MCTA 4362:Machine Learning
Assignment 1

Name:…… Matric No:……

Answer <u>ALL</u> Questions.

Question 1 (20 Marks)

Please refer to the data set attached which was obtained from a census done throughout various locations in California. The dataset contains the latitude and longitude of a particular suburb/region/location, how old the houses in that area are (housing median age), total number of rooms in that location, total number of bedrooms, the population, number of families (households), the median income of the families (median income), the proximity to the ocean (Near Bay, 1 hour away,etc) and the median house value in that area. Your objective is to develop regression models to approximate the median house value in that area (dependant variables) based on the independent variables that are available

A) Based on this data set select which parameters (independent variables) will you feel will be suitable/useful to form your matrix of features. Justify your answer.
B) Perform the necessary data preparation for applying supervised machine learning algorithms, Describe what data preparation was necessary (approximate missing data, removing outliers, One Hot Encoding, scaling, etc)
C) Perform Multiple Linear Regression on this data. Evaluate and describe your model.
D) Perform Polynomial Linear Regression on this data using an order for the polynomial features of your choice. Evaluate and describe your model.
E) Perform Support Vector Regression on this data. Evaluate and describe your model.
F) Based on your results, determine the best model for this data and justify your answer.

Question 2 (20 Marks)

Solar-based energy is becoming one of the most promising sources for producing power for residential, commercial, and industrial applications. Energy production based on solar photovoltaic (PV) systems has gained much attention from researchers and practitioners recently due to its desirable characteristics. However, the main difficulty in solar energy production is the volatility intermittent of photovoltaic system power generation, which is mainly due to weather conditions. For the large-scale solar farms, the power imbalance of the photovoltaic system may cause a significant loss in their economical profit. Accurate forecasting of the power output of PV systems in a short term is of great importance for daily/hourly efficient management of power grid production, delivery, and storage, as well as for decision-making on the energy market, facilitate early participation in energy auction markets and efficient resource planning.

A) Based on this data set select which parameters (independent variables) will you feel will be suitable/useful to form your matrix of features. Justify your answer.
B) Perform the necessary data preparation for applying supervised machine learning algorithms, Describe what data preparation was necessary (approximate missing data, removing outliers, One Hot Encoding, scaling, etc)
C) Perform Multiple Linear Regression on this data. Evaluate and describe your model.
D) Perform Polynomial Linear Regression on this data using an order for the polynomial features of your choice. Evaluate and describe your model.
E) Perform Support Vector Regression on this data. Evaluate and describe your model.
F) Based on your results, determine the best model for this data and justify your answer.