**ASSESMENT – 5**

**\*\* MACHINE LEARING WORK SHEET ANSWERS \*\***

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

**Ans :- .Typically, however a smaller or lower value for the RSS is ideal in any model since its means there’s less variation in the data set. In other words, the lower the sum of squared residuals, the better the regression model is at explaining the data.**

1. How can I improve my clustering perfor What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

**Ans :- TSS (Total Sum of Squares) - The sum of squares total, denoted SST, is the squared differences between the observeddependent variableand *its*mean. You can think of this as the dispersion of the observed variables around the**[**mean**](https://365datascience.com/tutorials/statistics-tutorials/measures-central-tendency/)**–**[**much like the variance in descriptive statistics**](https://365datascience.com/tutorials/statistics-tutorials/coefficient-variation-variance-standard-deviation/). n

**Formula** - ∑ (yi - ȳ)2

i=1

**SSR (Residual Sum of Squares)** **- The second term is the sum of squares due to regression, or SSR. It is the sum of the differences between the predicted value and the mean of the dependent variable. Think of it as a measure that describes how well our line fits the**[**data**](https://365datascience.com/tutorials/statistics-tutorials/numerical-categorical-data/)**.**  n

**Formula** - ∑ (ŷi - ȳ)2

i=1

**ESS (Explained Sum of Squares) - The last term is the sum of squares error, or SSE. The error is the difference between the observed value and the predicted value. We usually want to**[**minimize the error**](https://365datascience.com/tutorials/statistics-tutorials/ols-assumptions/)**. The smaller the error, the better the estimation power of the regression. Finally, I should add that it is also known as RSS or residual sum of squares. Residual as in: remaining or unexplained.**  n

**Formula** - ∑ ei2

i=1

1. What is the need of regularization in machine learning?

**Ans :- Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.**

1. What is Gini–impurity index?

**Ans :- The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favours mostly the larger partitions and are very simple to implement. In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.**

1. Are unregularized decision-trees prone to overfitting? If yes, why?

**Ans :- Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.**

1. What is an ensemble technique in machine learning?

**Ans :- Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning.**

1. What is the difference between Bagging and Boosting techniques?

**Ans :- Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.**

1. What is out-of-bag error in random forests?

**Ans :- The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained.**

1. What is K-fold cross-validation?

**Ans :- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.**

1. What is hyper parameter tuning in machine learning and why it is done?

**Ans :- Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors. Hyperparameter tuning takes advantage of the processing infrastructure of Google Cloud to test different hyperparameter configurations when training your model. It can give you optimized values for hyperparameters, which maximizes your model's predictive accuracy.**

1. What issues can occur if we have a large learning rate in Gradient Descent?

**Ans :- In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.**

1. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

**Ans :- Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary. Non-linear problems can’t be solved with logistic regression because it has a linear decision surface.**

1. Differentiate between Adaboost and Gradient Boosting.

**Ans :- AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.**

1. What is bias-variance trade off in machine learning?

**Ans :- In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.**

1. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

**Ans :- Linear - SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple : The algorithm creates a line or a hyperplane which separates the data into classes.**

**RBF - Radial Basis Functions (RBF) are real-valued functions that use supervised machine learning (ML) to perform as a non-linear classifier. Its value depends on the distance between the input and a certain fixed point.**

**Polynomial** **kernels - In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.**