Jahangirnagar University
Department of Computer Science and Engineering
PMSCS Semester Final Examination, Summer 2022

Course ID: PMSCS 686 | Course Title: Introduction to Data Science
Time: 3 Hours | Full Marks: 60

Answer any *SIX* of the following questions.

1. 2 + 4 + 4

a) "Data Science is a multi-disciplinary subject" – Prove that the statement is true with the necessary evidence.

b) Determine the responsibility of a data scientist in data processing activities?

c) Write down the differences between structured, semi-structured and unstructured data.

2. 2 + 4 + 4

a) Describe the fields where the data science can be applied effectively.

b) What do you understand by data aggregation? How can data aggregation be used a strong tool for generating meaningful features?

c) Why feature reduction technique is used in data science? Write down the names of different feature reduction techniques.

3. 2 + 3 + 5

a) What is hypothesis testing?

b) What criteria the null and alternate hypothesis must possess?

c) Convert the following word hypothesis into statistical hypothesis – "People who earn an A in C++ Programming are more likely to get a job in Google than those who do not earn an A".

4. 5 + 5

a) The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table. What will be the sales for the expenditure of $67?

Expenditure (x)	Weekly sales (y)
41	1250
54	1380
63	1425
54	1425
48	1450
46	1300
62	1400
61	1510
64	1575
71	1650

b) Given that the area under the normal curve between z-scores of -1.96 and +1.96 is [illegible]

thus, the 95% confidence interval is given by $\overline{X} \pm 1.96$ standard deviation of the sampling distribution. Find out the ***confidence interval*** assuming that $\bar{X} = 4.50$, $s_X = 0.55$, and $n = 35$.

5 + 5

5.

a) What is residual? Why residual plot analysis is important in statistical inference? Justify your answer with some examples.

b) Define the terms "overfitting" and "underfitting". Discuss the various approaches of controlling an overfitting.

2 + 4 + 4

6.

a) Write the important criteria of a logit function.

b) Suppose for a set of patients, if the probability of response for treatment group, Pr(response | trt) = 0.4 and the probability of response for placebo group, Pr(response | placebo) = 0.2. What will be the ODD Ratio between two groups?

c) For an Assisted Reproduction Technology (ART) clinics, one of the main outcomes is clinical pregnancy. There is much empirical evidence that the candidate mother's age is a significant factor that affects the chances of pregnancy success. A recent study examined the effect of the mother's age, along with clinical characteristics, on the odds of pregnancy success on the first ART attempt. The logistic regression model is represented as

$$\ln\left(\frac{\Pr(pregnancy)}{1-\Pr(pregnancy)}\right) = 2.5 - 1.15 * Age$$

(i) What is the effect of Age on Pregnancy?

(ii) What is the predicted probability of a 27-year-old having pregnancy success with first ART attempt?

7. 5 + 5

a) Why *k*-Nearest Neighbor (*k*-NN) classifier is termed as lazy classifier? Write down the problems that are encountered in Euclidian distance measure.

b) Write down the kd-tree construction algorithm. Hence, show that a kd-tree has a depth in the order of $O(log_2 n)$; where n is the number of records.

8. 5 + 5

a) What do you understand by purity in a data set? How do you determine the best split in decision tree construction?

Jahangirnagar University

Department of Computer Science and Engineering
PMSCS Semester Final Examination, Fall 2022

Course ID: PMSCS 686 Course Title: Introduction to Data Science
Time: 3 Hours Full Marks: 60

Answer any SIX questions.

2 + 4 + 4

a) Determine the responsibility of a data scientist in data processing activities.
b) What do you understand by data aggregation? How can data aggregation be used as a strong tool for generating the meaningful features in a dataset?
c) Why feature reduction technique is used in a data set? Write down the names of different feature reduction techniques.

2 + 3 + 5

a) What is hypothesis testing?
b) Write down the criteria of null and alternate hypothesis.
c) Convert the following word hypothesis into statistical hypothesis – "People who earn an A in JAVA Programming are more likely to get a job in Google than those who do not earn an A".

5 + 5

a) The weekly advertising expenditure (x) and weekly sales (y) are presented in the following Table. What will be the sales for the expenditure of $68?

