CIND 119
Introduction to Big Data Analytics
Assignment 1

Name - Ashwin David
Student Number - 500830814
Section - DK0
Due date - February 28, 2021

1. Download the breast-cancer-dataset.csv from your D2L Assignment 1 link. Complete the following tasks (5 points):

a. Read the file in SAS and display the contents using the import and print procedures. (1 point)

Code:

```
proc import
datafile = "V:\CIND119\Assignment_1\breast_cancer_dataset.csv"
out = breast_cancer
dbms = csv replace;
getnames = yes;
run;
proc print data = breast_cancer;
title "Breast Cancer Dataset";
run;

/* clean the data and remove duplicates */
data breast_cancer_clean;
set breast_cancer;

proc sort data = breast_cancer_clean out = breast_cancer_clean;
by node_caps;
run;

ods pdf file =
"V:\CIND119\Assignment_1\breast_cancer_dataset.rtf";
proc print data = breast_cancer_clean;
title "Breast Cancer Dataset";
run;
ods pdf close;
```

Answer:

Obs	class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig
1	no-recurrence-events	40-49	premeno	25-29	0-2	?	2
2	no-recurrence-events	60-69	ge40	25-29	3-5	?	1
3	no-recurrence-events	60-69	ge40	25-29	3-5	?	1
4	no-recurrence-events	50-59	ge40	30-34	9-1	?	3
5	no-recurrence-events	50-59	ge40	30-34	9-1	?	3
6	recurrence-events	70-79	ge40	15-19	9-1	?	1

7	recurrence-events	50-59	lt40	20-24	0-2	?	1
8	recurrence-events	50-59	lt40	20-24	0-2	?	1
9	no-recurrence-events	30-39	premeno	30-34	0-2	no	3
10	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
11	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
12	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
13	no-recurrence-events	40-49	premeno	0-4	0-2	no	2
14	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
15	no-recurrence-events	50-59	premeno	25-29	0-2	no	2
16	no-recurrence-events	60-69	ge40	20-24	0-2	no	1
17	no-recurrence-events	40-49	premeno	50-54	0-2	no	2
18	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
19	no-recurrence-events	40-49	premeno	0-4	0-2	no	3
20	no-recurrence-events	50-59	ge40	25-29	0-2	no	2
21	no-recurrence-events	60-69	lt40	10-14	0-2	no	1
22	no-recurrence-events	50-59	ge40	25-29	0-2	no	3
23	no-recurrence-events	40-49	premeno	30-34	0-2	no	3
24	no-recurrence-events	60-69	lt40	30-34	0-2	no	1
25	no-recurrence-events	40-49	premeno	15-19	0-2	no	2
26	no-recurrence-events	50-59	premeno	30-34	0-2	no	3
27	no-recurrence-events	60-69	ge40	30-34	0-2	no	3
28	no-recurrence-events	50-59	ge40	30-34	0-2	no	1
29	no-recurrence-events	50-59	ge40	40-44	0-2	no	2
30	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
31	no-recurrence-events	30-39	premeno	25-29	0-2	no	2
32	no-recurrence-events	50-59	premeno	40-44	0-2	no	2
33	no-recurrence-events	50-59	premeno	35-39	0-2	no	2
34	no-recurrence-events	40-49	premeno	25-29	0-2	no	2
35	no-recurrence-events	50-59	premeno	20-24	0-2	no	1
36	no-recurrence-events	60-69	ge40	25-29	0-2	no	3
37	no-recurrence-events	40-49	premeno	40-44	0-2	no	2
38	no-recurrence-events	60-69	ge40	30-34	0-2	no	2
39	no-recurrence-events	50-59	ge40	40-44	0-2	no	3
40	no-recurrence-events	50-59	premeno	15-19	0-2	no	2

41	no-recurrence-events	50-59	premeno	10-14	0-2	no	3
42	no-recurrence-events	50-59	ge40	10-14	0-2	no	1
43	no-recurrence-events	50-59	ge40	10-14	0-2	no	1
44	no-recurrence-events	30-39	premeno	30-34	0-2	no	2
45	no-recurrence-events	50-59	ge40	0-4	0-2	no	2
46	no-recurrence-events	50-59	ge40	15-19	0-2	no	1
47	no-recurrence-events	40-49	premeno	10-14	0-2	no	2
48	no-recurrence-events	40-49	premeno	30-34	0-2	no	1
49	no-recurrence-events	50-59	ge40	20-24	0-2	no	1
50	no-recurrence-events	60-69	ge40	25-29	0-2	no	2
51	no-recurrence-events	60-69	ge40	5-9	0-2	no	1
52	no-recurrence-events	40-49	premeno	10-14	0-2	no	2
53	no-recurrence-events	50-59	ge40	50-54	0-2	no	1
54	no-recurrence-events	50-59	ge40	30-34	0-2	no	1
55	no-recurrence-events	40-49	premeno	25-29	0-2	no	2
56	no-recurrence-events	50-59	premeno	25-29	0-2	no	1
57	no-recurrence-events	40-49	premeno	20-24	0-2	no	1
58	no-recurrence-events	40-49	premeno	20-24	0-2	no	1
59	no-recurrence-events	50-59	lt40	15-19	0-2	no	2
60	no-recurrence-events	30-39	premeno	20-24	0-2	no	2
61	no-recurrence-events	50-59	premeno	15-19	0-2	no	1
62	no-recurrence-events	70-79	ge40	20-24	0-2	no	3
63	no-recurrence-events	70-79	ge40	40-44	0-2	no	1
64	no-recurrence-events	70-79	ge40	40-44	0-2	no	1
65	no-recurrence-events	50-59	ge40	0-4	0-2	no	1
66	no-recurrence-events	50-59	ge40	5-9	0-2	no	2
67	no-recurrence-events	60-69	ge40	30-34	0-2	no	1
68	no-recurrence-events	60-69	ge40	15-19	0-2	no	1
69	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
70	no-recurrence-events	40-49	premeno	10-14	0-2	no	1
71	no-recurrence-events	50-59	ge40	0-4	0-2	no	1
72	no-recurrence-events	20-29	premeno	35-39	0-2	no	2
73	no-recurrence-events	40-49	premeno	25-29	0-2	no	1
74	no-recurrence-events	40-49	premeno	10-14	0-2	no	1

