**Chapter 1**

**Introduction**

The following section will give an overview of the research topic chosen. This section will include Research Background, Research Purpose, Hypothesis, and Research Questions, Significance of Research, Importance of Research, Research Boundaries and Research Outline.

**1.1 Research Background**

The main area of study proposed is Machine Learning. Machine learning is a branch of artificial intelligence which is centered on the concept that system can learn from data, gain the ability to identify patterns, thus leading to decisions with minimal human intervention. The iterative position in machine learning is crucial as models are subjected to new data, they adapt in an independent matter (Machine Learning: What it is and why it matters, 2021).

In this research, the research will focus on supervised machine learning. Super- vised machine learning is the pursuit for algorithms that use externally supplied occurrences to reason and create a general hypothesis, which then make expectations for future instances. The goal in supervised learning is to construct a short model of the distribution of group labels in predictor feature conditions. The classifier of result is then used to designate group labels in the testing occurrences where the values of the predictor features are recognized, yet the significance of the group level is unknown (Kotsiantis, 2007).

The researcher intends to focus his research on predicting auto dealer car sales. The research method for this research will be quantitative, as the data will collected from several auto dealers through an online survey and experiment. The experiment will make use of a past dataset from the internet as this will not be given from the dealers themselves. The data collected will then be analyzed via descriptive and inferential statistics techniques.

**1.2 Research Purpose**

The motivations of this research are trifold which contributes to the researcher’s personal motivation for the choice of the research theme: Firstly, the researcher has a great interest in the automotive industry alongside the understanding of market fluctuation. As a car enthusiast, the flexibility in prices of newly developed cars is of interest as it may possibly:

* Be a possibility of a future purchase.
* Benefit the organization and lead to more research in improving automobiles, improved and more sustainable production, and ultimately create an automobile of personal interest in long term or
* Detriment the profitability of the organization leading to less production, more costs incurred, and ultimately failing to create personal interest in their automobiles.

Secondly, machine learning is a technique of data evaluation which presets analytical model constructing. The concept is centered around the system can learn from data, obtain the ability to identify patterns, thus judgments are made with minimal human intervention.

Finally, Machine Learning has made it possible to produce models capable rapidly and automatically of larger and more complex data analysis, faster delivery, and more accurate results. With the construction of accurate models, businesses have an increased chance of identifying profitable chances and avoiding unknown risks (Machine Learning: What it is and why it matters, 2021). The application of machine learning has enhanced many individual studies such as neural networks, genetic learning, cancer prediction and prognosis, drug discovery and development, ecological modelling, animal behavior studies, and so on.

**1.2.1 Hypothesis and Research Questions**

The Hypothesis of this research is: Supervised machine learning, can predict auto dealer car sales.

The research questions in this study are:

RQ1: Does applying supervised machine learning technique increases auto dealer car sales?

RQ2: Can machine learning using a supervised learning method accurately predict auto dealer sales?

This section will present a detailed idea about how the proposed study will be conducted. This section will include Target Participants and Sample size, Data Collection Methods and Data Analysis.

The target participants are vehicle auto dealers. In this case with a population size of 116 auto dealers (Car Dealers in Malta Gozo, 2021), a 10% margin of error, and a 95% confidence, the sample size is of 53 respondents (Sample Size Calculator by Raosoft, Inc., 2021). This target group has been selected as it fit due to its relevance to the study in confirming the information discovered.

Online surveys to gather information from Auto dealers. Topics for the survey would be regarding the prices, market research, if such a system would be ideal for their use, would they suggest it etc. This will take place online. The experiment for this study will consist of understanding the idea on why decisions were made, the demographics behind these decisions, and the outcome. Upon receiving instances of these, the system will learn and adjust to scenarios, gaining capability of prediction in auto dealer sales.