Expenditure (x)	Weekly sales (y)
41	1248
54	1382
63	1435
54	1424
48	1440
46	1310
62	1405
61	1505
64	1555
71	1652

b) Given that the area under the normal curve between z-scores of -1.96 and +1.96 is .95 and thus, the 95% confidence interval is given by $(\bar{X} \pm 1.96)$ standard deviation of the sampling distribution. Find out the ***confidence interval*** assuming that $\bar{X} = 4.50$, $s_X = 0.55$, and $n = 35$.

5 + 5

a) What is residual? Why residual plot analysis is important in statistical inference? Justify your answer with some examples.
b) Define the terms "overfitting" and "underfitting". Discuss the various approaches of controlling an overfitting.

5 + 5

a) What do understand by "maximum likelihood"? For the independent events, show that

$$lnL = \sum_i y_i \beta x_i - \sum_i \ln(1 + \exp(\beta x_i))$$

For an Assisted Reproduction Technology (ART) clinics, one of the main outcomes is clinical pregnancy. There is much empirical evidence that the candidate mother's age is a significant factor that affects the chances of pregnancy success. A recent study examined the effect of the mother's age, along with clinical characteristics, on the odds of pregnancy success on the first ART attempt. The logistic regression model is represented as

$$\ln\left(\frac{\Pr(pregnancy)}{1-\Pr(pregnancy)}\right) = 2.5 - 1.15 * Age$$

(i) What is the effect of Age on Pregnancy?

(ii) What is the predicted probability of a 27-year-old having pregnancy success with first ART attempt?

2 + 4 + 4

(6)

(a) Why do we use *neighbor learning* technique instead of *rote learning* technique?

(b) What types of problem a *k*-NN algorithm may face when the value of K is too small or too big?

(c) "The *k*-NN machine learning classifier is lazy learner" – justify the statement.

3 + 3 + 4

a) Given the samples from two distinct classes, how do you decide about a decision boundary whether it is better or not?

b) A doctor knows that meningitis causes stiff neck 60% of the time with a prior probability of any patient having meningitis is 1/40,000 and a prior probability of any patient having stiff neck is 1/25. If a patient has stiff neck, what is the probability he/she has meningitis?

c) For a given data set, what are the process of finding the *probability of occurrences* when the attribute is *categorical* and when the attribute is *numerical* in Naïve Bayes Classifier?

5 + 5

(a) Describe Hunt's algorithm. Using the algorithm, find a decision tree for the following data set.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(b) What do you understand by Entropy measure? By using entropy, give some demonstration for deciding about the pure and impure data set.

Jahangirnagar University

Department of Computer Science and Engineering
PMSCS Semester Final Examination, Spring 2023

Course ID: PMSCS 686 Course Title: Introduction to Data Science
Time: 3 Hours Full Marks: 60

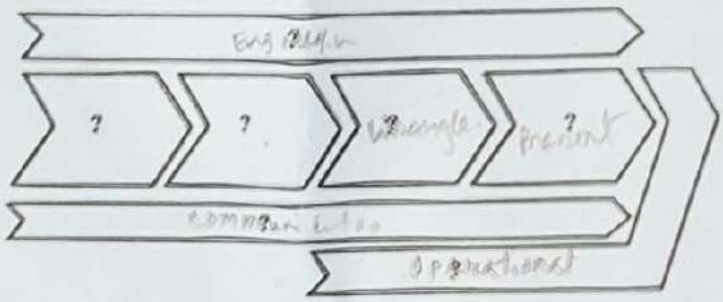
Answer any *SIX* questions.

1. 3 + 3 + 4

a) What do you understand by big data? Is there any relation between big data and data science?

b) Write a comparative study between Python and R in the context of data science project development.

c) Write down the name of components of the below value chain and describe them accordingly.



2. 2 + 4 + 4

a) Why does 50% of analytical projects fail? How the failure can be avoided?

b) Define the term "Outlier value" and "missing value". How you can treat those problems.

c) Though an increased number of feature in data set is very important, we still try to reduce the number of features. Why?

3. 5 + 5

a) What is sampling distribution? How can you determine that one sampling distribution is better than another sampling distribution?

b) Given that the area under the normal curve between z-scores of -1.96 and +1.96 is .95 and thus, the 95% confidence interval is given by $\bar{X} \pm 1.96$ standard deviation of the sampling distribution. Find out the ***confidence interval*** assuming that $\bar{X} = 4.50, S_{\bar{X}} = 0.55, and\ n = 35$.

