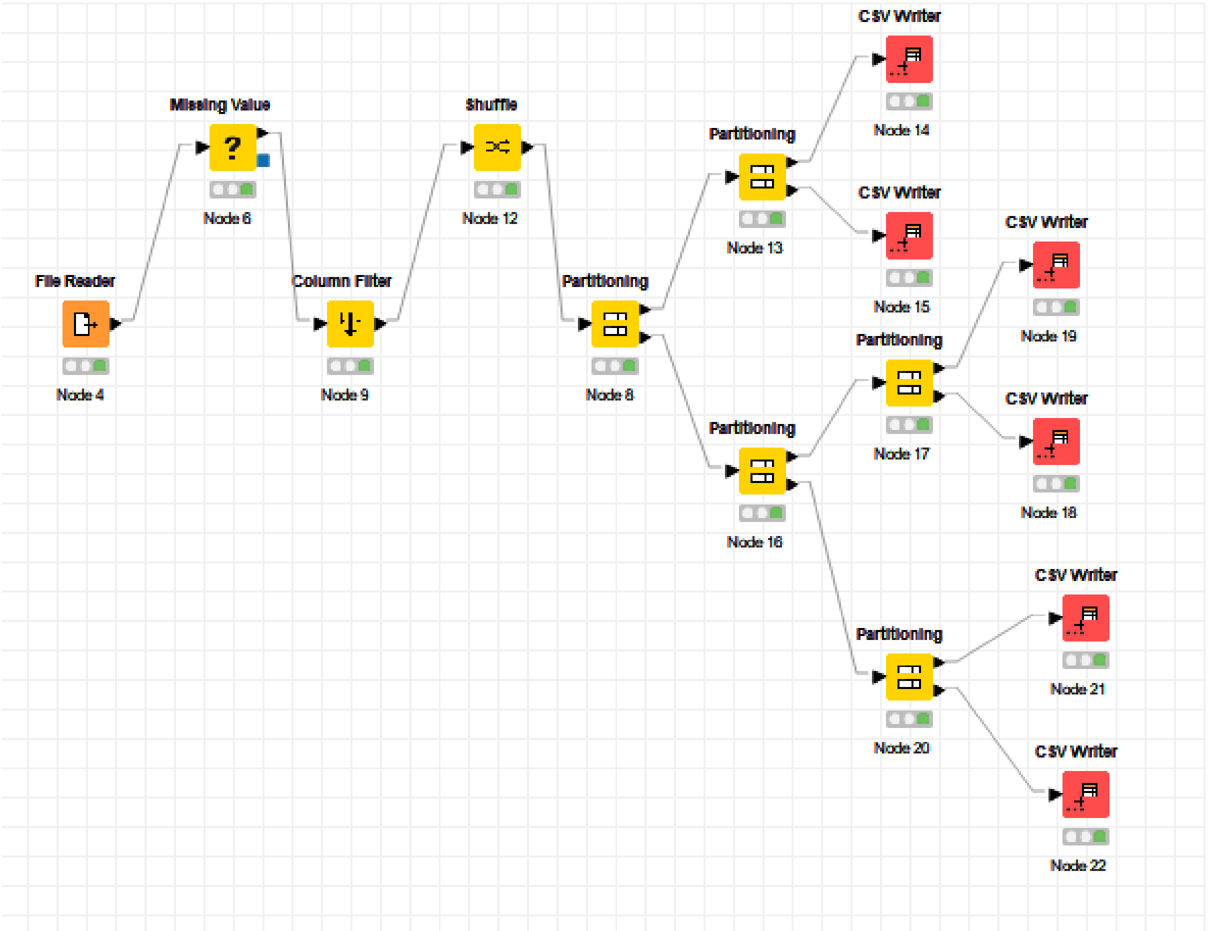


Figure 3: Shuffling and Partitioning the dataset

Using the node 6, I replaced the missing values with N/A so that knime does not detect the missing values. Then I used the column filter to select the attributes I wanted which are mentioned above. Then, I used the shuffle node to shuffle dataset according to the question. Later on, using node 8 partition, I split the output into two parts with the first one node 13 containing 2708 tuples and the rest (5416 tuples) going to the node 16 partition. From node 13, I separated by a relative value of 90% to node 14, the first csv writer output which I named “Training 1” and the rest of the data going to node 15 “Test1”. Similarly, I used partitioning node 17 and node 20 to create the rest of the files.

