(i) Give two examples of situations when logistic regression would give inaccurate predictions. Explain your reasoning. [5 marks]

If the number of training data points is less than the number of features variables, for example, 50 feature variables/independent variables, but only 10 rows of training data points, which most likely results from overfitting because the model can perfectly fits very few training data, but not generalize enough to general data due to lack of training data for getting the appropriate parameters for these feature variables and make model not generalized.

logistic regression requires the independent variables/feature variables not be too highly correlated with each other(no multicollinearity), otherwise the model can add more weights/duplicate to the similar variables(correlated), this can make the model gives inaccurate prediction. This non correlated requirement applies to the linear regression. As for logistic regression is the log of linear regression, this rule is applied. For example, if the size is an independent variable, the length should not be as an independent variable, as size and length are very related with each other.

(ii) Discuss some advantages and disadvantages of a kNN classifier vs an MLP neural net classifier. Explain your reasoning. [5 marks]

KNN
Advantage

- Non-parmetric, no assumption on the distribution of data.
- Simple, only provides parameter k, and the distance function(Euclidean, Manhattan distance)
- No training step, no need to fit a model first, lazy learning,
- Easy to understand and interpretable.

Disadvantage

slow when dataset is large. Heavy calculations and recalculation when new data comes in.
feature scaling: standardize the features in same scale during to distance calculation.
very sensitive to outliers: distance calculation with outliers can create biased result.

MLP neural net classifier
Advantage

- Good at handling both linear and non-linear problems.
- Very flexible and Strong: can be trained for very complex functions between input and outputs. Learn abstract features at different layers.
- No issue working on large dataset

Disadvantage

- Interpretability: multilayer of neurons is Blackbox. Not know the extent of impact of independent variables on the output dependent variables.
- Computation is intense

iii) In k-fold cross-validation a dataset is resampled multiple times. What is the idea behind this resampling i.e. why does resampling allow us to evaluate the generalisation performance of a machine learning model. Why are k = 5 or k = 10 often suggested as good choices? [10 marks]

The idea behind doing the resampling multiple times in k-fold cross-validation is to increase the training times with varieties of spitted training data and test data to avoid the impact of test and training data noises and model results fluctuation, or not enough training, which result in non-representative model parameters.

Through resampling dataset with k subsets, creating (k-1)/n training data size, and one test data (n/k points), fitting the model k times, then average over these results to smooth out the noise and bias, and improve the model generalization performance.

The common choices for k are k=5 or k=10, as we want to k is large enough to average out the noise improve model generalization with enough training needed, at the same time, increasing k increases the model training times, so don’t want to k too large. The criteria are to choose k value, which makes model doesn’t suffer from high variance and high bias, k=5 or 10 is a good tradeoff for the common scenarios and is recommended.

(iv) Discuss how lagged output values can be used to construct features for time series data. Illustrate with a small example. [5 marks]

Lagged output values are target values/past data points from previous period, and timestamps are attached to these values and forming the time-period-oriented sequence of data as time series features for prediction/forecasting for the future time. Given the lagged data up to k-1 time, for prediction of y(k), the feature vector with lagging values is [y(k-1), y (k-2, y(k-3) ….)). K represents the time scale to choose which past data points (for example, month, or year) to include.