75	no-recurrence-events	40-49	premeno	25-29	0-2	no	1
76	no-recurrence-events	50-59	ge40	20-24	0-2	no	3
77	no-recurrence-events	50-59	ge40	35-39	0-2	no	3
78	no-recurrence-events	60-69	ge40	50-54	0-2	no	2
79	no-recurrence-events	60-69	ge40	10-14	0-2	no	1
80	no-recurrence-events	40-49	premeno	25-29	0-2	no	2
81	no-recurrence-events	60-69	ge40	20-24	0-2	no	2
82	no-recurrence-events	50-59	premeno	15-19	0-2	no	2
83	no-recurrence-events	30-39	premeno	5-9	0-2	no	2
84	no-recurrence-events	50-59	ge40	10-14	0-2	no	1
85	no-recurrence-events	50-59	ge40	10-14	0-2	no	2
86	no-recurrence-events	30-39	premeno	25-29	0-2	no	1
87	no-recurrence-events	50-59	premeno	25-29	0-2	no	2
88	no-recurrence-events	40-49	premeno	25-29	0-2	no	2
89	no-recurrence-events	50-59	ge40	10-14	0-2	no	2
90	no-recurrence-events	60-69	ge40	10-14	0-2	no	1
91	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
92	no-recurrence-events	50-59	ge40	15-19	0-2	no	2
93	no-recurrence-events	40-49	premeno	20-24	0-2	no	1
94	no-recurrence-events	50-59	ge40	35-39	0-2	no	3
95	no-recurrence-events	60-69	ge40	25-29	0-2	no	2
96	no-recurrence-events	70-79	ge40	0-4	0-2	no	1
97	no-recurrence-events	50-59	ge40	20-24	0-2	no	3
98	no-recurrence-events	40-49	premeno	40-44	0-2	no	1
99	no-recurrence-events	30-39	premeno	0-4	0-2	no	2
100	no-recurrence-events	50-59	ge40	20-24	0-2	no	3
101	no-recurrence-events	50-59	ge40	25-29	0-2	no	2
102	no-recurrence-events	60-69	ge40	20-24	0-2	no	2
103	no-recurrence-events	50-59	premeno	10-14	0-2	no	1
104	no-recurrence-events	40-49	premeno	30-34	0-2	no	2
105	no-recurrence-events	60-69	ge40	30-34	0-2	no	2
106	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
107	no-recurrence-events	40-49	premeno	30-34	0-2	no	1
108	no-recurrence-events	30-39	premeno	25-29	0-2	no	2

109	no-recurrence-events	40-49	ge40	20-24	0-2	no	3
110	no-recurrence-events	50-59	ge40	30-34	0-2	no	3
111	no-recurrence-events	50-59	premeno	25-29	0-2	no	2
112	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
113	no-recurrence-events	40-49	premeno	10-14	0-2	no	2
114	no-recurrence-events	40-49	premeno	30-34	0-2	no	1
115	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
116	no-recurrence-events	30-39	premeno	40-44	0-2	no	2
117	no-recurrence-events	40-49	premeno	30-34	0-2	no	3
118	no-recurrence-events	60-69	ge40	30-34	0-2	no	1
119	no-recurrence-events	50-59	ge40	25-29	0-2	no	1
120	no-recurrence-events	50-59	ge40	15-19	0-2	no	1
121	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
122	no-recurrence-events	40-49	premeno	10-14	0-2	no	1
123	no-recurrence-events	40-49	premeno	35-39	0-2	no	2
124	no-recurrence-events	50-59	ge40	20-24	0-2	no	2
125	no-recurrence-events	30-39	premeno	15-19	0-2	no	1
126	no-recurrence-events	40-49	ge40	20-24	0-2	no	3
127	no-recurrence-events	30-39	premeno	10-14	0-2	no	1
128	no-recurrence-events	60-69	ge40	15-19	0-2	no	1
129	no-recurrence-events	60-69	ge40	20-24	0-2	no	1
130	no-recurrence-events	50-59	ge40	15-19	0-2	no	2
131	no-recurrence-events	50-59	ge40	40-44	0-2	no	3
132	no-recurrence-events	50-59	ge40	30-34	0-2	no	1
133	no-recurrence-events	60-69	ge40	10-14	0-2	no	1
134	no-recurrence-events	70-79	ge40	10-14	0-2	no	2
135	no-recurrence-events	40-49	premeno	30-34	6-8	no	2
136	no-recurrence-events	50-59	ge40	40-44	0-2	no	3
137	no-recurrence-events	60-69	ge40	30-34	0-2	no	2
138	no-recurrence-events	30-39	premeno	20-24	3-5	no	2
139	no-recurrence-events	30-39	premeno	40-44	3-5	no	3
140	no-recurrence-events	40-49	premeno	5-9	0-2	no	1
141	no-recurrence-events	30-39	premeno	40-44	0-2	no	2
142	no-recurrence-events	40-49	premeno	30-34	0-2	no	2

