**Research Assignment 3: Classification - Logistic Regression, Decision Tree, and Random Forest**

This paper explores the topic of classification in the context of supervised machine learning, examples of when data scientists use classification, and the differences between three popular implementations of classification: logistic regression, decision trees, and random forests.

According to Jones et al. (2020, Introduction to Clustering), supervised learning is a technique that detects meaningful information from data already labeled and structured in some manner. One form of supervised machine learning is classification, a modeling technique trained on labeled data to predict a discrete categorical output given some input (Ng, 2012). This discrete output value is predetermined and is the key difference between classification and regression techniques, which aims to produce a numerical value from a range of continuous values (Tatsat et al., 2020). An example of classification modeling is document sentiment analysis, which trains a model on positive, negative, and neutral customer feedback, and then predicts whether a new customer review fits under one of those three categories based upon the content of the review (Liu, 2015). Other examples include training models to identify whether a particular organism is obese given its weight as an input (Starmer, 2018a), or if a tumor is malignant or benign given its size as an input (Ng, 2012).

Due to the popularity of classification modeling, multiple implementations have been developed for particular use cases. The first example is logistic regression, a specialized use case of linear regression that performs classifications by predicting the probability, on a scale of zero to one, that an input belongs to a particular category (Starmer, 2018a). The resulting linear model appears as an “S” shape, with the top and bottoms of the model signifying zero and one. The “S” shape is fit to the training data via a method known as “Maximum likelihood” calculations (Starmer, 2018a). One use case of logistic regression is in the field of fabrication machining. When a batch of parts are being produced, maximizing the life of a tool is an important consideration. A logistic regression algorithm could be trained to predict whether a part would pass or fail inspection given the number of cuts since the last time the bit was replaced. The shop could then set a threshold number of runs before the bit must be replaced, to prevent a failed part.

Another way to implement classification is the Decision Tree model. Decision trees are chains of binary splits that resemble a tree structure, with branching out to “leaves” that represent the category of the data (Tatsat et al., 2020). A simple example is a model that takes the height and weight of a person and then predicts whether they are male or female (Tatsat et al., 2020). Decision tree models are created via a process called “recursive binary splitting*”* which divides the input data in many ways and then qualifies those divisions against a cost function that is appropriate for the model. The result is a tree structure that is split along the best possible divisions to arrive at each categorical conclusion (Tatsat et al., 2020).

Another implementation of classification is the random forest. Random forest is a method that uses a bagging algorithm to train multiple decision trees with slight variations, and then performs classification by aggregating the resulting classifications from all the decision trees (Starmer, 2018b). The random forest algorithm’s utility comes from the way that it trains each decision tree. It does not allow each decision tree to be trained on all variables and their values, and instead only allows a decision tree to be trained on a particular set of variables. The result is that each decision tree has a lower degree of correlation and provides a diversified approach to classification (Tatsat et al., 2020).

In comparison, each of these techniques has an advantage and disadvantage. Logistic regression is probabilistic and provides information about the likelihood that a particular input belongs to a specific classification. Decision trees are more deterministic, gauging the input variables and coming to a determination of the input’s classification. Given the example mentioned above of gauging the classification of a cancer diagnosis given the size of a tumor, the choice of model is critically important. A decision tree model may arrive at a decision to treat the tumor as cancerous and begin treatment. A probabilistic model such as logistic regression may indicate that the likelihood of the tumor being cancerous is only fifty-one percent and, therefore, may require further diagnosis before starting aggressive treatments. The random forest model is known as an ensemble technique, meaning that it combines many models to arrive at its classification. This provides more nuanced and accurate results, relative to the decision tree, but in doing so, the random forest model increases the complexity of the classification and begins to resemble a black box that makes controlling its implementation difficult (Tatsat et al., 2020).

**References**

Jones, A., Kruger, C., & Johnston, B. (2020). *The Unsupervised Learning Workshop*. Packt Publishing. <https://learning.oreilly.com/library/view/the-unsupervised-learning/9781800200708/B15923_02_Final_SMP.xhtml#_idParaDest-49>

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press. <https://www.cambridge.org/core/books/sentiment-analysis/document-sentiment-classification/2066A1E7C1F4251049AD2D43A8504BE6?utm_campaign=shareaholic&utm_medium=copy_link&utm_source=bookmark>

Ng, A. (2012). Supervised Machine Learning: Regression and Classification. Coursera. <https://www.coursera.org/learn/supervised-machine-learning-regression-and-classification>

Starmer, Josh. (2018a). *StatQuest: Logistic Regression* YouTube. https://www.youtube.com/watch?v=yIYKR4sgzI8

Starmer, Josh. (2018b). StatQuest: *Random Forests Part 1 – Building, Using and Evaluating*. YouTube. https://www.youtube.com/watch?v=J4Wdy0Wc\_xQ

Tatsat, H., Puri, S., Lookabaugh, B. (2020). *Machine Learning and Data Science Blueprints for Finance (Chapter 4: Supervised Learning: Models and Concepts)*. Taiwan: O'Reilly Media.