Week 4 Faculty Homework.

1. a. GIGO means that if the input data given to a ML model is poor (may be messey, biased, incomplete) the resulting model will be unreliable and perform poorly. Since ML models learn patterns from the input data, poor-quality data leads to poor-quality models and predictions.
   b. Missing values: algorithms can't handle missing entries so the models may become biased or even fail. Outliers: these extreme values can skew model results. Categorical data: ML models need numeric inputs so categorical variables need to encoded appropriately so that we don't end up misleading the model.
2. Inputation is suitable for numerical or categorical data's Mean, Median, Mode through filling missing values with substitute. It can introduce bias into the model. Deletion may be suitable when only a small amount of entries may be missing wherein we remove the rows or columns containing missing values. However, this can lead to losing valuable information especially when we're working with smaller datasets.
3. Feature scaling is important because some algorithms (like KNN, SVM) are sensitive to the magnitude of feature values. Features with larger ranges can dominate others, leading to biased models. KNN is sensitive whereas Decision Trees are not sensitive.
4. The primary purpose of splitting a dataset is to ensure that the ML model is trained to promote generalisation to the unseen data. Training set is used to fit the model. Validation Set is used to tune hyperparameters and evaluate the model during the development process. Test Set is used only once after the final model is chosen in order to estimate the real-world performance.
5. a. In overfitting, the model performs well on the training data but poorly on unseen data. It might memrise noise and some very specific patterns rather than generalizing.
   b. A separate test set is important because it simulates new, unseen data, which helps us detect whether the model has generalized well or has just memorized the training data.
6. A loss function shows (in a quantitative manner) how far off the model's predictions are from the real values. Training is need to minimise this loss ; thereby improving model accuracy.
7. Feature engineering is basically creating new features that can represent the underlying patterns in the data in a better manner. For example, creating a specific BMI feature by combining the height and weight of an individual could help improve the performance of a health-risk prediction model.
8. A single hold-out validation set may not represent the full variability in the data. Its evaluation might be unreliable if the split was particularly "easy" or "hard" due tos chance.

9. a. K-Fold Cross-Validation reduced dependency on single split as it uses multiple folds – each of which serve as validation once while the rest serve as training. Additionally, the results are averaged for a much more robust estimate.
   b. In 5-fold CV the model is trained 5 times, once for each fold acting as the validation set.

10. External validation uses completely new data (from different time period or location) to test the final model's generalisability as it ensures that the model isn't just accurate when limited to the original dataset.

11. Data Leakage occurs when information from outside the training dataset is unintentionally used in the creation of the model. For example, applying scaling before splitting data would falsely inflate model performance and hence, should be avoided. This can be done by applying preprocessing steps after splitting.

12. The main aim of model deployment is to make the model usable by real world people or systems so that it may make predictions on new, unseen data. Such as through APIs.

13. Using tools like pickle to save a trained model is very important so that the model can later be reloaded for making predictions without having to retrain it. This process is important for sharing and deployment.

14. Batch Predictions: running the model once daily to score all customer data for likelihood to click. May be suitable when real-time updates aren't required. Real-time predictions: Using an API to give instant predictions as users visit the website – may be suitable for live personalisation or ad targeting.

15. The 'Works on My Machine' problem refers to code that runs well locally but breaks elsewhere due to dependency or version differences. Docker solves this by packaging the app, its dependencies, and environment into a portable container, which ensures consistent performance everywhere.