Task2:

For this task, I used the regression model as the prediction model. I used the Logistic Regression Lerner and the Logistic Regression Predictor to create the regression model. The regression model and the confusion matrix is attached below as screenshots.

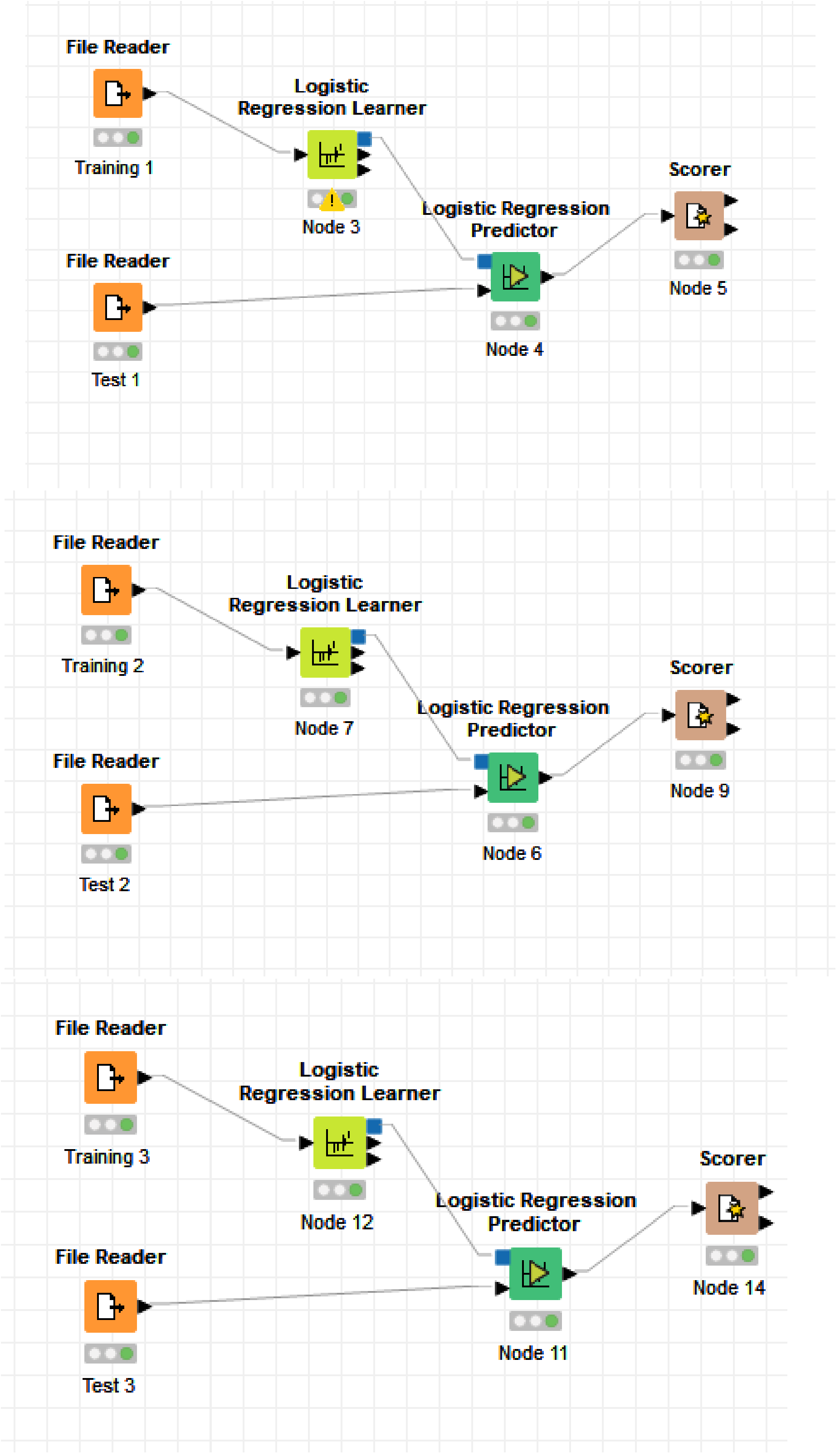
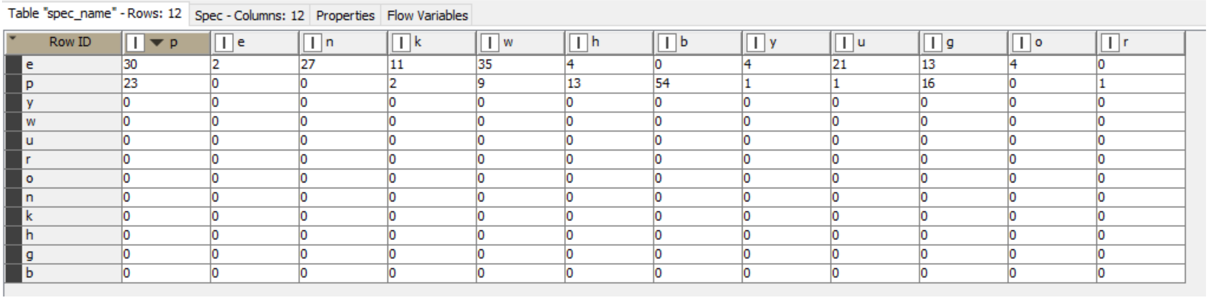