143	no-recurrence-events	60-69	ge40	10-14	0-2	no	1
144	no-recurrence-events	40-49	premeno	45-49	0-2	no	2
145	no-recurrence-events	60-69	ge40	50-54	0-2	no	2
146	no-recurrence-events	30-39	premeno	20-24	0-2	no	3
147	no-recurrence-events	50-59	lt40	30-34	0-2	no	3
148	no-recurrence-events	50-59	ge40	35-39	15-	no	3
149	no-recurrence-events	60-69	ge40	15-19	0-2	no	3
150	no-recurrence-events	30-39	lt40	15-19	0-2	no	3
151	no-recurrence-events	60-69	ge40	40-44	3-5	no	2
152	no-recurrence-events	50-59	premeno	30-34	0-2	no	1
153	no-recurrence-events	50-59	ge40	30-34	0-2	no	1
154	no-recurrence-events	40-49	premeno	35-39	0-2	no	1
155	no-recurrence-events	40-49	premeno	25-29	0-2	no	3
156	no-recurrence-events	60-69	ge40	10-14	0-2	no	2
157	no-recurrence-events	40-49	premeno	20-24	3-5	no	2
158	no-recurrence-events	40-49	premeno	20-24	3-5	no	2
159	no-recurrence-events	50-59	premeno	10-14	0-2	no	2
160	no-recurrence-events	40-49	ge40	30-34	0-2	no	2
161	no-recurrence-events	30-39	premeno	15-19	0-2	no	1
162	no-recurrence-events	40-49	premeno	30-34	0-2	no	2
163	no-recurrence-events	40-49	ge40	25-29	0-2	no	2
164	no-recurrence-events	60-69	ge40	10-14	0-2	no	2
165	no-recurrence-events	50-59	premeno	25-29	3-5	no	2
166	no-recurrence-events	40-49	premeno	20-24	0-2	no	3
167	no-recurrence-events	40-49	premeno	25-29	0-2	no	1
168	no-recurrence-events	40-49	premeno	20-24	6-8	no	2
169	no-recurrence-events	50-59	ge40	25-29	0-2	no	1
170	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
171	no-recurrence-events	40-49	premeno	10-14	0-2	no	2
172	no-recurrence-events	40-49	premeno	15-19	12-	no	3
173	no-recurrence-events	40-49	premeno	25-29	0-2	no	2
174	no-recurrence-events	30-39	premeno	10-14	0-2	no	2
175	no-recurrence-events	50-59	ge40	35-39	0-2	no	2
176	no-recurrence-events	50-59	premeno	10-14	3-5	no	1

177	no-recurrence-events	40-49	premeno	10-14	0-2	no	2
178	no-recurrence-events	50-59	premeno	25-29	0-2	no	1
179	no-recurrence-events	60-69	ge40	25-29	0-2	no	3
180	recurrence-events	50-59	premeno	15-19	0-2	no	2
181	recurrence-events	40-49	premeno	40-44	0-2	no	1
182	recurrence-events	50-59	ge40	35-39	0-2	no	2
183	recurrence-events	50-59	premeno	25-29	0-2	no	2
184	recurrence-events	30-39	premeno	0-4	0-2	no	2
185	recurrence-events	50-59	ge40	30-34	0-2	no	3
186	recurrence-events	50-59	premeno	25-29	0-2	no	2
187	recurrence-events	50-59	premeno	30-34	0-2	no	3
188	recurrence-events	40-49	premeno	35-39	0-2	no	1
189	recurrence-events	40-49	premeno	20-24	0-2	no	2
190	recurrence-events	50-59	ge40	20-24	0-2	no	2
191	recurrence-events	40-49	premeno	30-34	0-2	no	3
192	recurrence-events	50-59	premeno	25-29	0-2	no	1
193	recurrence-events	60-69	ge40	40-44	0-2	no	2
194	recurrence-events	40-49	ge40	20-24	0-2	no	2
195	recurrence-events	50-59	ge40	20-24	0-2	no	2
196	recurrence-events	40-49	premeno	15-19	0-2	no	2
197	recurrence-events	60-69	ge40	30-34	0-2	no	3
198	recurrence-events	30-39	premeno	15-19	0-2	no	1
199	recurrence-events	40-49	premeno	25-29	0-2	no	3
200	recurrence-events	30-39	premeno	30-34	0-2	no	1
201	recurrence-events	60-69	ge40	25-29	0-2	no	3
202	recurrence-events	60-69	ge40	20-24	0-2	no	3
203	recurrence-events	40-49	ge40	20-24	3-5	no	3
204	recurrence-events	50-59	premeno	30-34	0-2	no	3
205	recurrence-events	60-69	ge40	45-49	0-2	no	1
206	recurrence-events	40-49	premeno	30-34	3-5	no	2
207	recurrence-events	30-39	premeno	30-34	3-5	no	3
208	recurrence-events	60-69	ge40	30-34	0-2	no	3
209	recurrence-events	40-49	premeno	25-29	0-2	no	2
210	recurrence-events	40-49	premeno	25-29	0-2	no	2