The data analysis will consist of Chi squared, Cross-tabulation, T-test, and Anova. Chi squared, which is commonly used to assess tests of independence when using Cross tabulation. The test of independence assesses whether there is an association between the two variables via witnessing a repetition of reactions expected if the variables were truly objective of one other (Using Chi-Square Statistic in Research - Statistics Solutions, 2021). The T-test being an inferential statistic, it shall be used to determine any substantial difference between two group means, potentially related in certain features (T-Test Definition, 2021). Anova splits observed cumulative variability observed inside a data set into two parts: systematic factors and random factors. Systematic factors have a statistical influence on the data set given, while random factors do not (How Analysis of Variance (ANOVA) Works, 2021).

**1.2.2 Significance of this Research**

The research will be carried out to create a machine learning environment in which car dealership prices and sales in Malta are analyzed and learned. Ultimately this would lead to the machine predicting the future sales of car dealership cars, with the use of supervised regression learning. This will aid risk management and provide advantages and disadvantages of important business decisions, as well as understanding the organization’s current situation and future situations to make the correct business decisions.

Dealerships are prone to interest themselves in such a product due to the advantages in aiding future business decisions. Finally, this study will also compare the advantages and disadvantages of making use of such a system, while also comparing risks and whether it is worth taking them. Furthermore, this research can be used as a reference by other researchers’ conducting similar studies to obtain the applicable and justifiable resource about the use of supervised machine learning or it can be also used as a reading material for any person who is interested in this area of study.

**1.3 Importance of Research**

The contributions made by this study will involve additional insight to the effects of implementations and understanding of future new car markets from auto dealers in Malta. This study will help the understanding of the upcoming new car sale sales and the decision changes auto dealers are prone to with the use of machine learning prediction. Prices are constantly fluctuating, and, in some instances, second-hand cars are more expensive than new cars due to an inflation which has been caused by popularity, rarity, manufacturers concluding production, etc. The Artificial intelligence will learn through a series of examples the patterns while taking into consideration the effects of both independent and dependent variables, estimate an accurate price point which will then reveal which factors are imperative, what changes or precautions can be made, and what data can be altered. Further insight into this topic will avoid future problems, understand if machine learning predictions are worth the effort and price and if this will be profitable.

**1.4 Research Boundaries**

This study will remain within the supervised machine learning environment. A supervised learning model has been chosen for omitting manual classification work and for making predictions based on labeled data in the future. Ideally the use of supervised machine learning is in cases where both input and the corresponding output are known, where labeled data is used in training the machine. The supervised learning method takes direct feedback to help understand whether the correct answer has been predicted. Following the data input this machine learning method is to be supervised to generate a more accurate prediction.

The target market has been chosen due to the researcher’s interest in the auto- motive industry along-with the recognition of market fluctuation. This will aid in interpreting the flexibility in prices of newly developed cars, possibly to be a benefit of an organization, improving research on automobiles, and feasible production.

**1.5 Research Outline**

This research will contain the following chapters:

1. Introduction:

2. Literature Review:

1. Types of Machine Learning
2. Supervised Machine Learning
3. Development Methods of Supervised Machine Learning
4. Benefits of Supervised Machine Learning
5. Predictive Analysis Approach and Current Applications
6. Suggested Development Solution
7. Conclusion

3. Research Methodology:

4. Analysis of Results and Discussion

5. Conclusions and Recommendations

**1.6 Conclusion**

This section will give an overview of the literature that is related to the research topic

chosen. This section will include Machine Learning, Types of Machine Learning, Supervised

Machine Learning, Development Methods of Supervised Machine Learning, Benefits of

Supervised Machine Learning, Predictive Analysis Approach and Current Applications,

Suggested Development Solution.

**Chapter 2**

**Types of Machine Learning**

By feeding data to the machine learns in four possible ways, supervised, unsupervised, semi-supervised, and reinforcement learning. Developing an algorithm takes both input and output in machine learning, while the algorithm is expected in the production of machine learning (Mamgain, Kumar, Nayak and Vipsita, 2018).

Supervised learning algorithms take direct feedback for the prediction. Supervised learning can be categorized in classification and regression methods. K-nearest neighbor (KNN), Decision Tree (DT) using Support Vector Machine (SVM), Logistic Regression, Linear Regression, Naïve Bayes (NB) etc., are popular algorithms of supervised learning (Chauhan, et al., 2021).