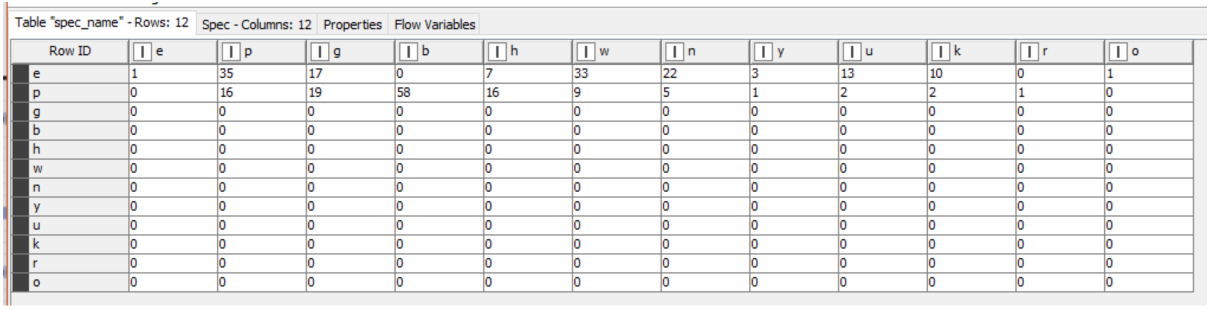
Figure 4: Regression Model for Mushroom

Confusion Matrix:



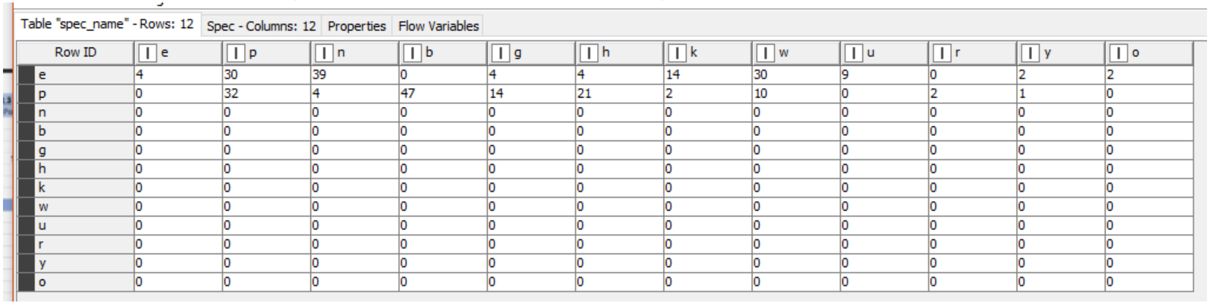
Row ID	p	e	n	k	w	h	b	y	u	g	o	r
e	30	2	27	11	35	4	0	4	21	13	4	0
p	23	0	0	2	9	13	54	1	1	16	0	1
y	0	0	0	0	0	0	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5: Confusion Matrix-Regression Model Test 1



Row ID	e	p	g	b	h	w	n	y	u	k	r	o
e	1	35	17	0	7	33	22	3	13	10	0	1
p	0	16	19	58	16	9	5	1	2	2	1	0
g	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0