211	recurrence-events	30-39	premeno	35-39	0-2	no	3
212	recurrence-events	60-69	ge40	20-24	3-5	no	2
213	recurrence-events	50-59	ge40	25-29	6-8	no	3
214	recurrence-events	30-39	premeno	30-34	9-1	no	2
215	recurrence-events	40-49	premeno	25-29	0-2	no	3
216	recurrence-events	40-49	premeno	50-54	0-2	no	2
217	recurrence-events	30-39	premeno	40-44	0-2	no	1
218	recurrence-events	60-69	ge40	50-54	0-2	no	3
219	recurrence-events	40-49	premeno	30-34	0-2	no	1
220	recurrence-events	50-59	ge40	30-34	3-5	no	3
221	recurrence-events	60-69	ge40	25-29	3-5	no	2
222	recurrence-events	60-69	ge40	25-29	0-2	no	3
223	recurrence-events	30-39	premeno	35-39	0-2	no	3
224	recurrence-events	40-49	premeno	25-29	0-2	no	2
225	recurrence-events	50-59	premeno	25-29	0-2	no	3
226	recurrence-events	30-39	premeno	30-34	0-2	no	2
227	recurrence-events	30-39	premeno	20-24	0-2	no	3
228	recurrence-events	60-69	ge40	20-24	0-2	no	1
229	recurrence-events	40-49	ge40	30-34	3-5	no	3
230	recurrence-events	50-59	ge40	30-34	3-5	no	3
231	no-recurrence-events	30-39	premeno	30-34	6-8	ye	2
232	no-recurrence-events	30-39	premeno	25-29	6-8	ye	2
233	no-recurrence-events	50-59	premeno	25-29	0-2	ye	2
234	no-recurrence-events	40-49	premeno	35-39	9-1	ye	2
235	no-recurrence-events	40-49	premeno	35-39	9-1	ye	2
236	no-recurrence-events	40-49	premeno	40-44	3-5	ye	3
237	no-recurrence-events	50-59	ge40	40-44	3-5	ye	2
238	no-recurrence-events	50-59	premeno	20-24	3-5	ye	2
239	no-recurrence-events	60-69	ge40	45-49	6-8	ye	3
240	no-recurrence-events	50-59	premeno	30-34	3-5	ye	2
241	no-recurrence-events	50-59	ge40	25-29	15-	ye	3
242	no-recurrence-events	60-69	ge40	30-34	3-5	ye	3
243	no-recurrence-events	50-59	ge40	25-29	3-5	ye	3
244	no-recurrence-events	40-49	premeno	30-34	3-5	ye	2

245	no-recurrence-events	40-49	ge40	40-44	15-	ye	2
246	no-recurrence-events	30-39	premeno	20-24	3-5	ye	2
247	no-recurrence-events	60-69	ge40	30-34	6-8	ye	2
248	no-recurrence-events	50-59	ge40	20-24	3-5	ye	2
249	no-recurrence-events	50-59	premeno	25-29	3-5	ye	2
250	no-recurrence-events	40-49	premeno	35-39	0-2	ye	3
251	no-recurrence-events	40-49	premeno	35-39	0-2	ye	3
252	no-recurrence-events	50-59	ge40	20-24	0-2	ye	2
253	no-recurrence-events	50-59	ge40	30-34	6-8	ye	2
254	no-recurrence-events	50-59	premeno	50-54	0-2	ye	2
255	no-recurrence-events	50-59	ge40	15-19	0-2	ye	2
256	recurrence-events	30-39	premeno	25-29	3-5	ye	3
257	recurrence-events	40-49	premeno	30-34	15-	ye	3
258	recurrence-events	60-69	ge40	40-44	3-5	ye	3
259	recurrence-events	50-59	premeno	50-54	9-1	ye	2
260	recurrence-events	50-59	premeno	25-29	3-5	ye	3
261	recurrence-events	40-49	premeno	20-24	3-5	ye	2
262	recurrence-events	40-49	premeno	15-19	15-	ye	3
263	recurrence-events	50-59	ge40	20-24	3-5	ye	3
264	recurrence-events	40-49	premeno	30-34	12-	ye	3
265	recurrence-events	30-39	premeno	15-19	6-8	ye	3
266	recurrence-events	50-59	ge40	30-34	9-1	ye	3
267	recurrence-events	60-69	ge40	35-39	6-8	ye	3
268	recurrence-events	30-39	premeno	20-24	3-5	ye	2
269	recurrence-events	40-49	premeno	30-34	0-2	ye	3
270	recurrence-events	40-49	premeno	30-34	6-8	ye	3
271	recurrence-events	40-49	premeno	20-24	3-5	ye	2
272	recurrence-events	50-59	ge40	30-34	6-8	ye	2
273	recurrence-events	40-49	ge40	25-29	12-	ye	3
274	recurrence-events	30-39	premeno	35-39	9-1	ye	3
275	recurrence-events	40-49	premeno	30-34	3-5	ye	2
276	recurrence-events	60-69	ge40	20-24	24-	ye	3
277	recurrence-events	50-59	ge40	30-34	6-8	ye	3
278	recurrence-events	40-49	premeno	15-19	0-2	ye	3

