1. (16%) Below is a subset of relations from COMPANY schema. The keys have been underlined.
EMPLOYEE(EMPNAME, **EmpID**, ADDRESS, SALARY, SupervisorEmpID, DNUMBER)
DEPARTMENT(DNAME, **DNUMBER**, MANAGERID)
WORKS_ON(**EmpID, PNUMBER**, HOURS)
PROJECT(PROJNAME, **PNUMBER**, DNUMBER)
DEPENDENT (**EmpID, DEPENDENTNAME**, SEX, BDATE, RELATIONSHIP)

In a company, each employee works for a department and may work on several projects. The WORKS_ON table keeps track the hours that an employee works on each project. The EMPLOYEE table also keeps track of the direct supervisor of each employee. Supervisors are also employees. An employee (supervisor) may be the direct supervisor of several employees. A department controls several projects. A project is controlled by a department. Each employee may have a number of DEPENDENTs.

Express the following Query in SQL statements.

Query: For each employee who is not a manager and has more than two dependents, list the EmpID and name of the employee, the name of his/her supervisor and the number of projects that the employee works on.

2. Suppose that we have an ordered file with r = 80,000 records stored on a disk with block size B = 512 bytes. File records are of fixed size with record length R = 100 bytes.

(a) (4%) How many block accesses would be needed to do a binary search on the data file? How many block accesses in average would be needed to do a linear search for non-ordering field on the data file?

(b) (4%) Assume that we have constructed a primary index for the file that the ordering key field of the file is 10 bytes long and a block pointer that is 10 bytes long. What's the total number of blocks needed for the index file? How many block accesses would be needed to search for a record using the primary index?

(c) (4%) Assume that we have constructed a secondary index on a non-ordering key field of the file that is 10 bytes long and a block pointer that is 10 bytes long. What's the total number of blocks needed for the index file? How many block accesses would be needed to search for a record using the secondary index?

The above index, which is constructed based on a non-ordering key field, is called a first level index. Assume that a second-level index has been constructed based on the first-level index to create a multi-level index.

(d) (5%) What's the total number of index entries for the second-level index. How many block accesses would be needed to search for a record using the multilevel index?

3. **(10%)** Vocabulary (Matching) Questions

Term **Descriptive Phrase**

(a) _____ A software layer that manipulates a database in response to requests from applications.

(b) _____ A data mining technique applied when trying to identify any underlying heterogeneity within housing patterns in a community.

(c) _____ A data mining technique applied when trying to identify common properties between different groups of shoppers.

(d) _____ A description of only the portion of a database available to a particular user.

(e) _____ A data mining technique would be applied when trying to identify traits that characterize the citizens of a democracy who fail to vote

List of possible answers:

A. Class discrimination	B. Class description	C. Sequential pattern analysis
D. Relational operation	E. Outlier analysis	F. Data independence
G. Association analysis	H. Relation	I. Subschema
J. Schema	K. Cluster analysis	L. Data warehouse
M. Distributed Database	N. DBMS	O. Relational database model

4. **(a) (5%)** Why is the straightforward "goto" statement no longer popular in high-level programming languages?

(b) (5%) Draw a flowchart representing the structure expressed by the following for statement.

```
for (int x = 2; x < 8; ++x)
{ . . . }
```

5. **(a) (8%)** Explain briefly the concept of a "class" and the concept of an "object" in an object-oriented programming environment.

(b) (5%) Why is data mining not conducted on "online" database?

6. The experts focusing on policy making and strategic planning management claim that the results of political opinion polls may affect voting behavior dramatically. In a governor election, one candidate argued against the fake news and fake polls. A netizen A claimed that he used historical polls data to model the final results of the election in 2010, 2014, and 2018. He input all these historical data into a deep neural network model and trained for 15,000 epochs. Several data scientists and other netizens present their comments as follows:
B indicated that the number of samples was too small to use deep learning model since it could obtain overfitting results. Some simple models, such as linear regression, were suggested to solve this problem.
C said that since the problem regarding time series analysis, long short-term memory (LSTM) network could be applied to enhance the prediction performance.
D suggested to use generative adversarial network (GAN) techniques to solve the data inadequacy problem.

(a) (6%) Please briefly describe the overfitting problem and propose some methods to prevent it.
(b) (6%) What is the difference between "linear regression" and "deep neural network"? Do you agree with B and explain your answer?
(c) (6%) Please briefly describe generative adversarial network (GAN) and propose your own GAN-based strategy to improve this prediction model.
(d) (4%) What is long short-term memory (LSTM) network?

7. In artificial intelligence research, confusion matrix is a table that is often used to describe the performance of the classification model. Given an example confusion matrix for a binary classifier as below, please define the equation of the following terms by TN, FP, FN, and TP.

(a) (4%) Accuracy
(b) (4%) False Positive Rate
(c) (4%) F_1 score

		Predicted	
		Negative	Positive
Ground Truth	Negative	**TN**	**FP**
	Positive	**FN**	**TP**