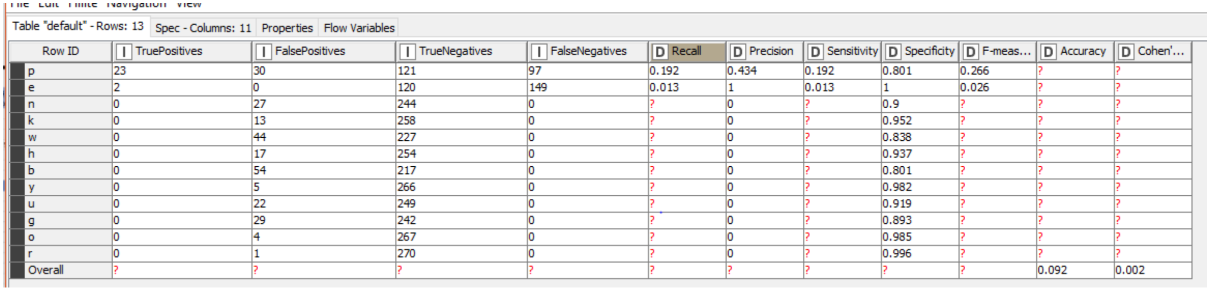
Figure 6: Confusion Matrix-Regression Model Test 2



Row ID	e	p	n	b	g	h	k	w	u	r	y	o
e	4	30	39	0	4	4	14	30	9	0	2	2
p	0	32	4	47	14	21	2	10	0	2	1	0
n	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0

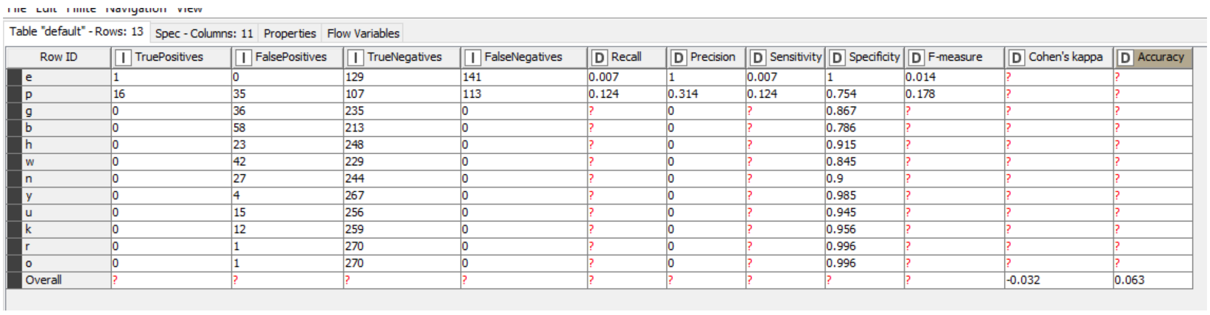
Figure 7: Confusion Matrix-Regression Model Test 3

Accuracy Statistics:



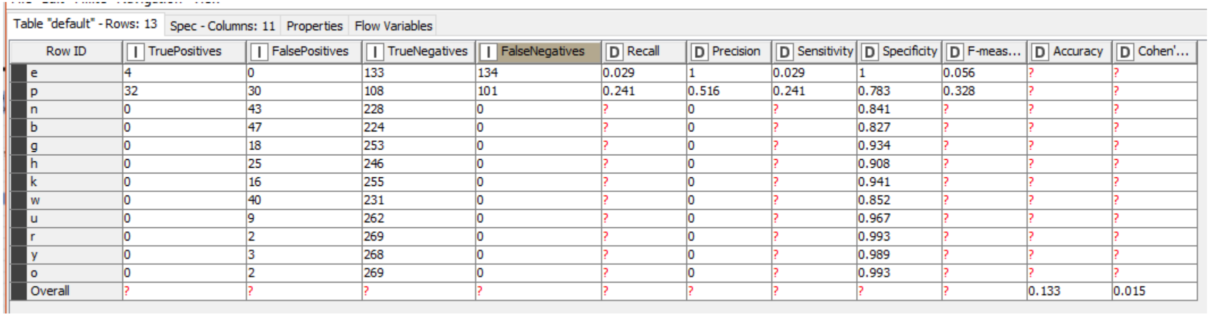
Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
p	23	30	121	97	0.192	0.434	0.192	0.801	0.266	?	?
e	2	0	120	149	0.013	1	0.013	1	0.026	?	?
n	0	27	244	0	?	0	?	0.9	?	?	?
k	0	13	258	0	?	0	?	0.952	?	?	?
w	0	44	227	0	?	0	?	0.838	?	?	?
h	0	17	254	0	?	0	?	0.937	?	?	?
b	0	54	217	0	?	0	?	0.801	?	?	?
y	0	5	266	0	?	0	?	0.982	?	?	?
u	0	22	249	0	?	0	?	0.919	?	?	?
g	0	29	242	0	?	0	?	0.893	?	?	?
o	0	4	267	0	?	0	?	0.985	?	?	?
r	0	1	270	0	?	0	?	0.996	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.092	0.002