279	recurrence-events	60-69	ge40	30-34	0-2	ye	2
280	recurrence-events	60-69	ge40	30-34	3-5	ye	2
281	recurrence-events	40-49	premeno	25-29	9-1	ye	3
282	recurrence-events	30-39	premeno	25-29	6-8	ye	3
283	recurrence-events	60-69	ge40	10-14	6-8	ye	3
284	recurrence-events	50-59	premeno	35-39	15-	ye	3
285	recurrence-events	50-59	ge40	40-44	6-8	ye	3
286	recurrence-events	50-59	ge40	40-44	6-8	ye	3

Obs	breast	breast_quad	irradiat
1	left	right_low	ye
2	right	left_up	ye
3	right	left_low	ye
4	left	left_up	ye
5	left	left_low	ye
6	left	left_low	ye
7	left	left_up	no
8	left	left_low	no
9	left	left_low	no
10	right	right_up	no
11	left	left_low	no
12	right	left_up	no
13	right	right_low	no
14	left	left_low	no
15	left	left_low	no
16	left	left_low	no
17	left	left_low	no
18	right	left_up	no
19	left	central	no
20	left	left_low	no
21	left	right_up	no
22	left	right_up	no
23	left	left_up	no
24	left	left_low	no

25	left	left_low	no
26	left	left_low	no
27	left	left_low	no
28	right	right_up	no
29	left	left_low	no
30	left	left_low	no
31	right	left_low	no
32	left	left_up	no
33	right	left_up	no
34	left	left_up	no
35	left	left_low	no
36	right	left_up	no
37	right	left_low	no
38	left	left_low	no
39	right	left_up	no
40	right	left_low	no
41	left	left_low	no
42	right	left_up	no
43	left	left_up	no
44	left	left_up	no
45	left	central	no
46	right	central	no
47	left	left_low	no
48	left	left_low	no
49	right	left_low	no
50	left	left_low	no
51	left	central	no
52	left	left_up	no
53	right	right_up	no
54	left	left_up	no
55	right	left_low	no
56	right	left_up	no
57	right	right_up	no
58	right	left_low	no

59	left	left_low	no
60	left	right_low	no
61	left	left_low	no
62	left	left_up	no
63	right	left_up	no
64	right	right_up	no
65	right	central	no
66	right	right_up	no
67	left	left_up	no
68	right	left_up	no
69	left	central	no
70	right	right_low	no
71	left	left_low	no
72	right	right_up	no
73	left	right_low	no
74	right	left_up	no
75	right	right_low	no
76	left	left_up	no
77	left	left_low	no
78	left	left_low	no
79	left	left_low	no
80	right	left_up	no
81	left	left_up	no
82	right	right_low	no
83	left	right_low	no
84	left	left_low	no
85	left	left_low	no
86	left	central	no
87	left	left_low	no
88	right	central	no
89	right	left_low	no
90	left	left_up	no
91	right	left_low	no
92	right	left_low	no

93	left	right_low	no
94	left	left_up	no
95	right	left_low	no
96	left	right_low	no
97	right	left_up	no
98	right	left_up	no
99	right	central	no
100	left	left_up	no
101	right	left_up	no
102	right	left_up	no
103	left	left_low	no
104	right	right_low	no
105	left	left_up	no
106	right	left_up	no
107	left	right_up	no
108	left	left_low	no
109	left	left_low	no
110	right	left_low	no
111	right	right_low	no
112	left	right_low	no
113	right	left_low	no
114	right	left_up	no
115	left	left_up	no
116	right	right_up	no
117	right	right_up	no
118	right	left_up	no
119	left	left_low	no
120	right	central	no
121	right	left_up	no
122	right	left_up	no
123	right	right_up	no
124	right	left_up	no
125	left	left_low	no
126	left	left_up	no

127	right	left_low	no
128	left	right_low	no
129	left	left_low	no
130	right	right_up	no
131	left	left_up	no
132	right	left_low	no
133	right	left_low	no
134	left	central	no
135	left	left_up	no
136	left	right_up	no
137	left	left_low	ye
138	right	central	no
139	right	right_up	ye
140	left	left_low	ye
141	left	left_low	ye
142	left	right_low	no
143	left	left_up	no
144	left	left_low	ye
145	right	left_up	ye
146	left	central	no
147	right	left_up	no
148	left	left_low	no
149	right	left_up	ye
150	right	left_up	no
151	right	left_up	ye
152	left	central	no
153	right	central	no
154	left	left_low	no
155	right	left_up	ye
156	right	left_up	ye
157	right	left_up	no
158	right	left_low	no
159	right	left_up	no
160	left	left_up	ye