Unsupervised learning algorithms do not receive feedback for forecast. This learning finds the hidden patterns in the data. Unsupervised learning methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) are mainly applied for the dimensionality reduction. Two simple concepts are used in unattended machine learning, PCA and Cluster Analysis. PCAs eliminate extremely associated features using covariance matrices, eigenvalues, and eigenvectors. K-Mean, Self-Organizing Model (SOM), PCA and LDA, etc. are well-known unsupervised learning algorithms. The K-means clustering algorithm is implemented as pre-processing and outlier detection steps (Chauhan, et al., 2021).

Semi-supervised classification algorithms may be categorized into two classifications based on their fundamental assumptions. An algorithm is said to fulfill the manifold assumption if it operates the fact that the data lie on a low-dimensional manifold in the input space. Usually, the underlying geometry of the data is captured by constituting the data as a graph, with samples as the vertices, and the pairwise similarities between the samples as edge weights (Mallapragada, et al., 2009).

Reinforcement Learning method is motivated by the human learning process in which a lesson in learnt from the surrounding environment. Knowledge is obtained by interacting with the environment and adjusting behaviors with the feedback presented by the environment. The feedback received consists of information that would be helpful to decisions made and achieving goals set (Wu & Liao, 2019)

**Supervised Machine Learning**

In the process of supervised machine learning via regression, data from the previous annual overall are taken into consideration and considered the same. This includes both dependent and independent factors. In the process of applying Machine Learning to organizations, dependent factors are taken as natural resources, capital, labor, and all kinds of goods, including normative and positive statements, while independent factors are taken as recession, environmental disasters, political issues, market trends etc. (Sharma, Khater and Vashisht, 2021).

The supervised learning requires a former estimate of the result. In this learning method, a trainer assesses the program’s response to a set of inputs. Sets of inputs and outputs must be introduced to the program during the learning period. The program takes an input and produces the subsequent output, which it then contrasts to the correct output. As a conclusion, the program constructs an internal representation of inputs and outputs. The program is trained with a set of input-output pairs that describes a prior known state. The program discovers a function that characterizes input-output relations (Kovács and Terstyánszky, 1999).

As stated by Doreswamy, Gad and Manjunatha (2017), the supervised machine learning method by assessing the prediction error of five methods: linear regression, SVM, random forest, KNN and kernel ridge. Various tests are then executed to assess the prediction error of the five methods. Missing values within the datasets are handled by omitting the entire row which contains a missing value, while also omitting the missing data (Mamgain et al., 2018).

The issue with ignoring data and omitting missing data is the inaccuracy of the result. The assumption of missing data as well as total ignorance of specific entries together will lose precision thus leading to inaccurate results. Inaccurate data leads to inaccurate results, meaning if the 463 stocks out of 500 were predictable, a difference which potentially changes the margin of error and reliability of the study.

**Development Methods of Supervised Machine Learning**

KNN is one of the simple non-parametric methods. The KNN rule rationale is such that: KNNs are found for a query pattern, the most exemplified class is allocated to the query pattern via majority voting, among these nearest neighbors (Gou et al., 2012). Based on the Euclidean equation formula, KNN turns into a favorable option due to the capabilities offered by the method. KNN handles noise efficiently, simple, easy to use, and uses computerized data (Okfalisa et al., 2017).

DT is a flowchart-like structure, each internal node represents a test on an attribute, [i.e.] whether a coin flip comes heads or tails. SVMs are supervised learning patterns with related learning algorithms that examine data for categorization and regression analysis. This learning pattern makes use of the Kernel Trick, which is a method that represents data merely through a set of pairwise resemblance evaluations between the original data interpretations, ultimately to find an ideal boundary between the potential outputs. DT using SVM contains a starting point root which holds all the records of the data set. As tested by Elaidi et al. (2018), a two-class SVM was used to construct a DT, where each SVM produces the two most homogeneous clusters possible, after which the hyperplane that split the formed groups is established. This pattern was repeated until the stopping criterion was reached, resulting in the construction of a binary DT.