4. 5 + 5

a) What is the significance of hypothesis testing? Convert the following word hypothesis into statistical hypothesis – "People who earn an A in Java Programming are more likely to get a job in Google than those who do not earn an A".

b) The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table. What will be the sales for the expenditure of $67?

Expenditure (x)	Weekly sales (y)
41	1250
54	1380
63	1425
54	1425
48	1450
46	1300
62	1400
61	1510
64	1575
71	1650

5. 5 + 5

a) Why residual plot analysis is very important in regression model development? Discuss various types of residual plots.

b) When variable transformation is mandatory? Give some popular variable transformation functions used in data analytics.

6. 3 + 2 + 5

a) Suppose for a set of patients, if the probability of response for treatment group, Pr(response | trt) = 0.42 and the probability of response for placebo group, Pr(response | placebo) = 0.24. What will be the ODD Ratio between two groups?

b) With some suitable example, describe the maximum likelihood function.

c) Derive the equation for likelihood, L for N number of independent events, and show that an increase in regression co-efficient, β will surely increase the likelihood, L.

7. 4 + 2 + 4

a) Differentiate between supervised learning and unsupervised learning methods. Write down the strategies of dividing a data set for training and testing purpose.

b) Describe the importance of a margin in a support vector machine. Describe the process when the problem is not linearly separable.

c) For the below training set of records in TABLE-1 and corresponding naïve Bayes classifier, find out the class of a given record X= (Refund=No, Married, Income=120). {Formula of normal distribution is:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

TABLE-1

naive Bayes Classifier:

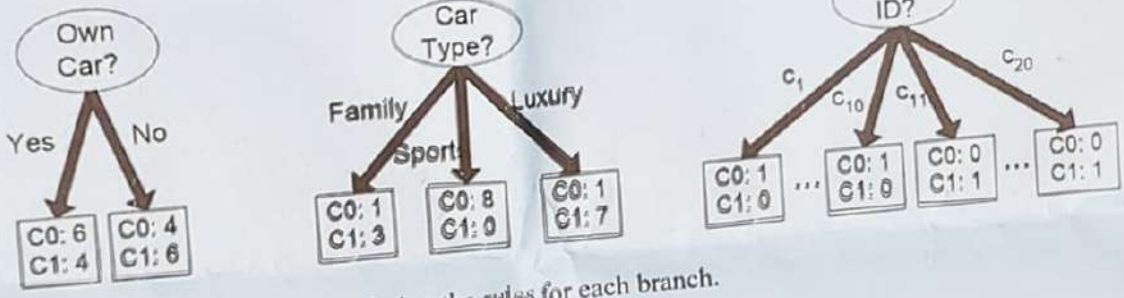
P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

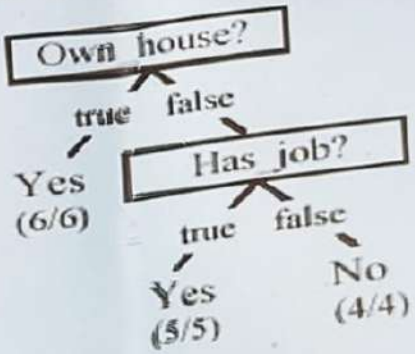
5 + 5

8.

a) For the below test cases, which one is the best? Justify your statement.



b) For the below decision tree, derive the rules for each branch.



Describe its representation in memory.

JAHANGIRNAGAR UNIVERSITY

Department of Computer Science and Engineering

PMSCS Semester Final Examination, Summer - 2023

Course ID: PMSCS 686 (Introduction to Data Science)

Time: 3 Hours Full Marks: 60

Answer any *SIX* questions.

1. $2+4+4$

a) Who is data Scientist? What is his responsibilities?

b) Write down the differences between structured, semi-structured and unstructured data. What type of operations are performed in data cleansing?

c) Describe the various data science applications in different fields of Science and Technology.

2. $2+4+4$

a) Define the term "outlier" and "missing value". How can you treat those problems?

b) Describe the steps for data preparation. What do you understand by data aggregation?

c) Why feature reduction technique is used in data science? Write down the names of different feature reduction techniques. How data set is used in trifold partition technique?

3. $5+5$

a) The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table. What will be the sales for the expenditure of $70?