161	left	left_low	no
162	right	right_up	ye
163	left	left_low	no
164	left	left_low	no
165	right	left_up	ye
166	right	left_low	ye
167	right	left_low	ye
168	right	left_low	ye
169	left	right_low	no
170	left	left_up	ye
171	right	left_up	no
172	right	right_low	ye
173	left	left_up	ye
174	left	right_low	no
175	left	left_up	no
176	right	left_up	no
177	left	left_low	ye
178	left	left_low	no
179	right	left_low	no
180	left	left_low	no
181	left	left_low	no
182	left	left_low	no
183	left	right_up	no
184	right	central	no
185	left	?	no
186	left	right_up	no
187	left	right_up	no
188	right	left_up	no
189	left	left_low	no
190	right	central	no
191	right	right_up	no
192	right	left_up	no
193	right	left_low	no
194	right	left_up	no

195	left	left_up	no
196	left	left_up	no
197	right	central	no
198	right	left_low	no
199	left	right_up	no
200	right	left_up	no
201	left	right_low	ye
202	right	left_low	no
203	right	left_low	ye
204	right	left_up	ye
205	right	right_up	ye
206	right	left_up	no
207	right	left_up	ye
208	right	left_up	ye
209	right	left_low	no
210	right	left_low	no
211	left	left_low	no
212	left	left_low	ye
213	left	left_low	ye
214	right	left_up	ye
215	left	left_up	no
216	right	left_low	ye
217	left	left_up	no
218	right	left_up	no
219	left	left_low	ye
220	right	left_up	no
221	right	right_up	no
222	left	left_up	no
223	left	left_low	no
224	left	left_low	ye
225	right	left_low	ye
226	left	left_up	no
227	left	left_up	ye
228	right	left_up	no

229	left	left_low	no
230	left	left_low	no
231	right	right_up	no
232	right	left_up	ye
233	left	left_up	no
234	right	left_up	ye
235	right	right_up	ye
236	right	left_up	ye
237	left	left_low	no
238	left	left_low	no
239	left	central	no
240	left	left_low	ye
241	right	left_up	no
242	left	left_low	no
243	right	left_up	no
244	right	left_low	no
245	right	left_up	ye
246	right	left_up	ye
247	right	right_up	no
248	right	left_up	no
249	left	left_low	ye
250	right	left_up	ye
251	right	left_low	ye
252	right	left_up	no
253	left	left_low	no
254	right	left_up	ye
255	left	central	ye
256	left	left_low	ye
257	left	left_low	no
258	right	left_low	no
259	right	left_up	no
260	left	left_low	ye
261	right	right_up	ye
262	left	left_low	no

263	right	right_up	no
264	left	left_up	ye
265	left	left_low	ye
266	left	right_low	ye
267	left	left_low	no
268	left	left_low	no
269	right	right_up	no
270	right	left_up	no
271	left	left_low	ye
272	left	right_low	ye
273	left	right_low	ye
274	left	left_low	no
275	left	right_up	no
276	left	left_low	ye
277	left	right_low	no
278	right	left_up	no
279	right	right_up	ye
280	left	central	ye
281	right	left_up	no
282	left	right_low	ye
283	left	left_up	ye
284	right	right_up	no
285	left	left_low	ye
286	left	left_low	ye

b. Develop a decision tree-based classification model using the hpsplit procedure of SAS. (2 points)

Code:

```
/* Q1. b) Develop a decision tree-based
classification model using the hpsplit procedure of SAS. */
ods graphics on;
ods rtf file =
"V:\CIND119\Assignment_1\breast_cancer_decision_tree.rtf";
proc hpsplit data = breast_cancer_clean;

class class age menopause tumor_size inv_nodes node_caps
deg_malig breast breast_quad irradiat;

model class = age menopause tumor_size inv_nodes node_caps
deg_malig breast breast_quad irradiat;
grow entropy;
```

```
prune costcomplexity;
run;
ods rtf close;
```

Answer:

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	**Engine**	**Role**	**Path**
WORK.BREAST_CANCER_CLEAN	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	10
Tree Depth	2
Number of Leaves Before Pruning	51
Number of Leaves After Pruning	3
Model Event Level	no-recurrence-events