Following Nasteski (2017), Linear Regression represents a correlation among a constant scalar dependent variable *Y* and one or more explanatory variables (such as independent variable, input variables etc.) represented as *X* making use of a linear function. Use is made in supervised learning models, implying a model is trained on a set of labeled data and the model is then used to predict labels on unlabeled data. Figure 1 shows the calculated model (red line) make use of training data (blue points) to fit points as accurately as possible.

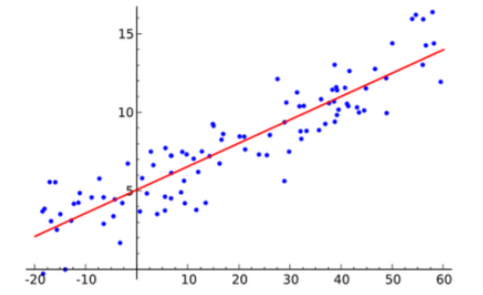


Figure 1 – Visual Representation of linear regression (Nasteski, 2017)

NB is another method of supervised learning as well as the statistical method for classification. NB uses the method of assuming an underlying probabilistic model, allows capturing ambiguity about the model in a principled way by establishing possibilities of the results. Classifications of this type prove practical learning algorithms and can merge monitored data. The NB method is used for the purpose of solving predictive problems (Nasteski, 2017).

Corresponding to NB, Logistic Regression extracts a set of weighted features from an input, logs and combines the result linearly, implying each characteristic is to be multiplied by a weight and subsequently added up. This regression analysis method envisages the prospect of an event occurring by fitting data to a logistic function (Nasteski, 2017). In contrast to NB, logistic regression is a discriminative classifier, it uses characteristics to predict events, takes different weights of data and creates uniqueness is each trait. While NB does not create a unique weight among traits in data, leading it to become a generative classifier.

Each Development method has alternative uses and ideal uses, which are useful in varieties of studies depending on the ideal comparison which is to be made. NB would be ideal is text classification and problems which make use of multiple classes. In contrast, Logistic Regression is ideally used in statistical software, as it will aid the prediction in the possibility of an event occurring. KNN makes highly accurate predictions due to its non-parametric use and would be ideal in data classification of more than two categories. Since SVM uses the kernel trick to transform data, this is ideally applied in studies consisting of multiple classes, both linear and non-linear studies can make use of SVM classification. Linear Regression is fit on a straight line, implying discrepancies between actual and predicted output values are minimized, businesses are capable using Linear Regression to evaluate trends and make forecasts.

An example of developing a supervised machine learning program is a program that will differentiate two different fruits, in this case an apple and an orange. The basis of these two measurements (x₁ and x₂ respectively) are taken from Figure 2. In supervised learning, known labeled examples are used, which are known as a ‘training set’ (the colored datapoints in Figure 2), to gain the capability to discriminate amid the two data classes by discovering a function (Schrider and Kern, 2021).

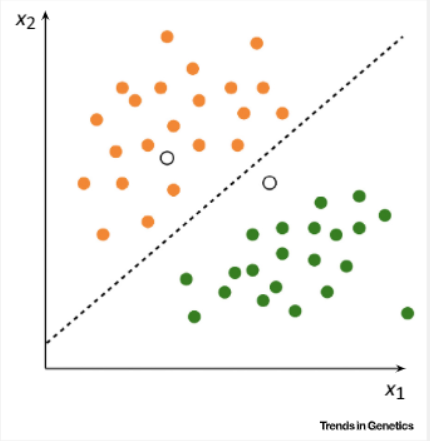


Figure 2 – Training Set (Schrider and Kem, 2021)

**Benefits of Supervised Machine Learning**

Several applications exist for supervised machine learning, depending on the requirements of the program. SVMs execute much better when handling multidimension and uninterrupted features. In contrast, logistic regression systems tend to perform better when dealing with discrete/categorical features. For SVMs, a sizeable sample size is essential to attain its full prediction precision whereas NB may need a rather small dataset. NB does not require a large storage space during both training and classification phases. KNN deals with majority voting, meaning the classification method is sensitive to minor features which make the system very accurate (Osisanwo et al., 2017). Supervised machine learning benefits from such a wide variety of development methods, implying that there is a wide variety of studies which can make use of supervised machine learning, if the relative development method is chosen correctly.