Expenditure (x)	Weekly sales (y)
41	1250
54	1380
63	1425
54	1425
48	1450
46	1300
62	1400
61	1510
64	1575
71	1650

b) Given that the area under the normal curve between z-scores of -1.96 and +1.96 is .95 and thus, the 95% confidence interval is given by $\bar{X} \pm 1.96$ standard deviation of the sampling distribution. Find out the ***confidence interval*** assuming that $\bar{X} = 4.50, S_X = 0.55, and\ n = 35.$

5 + 5

4. a) What is hypothesis testing? What criteria the null and alternate hypothesis must possess? Convert the following word hypothesis into statistical hypothesis – "People who earn a grade A in computer science are more likely to be admitted to graduate school than those who do not earn a grade A".

b) Write down the importance of the analysis of residuals. Show that residual plots may help to develop a practical linear regression model.

5 + 5

5. a) Derive the equation for likelihood, L for N number of independent events, and show that an increase in regression co-efficient, β will increase the likelihood, L.

b) For an Assisted Reproduction Technology (ART) clinics, one of the main outcomes is clinical pregnancy. There is much empirical evidence that the candidate mother's age is a significant factor that affects the chances of pregnancy success. A recent study examined the effect of the mother's age, along with clinical characteristics, on the odds of pregnancy success on the first ART attempt. The logistic regression model is represented as

$$\ln\left(\frac{\Pr(pregnancy)}{1-\Pr(pregnancy)}\right) = 2.5 - 1.15 * Age$$

(i) What is the effect of Age on Pregnancy?

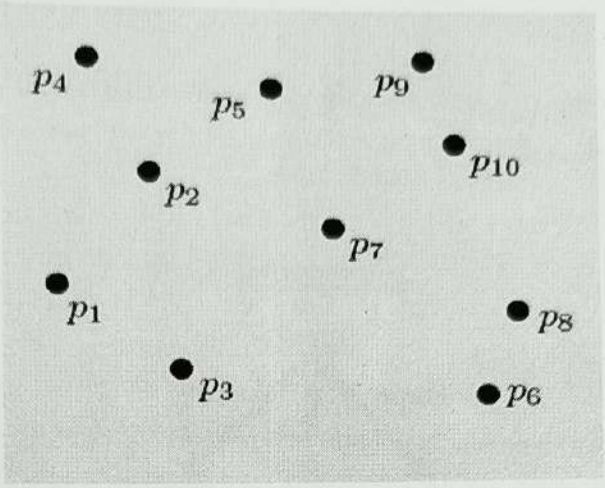
(ii) What is the predicted probability of a 27 year old having pregnancy success with first ART attempt?

5 + 5

6. a) Describe the k-means nearest neighbor classification. Write down the problems that are encountered in Euclidian distance measure.

b) Mention kd-tree construction algorithm; and hence, construct a kd-tree from the following data points in a 2 dimensional space.

Marks: 2



7. $5 + 5$

a) How do you choose a better classification boundary in support vector machine? What will be the necessary measures for a non-linear decision boundary?

b) For the above training set of records in below TABLE, find out the class of a given record X = (Give Birth=yes, Can Fly=no, Live in Water=yes, Have Legs=no).

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

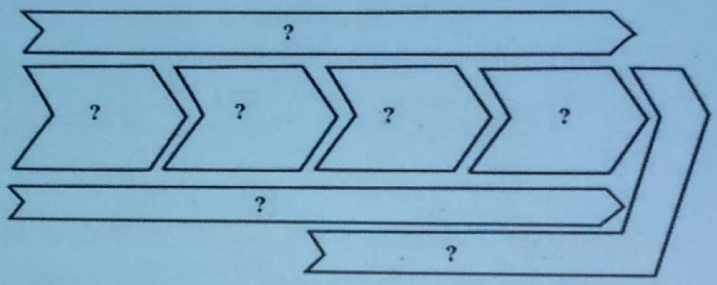
JAHANGIRNAGAR UNIVERSITY
Department of Computer Science and Engineering
PMSCS Term Final Examination, Fall - 2023
Course ID: PMSCS 686 (Introduction to Data Science)
Time: 3 Hours Full Marks: 60

Answer any *SIX* questions.

1. 3 + 7

a) What is big data? Is there any relation between big data and data science?

b) Write down the name of components of the below value chain and describe them accordingly.