Number of Observations Read	286
Number of Observations Used	286

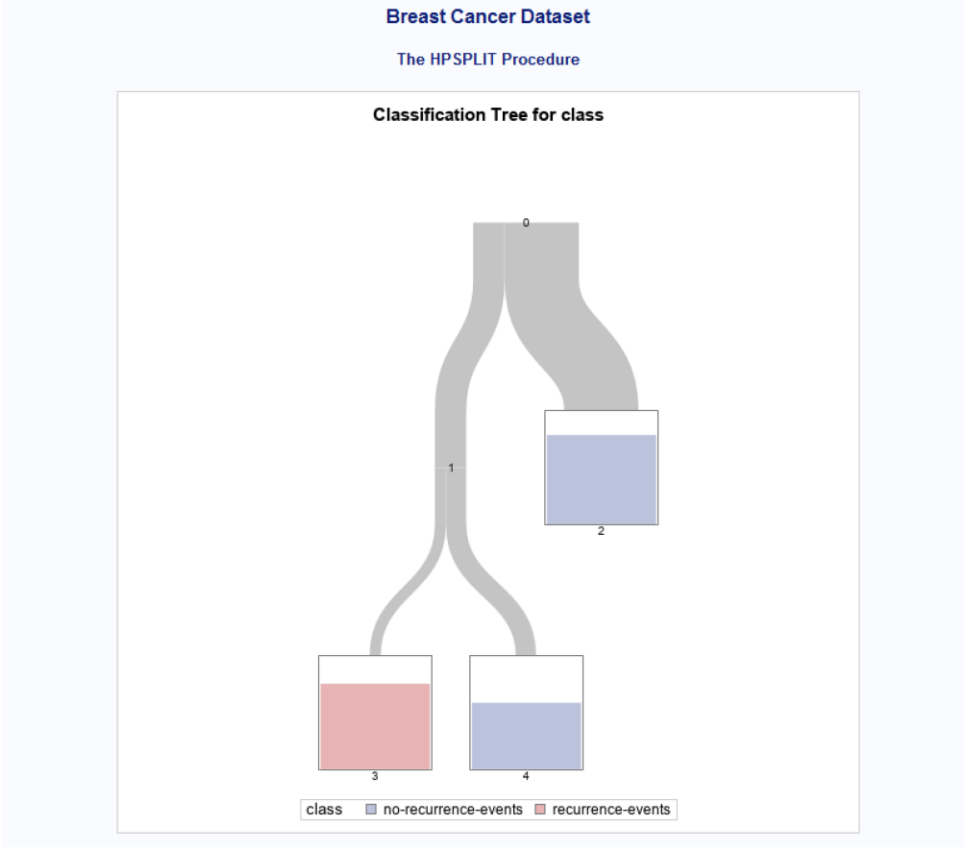

Figure 1: Model for classification Tree

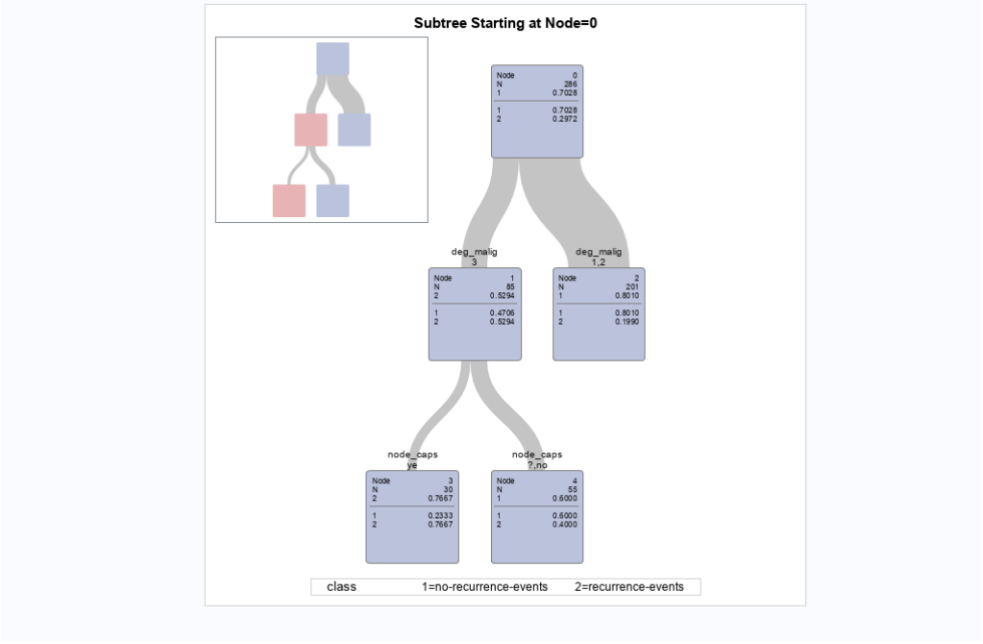

Figure 2: Decision Tree for Breast Cancer Dataset

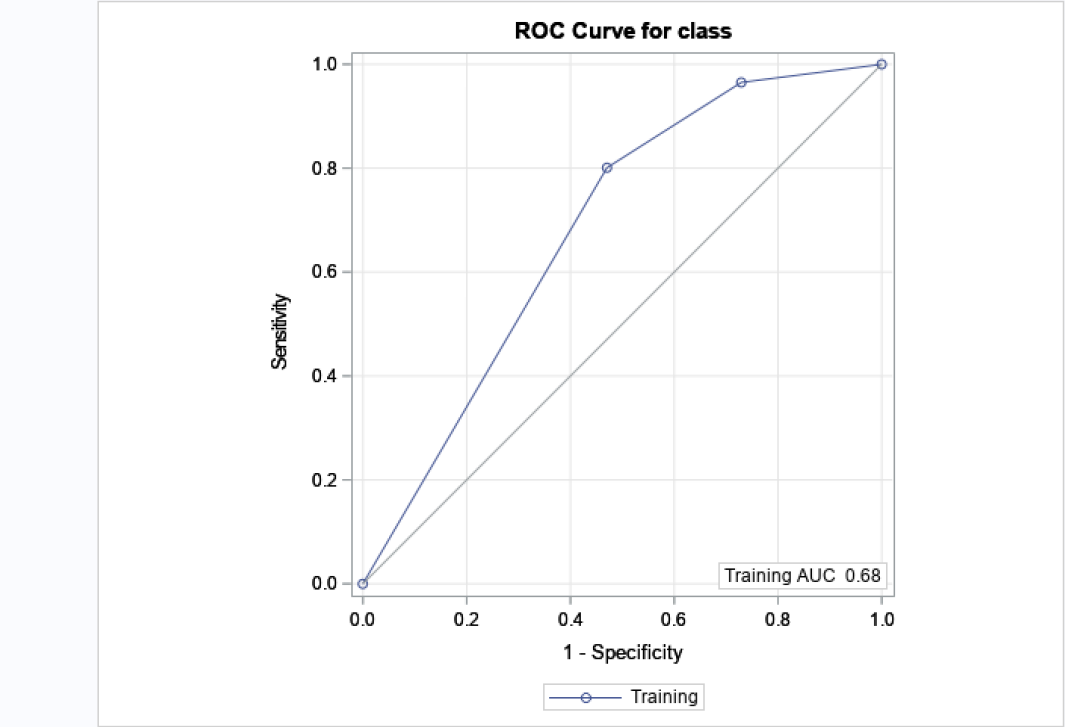

Figure 3: ROC curve for variable class

c. Navigate the contents of Results View by clicking on HPSplit breast-cancer-dataset, and then by selecting Model Assessment. Examine the confusion matrix, fit statistics, and variable importance. (2 points)

Answer:

Model-Based Confusion Matrix			
	Predicted		Error Rate
Actual	no-recurrence-events	recurrence-events	
no-recurrence-events	194	7	0.0348
recurrence-events	62	23	0.7294

Model-Based Fit Statistics for Selected Tree								
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
3	0.1769	0.2413	0.9652	0.2706	0.7749	0.3539	101.2	0.6829

Variable Importance			
	Training		
Variable	Relative	Importance	Count
deg_malig	1.0000	3.6115	1
node_caps	0.6326	2.2846	1

2. Using the confusion matrix, compute the following assessment metrics accuracy, recall, and precision (see lecture for formulas). (5 points)
   Condition for marks: 3 points for accuracy, 1 point for precision, and 1 point for recall.

   Answer:
   Based on the variable "class" that has values of either recurring events or nonrecurring events the confusion matrix was built and the statistics such as accuracy, recall and precision were calculated.

$$Accuracy = a = \frac{TP + TN}{T} = \frac{194 + 23}{286} = 0.7587$$

$$Recall = TPR = \frac{TP}{TP + FN} = \frac{194}{194 + 7} = 0.9652$$

$$Precision = P = \frac{TP}{TP + FP} = \frac{194}{194 + 62} = 0.7578$$

3. Change the grow algorithm to "gini" and recompute the metrics from question 2. Does entropy build a more accurate classifier or gini? (5 points)