The precision of specific labels allocated by a supervised machine learning model can be justified using the current labels included in the training set. These specific labels are typically allocated by an individual. In this regard, it is a less objective certification standard than the more meticulous dictionary-centered approach of using certain keywords. Simultaneously, a supervised learning method may generate high quality findings at significantly reduced costs, when associated to the costs of emerging comprehensive dictionaries (Collingwood and Wilkerson, 2011).

In addition to accurate prediction, one of the main benefits of using supervised machine learning is the ability to circumvent using ideal, parametric simulations of the data when labeled training data can be attained from practical observation. In cases where practically developed training sets are unavailable, simulations can be used as an alternative to produce training sets. Further, supervised machine learning can be trained to identify occurrences as they are in nature, as opposed to the phenomena shown in the model (Schrider and Kern,2021).

**Predictive Analysis Approach and Current Applications**

The predictive analysis approach is when data, statistical algorithms, and machine learning techniques are used to detect the probability of impending results based on past data. The expected outcome of this approach is to surpass what has happened to provide the best evaluation of what the impending outcome shall be (SAS, 2019).

Real time data facilitates fault prediction, they are considerably more beneficial to the testing and evaluation process. The process has been used for models of sorts, such as limiting data for prediction, historic data to predict faults, fault modeling etc. (Jones and Engler, 2010). Jones and Engler (2010), argued that Real time data analysis facilities are missing the aspect of real time data to produce fault predictions in real-time, their literature has presented a current application of the Predictive Analysis system used in a real-time environment. Their system Predictive Analysis Collaboration Object (PACO) runs an analysis on test data used both in real-time and offline, facilitating optimal prediction accuracy. The creation of PACO was to secure data from one or more test stations, to be used in analysis for fault predictions. PACO is made up of 5 components, namely, PACO listener, PACO server, PACO interference engine, PACO SQL database, and PACO runtime monitor. Through such a design PACO can customize an installation to multiple system configurations.

Other applications of predictive analysis systems presently used are in healthcare in which estimations on expected health conditions or results are made, similarly education and artificial intelligence, where deep learning studies are withheld in attempts to grasp a further understanding and gain more knowledge to apply in studies, arts, hobbies, work, and overall create an easier living with these prediction analysis tools.

**Suggested Development Solution**

The automotive industry has become an essential part of the world economy, it is commonly observed new cars at the dealerships priced lower than second-hand cars. Setting a high-level asking price decreases chances of attracting potential buyers, which deters the buyer from visiting the dealership. In contrast, setting a low-level asking price will accelerate sales at the cost of lowering the profit of the dealership (Jerenz, 2008).

In the era of technology, artificial intelligence is an emerging technology with an essential part in prominent projects of today’s world. Machine learning plays a core part in artificial intelligence as it provides self-learning and self-improvement via the machine itself (Mamgain, Kumar, Nayak and Vipsita, 2018). The focus of the development of this study is on the creation of programs which can be provided data and learn it by itself patterns, behaviors, trends etc. Moreover, such a system once tested and accuracy is confirmed reduces uncertainty while also anticipating change in the market, affecting investors, shareholders and customers (dairu and Shilong, 2021).

**Conclusion**

This section gave an overview of the literature that is related to the research topic chosen. This section included: Machine Learning, Types of Machine Learning, Supervised Machine Learning, Development Methods of Supervised Machine Learning, Benefits of Supervised Machine Learning, Predictive Analysis Approach and Current Applications, Suggested Development Solution.

In this research, the research will focus on supervised machine learning. Supervised machine learning is the pursuit for algorithms that use externally supplied occurrences to reason and create a general hypothesis, which then make expectations for future instances. The classifier of result is then used to designate group labels in the testing occurrences where the values of the predictor features are recognized, yet the significance of the group level is unknown (Kotsiantis, 2007).