Figure 8: Accuracy Statistics-Regression Model Test 1



Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Cohen's kappa	Accuracy
e	1	0	129	141	0.007	1	0.007	1	0.014	?	?
p	16	35	107	113	0.124	0.314	0.124	0.754	0.178	?	?
g	0	36	235	0	?	0	?	0.867	?	?	?
b	0	58	213	0	?	0	?	0.786	?	?	?
h	0	23	248	0	?	0	?	0.915	?	?	?
w	0	42	229	0	?	0	?	0.845	?	?	?
n	0	27	244	0	?	0	?	0.9	?	?	?
y	0	4	267	0	?	0	?	0.985	?	?	?
u	0	15	256	0	?	0	?	0.945	?	?	?
k	0	12	259	0	?	0	?	0.956	?	?	?
r	0	1	270	0	?	0	?	0.996	?	?	?
o	0	1	270	0	?	0	?	0.996	?	?	?
Overall	?	?	?	?	?	?	?	?	?	-0.032	0.063

Figure 9: Accuracy Statistics-Regression Model Test 2



Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
e	4	0	133	134	0.029	1	0.029	1	0.056	?	?
p	32	30	108	101	0.241	0.516	0.241	0.783	0.328	?	?
n	0	43	228	0	?	0	?	0.841	?	?	?
b	0	47	224	0	?	0	?	0.827	?	?	?
g	0	18	253	0	?	0	?	0.934	?	?	?
h	0	25	246	0	?	0	?	0.908	?	?	?
k	0	16	255	0	?	0	?	0.941	?	?	?
w	0	40	231	0	?	0	?	0.852	?	?	?
u	0	9	262	0	?	0	?	0.967	?	?	?
r	0	2	269	0	?	0	?	0.993	?	?	?
y	0	3	268	0	?	0	?	0.989	?	?	?
o	0	2	269	0	?	0	?	0.993	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.133	0.015

Figure 10: Accuracy Statistics-Regression Model Test 3

Task 3:

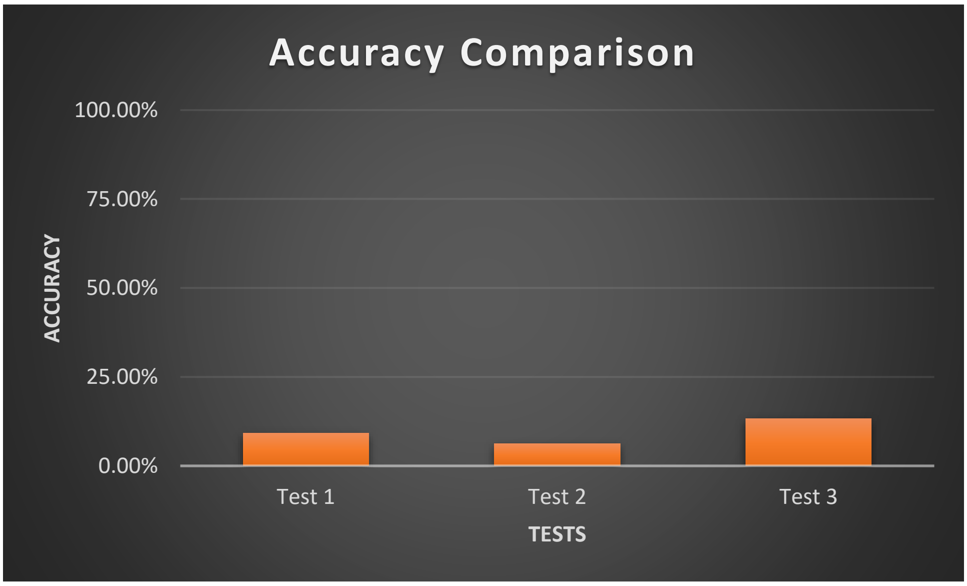


Figure 11: Accuracy Comparison- Bar chart

The accuracy comparison among the 3 test results is given above in the form of a bar chart. We can come to a conclusion looking at the model that regression model is not suitable for this dataset looking at the accuracy rate.

Referencing:

1. Mushroom Classification, viewed by 14 May 2021 < https://www.kaggle.com/imchentouf/mushroom-classification?fbclid=IwAR2y57i2Wm8-NzXazHz4HiGpUdNgIQlfxVa3mWI2jWshT5zA3vDW02scpuM >