2. 3 + 3 + 4

a) Why does 50% of analytical projects fail? How the failure can be avoided?

b) Define the terms "Outlier" and "missing value" and their possible solutions.

c) Though an increased number of feature in data set is very important, we try to reduce the number of features. Why?

3. 2 + 3 + 5

a) Why hypothesis testing is used for?

b) What criteria the null and alternate hypothesis must possess?

c) Convert the following word hypothesis into statistical hypothesis – "People who eat breakfast will run a race faster or slower than those who do not eat breakfast".

4. 3 + 7

a) What do you understand by variable transformation? Why do we use variable transformation in linear regression analysis?

b) For the weekly advertising expenditure and weekly sales table as given below. The manageme team of a company is interested in testing whether or not there is a linear association between advertising expenditure and weekly sales, using a linear regression model. Use $\alpha = .05$. [*T*-distribution table is attested at *Appendix-A*].

Expenditure (x)	Weekly sales (y)
41	1250
54	1380
63	1425
54	1425
48	1450
46	1300
62	1400
61	1510
64	1575
71	1650

5. 5 + 5

a) Suppose for a set of patients, if the probability of response for treatment group, Pr(response | trt) = 0.4 and the probability of response for placebo group, Pr(response | placebo) = 0.2. What will be the ***ODD Ratio*** between two groups?

b) For an Assisted Reproduction Technology (ART) clinics, one of the main outcomes is clinical pregnancy. There is much empirical evidence that the candidate mother's age is a significant factor that affects the chances of pregnancy success. A recent study examined the effect of the mother's age, along with clinical characteristics, on the odds of pregnancy success on the first ART attempt. The logistic regression model is represented as

$$\ln\left(\frac{\Pr(pregnancy)}{1-\Pr(pregnancy)}\right) = 2.5 - 1.15 * Age$$

(i) What is the effect of Age on Pregnancy?

(ii) What is the predicted probability of a 27-year-old having pregnancy success with first ART attempt?

6. 5 + 5

a) Why *k*-Nearest Neighbor (*k*-NN) classifier is termed as lazy classifier? Write down the problems that are encountered in Euclidian distance measure.

b) Write down the kd-tree construction algorithm. Hence, show that a kd-tree has a depth in the order of $O(log_2 n)$; where n is the number of records.

7. 3 + 7

a) Write down the working principle of a Support Vector Machine (SVM). How does the SVM absorb the noise components?

b) Applying ***Naïve Bayes*** theorem, determine whether the below test animal is Mammal or Not.

Test animal:

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

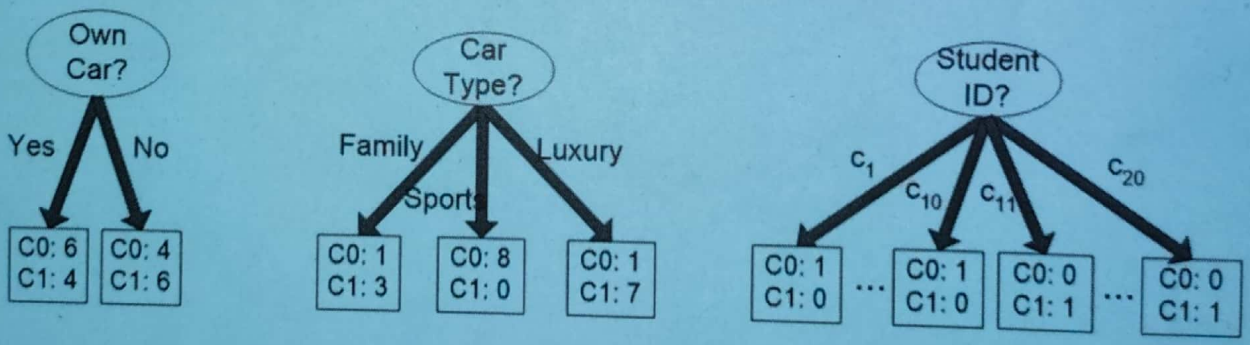
Data Set is given as below:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

8.

3 + 7

a) For the below test cases, which classification is the best? Justify your statement.



b) From the table below, select the first one among four attributes for constructing a decision t
on the basis of information gain of each attributes.

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

Appendix-A: Table for T-distribution

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										