**Reference list:**

1. Chauhan, T., Rawat, S., Malik, S. and Singh, P. (2021). *Supervised and Unsupervised Machine Learning based Review on Diabetes Care*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/9442021> [Accessed 30 Sep. 2021].
2. Collingwood, L. and Wilkerson, J. (2011). *Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods*. [online] *ScholarWorks*. Available at: <https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1005&context=jitpc2011> [Accessed 2 Oct. 2021].
3. dairu, X. and Shilong, Z. (2021). *Machine Learning Model for Sales Forecasting by Using XGBoost*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/abstract/document/9342304> [Accessed 20 Apr. 2021].
4. Demiriz, A. (2018). *Used Car Pricing and Beyond: A Survival Analysis Framework*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/8665680> [Accessed 30 Sep. 2021].
5. Doreswamy, Gad, I. and Manjunatha, B.R. (2017). *Performance evaluation of predictive models for missing data imputation in weather data*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/8126025> [Accessed 5 Oct. 2021].
6. Elaidi, H., Elhaddar, Y., Benabbou, Z. and Abbar, H. (2018). *An idea of a clustering algorithm using support vector machines based on binary decision tree*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/8354024> [Accessed 1 Oct. 2021].
7. Gou, J., Yi, Z., Du, L. and Xiong, T. (2012). A Local Mean-Based k-Nearest Centroid Neighbor Classifier. *The Computer Journal*, [online] 55(9), pp.1058–1071. Available at: <https://ieeexplore.ieee.org/document/8140321>.
8. Jerenz, A.. (2008). Revenue Management and Survival Analysis in the Automobile Industry. Revenue Management and Survival Analysis in the Automobile Industry. 1-168.
9. Jones, T. and Engler, J. (2010). *PACO: A predictive analysis system for manufacturing test*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/5613625> [Accessed 2 Oct. 2021].
10. Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, [online] 26(3), pp.159–190. Available at: <http://www.informatica.si/index.php/informatica/article/viewFile/148/140> [Accessed 24 Mar. 2019].
11. Kovács, L. and Terstyánszky, G.Z. (1999). *Diagnosising faults by supervised and unsupervised learning*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/7099999> [Accessed 30 Sep. 2021].
12. Mallapragada, P.K., Jin, R., Jain, A.K. and Liu, Y. (2009). SemiBoost: Boosting for Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11), pp.2000–2014.
13. Mamgain, S., Kumar, S., Nayak, K.M. and Vipsita, S. (2018). *Car Popularity Prediction: A Machine Learning Approach*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/8697832> [Accessed 30 Sep. 2021].
14. Nasteski, V. (2017). *An Overview of the Supervised Machine Learning Methods*. [online] Research Gate. Available at: <https://www.researchgate.net/profile/Vladimir-Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf> [Accessed 1 Oct. 2021].
15. Okfalisa, Gazalba, I., Mustakim and Reza, N.G.I. (2017). Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*.
16. Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT)*, [online] 48(3). Available at: <https://www.researchgate.net/profile/J-E-T-Akinsola/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison/links/596481dd0f7e9b819497e265/Supervised-Machine-Learning-Algorithms-Classification-and-Comparison.pdf> [Accessed 2 Oct. 2021].
17. SAS (2019). *Predictive Analytics: What it is and why it matters*. [online] Sas.com. Available at: <https://www.sas.com/en_us/insights/analytics/predictive-analytics.html>.
18. Schrider, D.R. and Kern, A.D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, 34(4), pp.301–312.
19. Sharma, P., Khater, S. and Vashisht, V. (2021). *Sales Forecast of Manufacturing Companies using Machine Learning navigating the Pandemic like COVID-19*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/9357751>.
20. Wu, W. and Liao, M. (2019). *Reinforcement Fuzzy Tree: A Method extracting Rules from Reinforcement Learning Models*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/8940165> [Accessed 30 Sep. 2021].