   Code:

```
/* Q3. Change the grow algorithm to "gini" and recompute the metrics
from question 2. Does entropy build a more accurate classifier or
gini?
(5 points) */

ods graphics on;
ods rtf file =
"V:\CIND119\Assignment_1\breast_cancer_decision_tree_gini.rtf";
proc hpsplit data = breast_cancer_clean;

class class age menopause tumor_size inv_nodes node_caps
deg_malig breast breast_quad irradiat;

model class = age menopause tumor_size inv_nodes node_caps
deg_malig breast breast_quad irradiat;
grow gini;
prune costcomplexity;
run;
ods rtf close;
```

Answer:

Model-Based Confusion Matrix			
Actual	Predicted		Error Rate
	no-recurrence-events	recurrence-events	
no-recurrence-events	191	10	0.0498
recurrence-events	58	27	0.6824

Model-Based Fit Statistics for Selected Tree								
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
3	0.1769	0.2378	0.9502	0.3176	0.7751	0.3538	101.2	0.6836

Variable Importance			
Variable	Training		Count
	Relative	Importance	
deg_malig	1.0000	3.6115	1
inv_nodes	0.6349	2.2931	1

$$Accuracy = a = \frac{TP + TN}{T} = \frac{191 + 27}{286} = 0.7622$$

$$Recall = TPR = \frac{TP}{TP + FN} = \frac{191}{191 + 10} = 0.9502$$

$$Precision = P = \frac{TP}{TP + FP} = \frac{191}{191 + 58} = 0.7671$$

When comparing the metrics obtained from different growth statistics, gini and entropy, the data obtained in the confusion matrices from question 2 and 3 were used to analyze the preciseness of these metrics. The accuracy of entropy relayed a value of 0.7587 while the growth, using the classifier gini produced an accuracy of 0.7622 with a better precision. The values for all metrics deviated about $1/10^{th}$ of a value providing enough evidence that the classifier gini was more accurate and provided better metrics as opposed to the growth classifier entropy.