**RESEARCH PROPOSAL FORM**

*(also referred to as the ‘Statement of Intent Form’, or SOI)*

***To be submitted by the researcher to the Institute Research Sub-Committee (IRC)***

|  |  |
| --- | --- |
| **Research Title:**  Applying Supervised Machine Learning Technique to Predict Auto Dealer Car Sales | |
| **Institute name**  Institute of Information & Communication Technology | |
| **Course / Programme:**  MCAST Bachelor of Science (Honours) in Business Analytics | |
| **Level and year of study**  Level 6 - Year 3 (2021-2022) | |
| **Main area of study being proposed:**  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. 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 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. | |
| **Name of Researcher:**  Clive Smith | **Researcher’s I.D. Number:**  00801L |
| **Signature of Researcher** | **Date of submission of Form**  2nd July |
| **Name of Tutor (or Recommended Tutor):**  Jean Paul Tabone | |

|  |
| --- |
| **Personal Motivation for the Choice of Research Theme.** |
| 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. |
|  |

|  |
| --- |
| **Outline of Key Literature and Theoretical Framework or Propositions.** |
| 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.   1. **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).   1. **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).   1. **Development Methods of Supervised Machine Learning**   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 1. In supervised learning, known labeled examples are used, which are known as a ‘training set’ (the colored datapoints in Figure 1), to gain the capability to discriminate amid the two data classes by discovering a function (Schrider and Kern, 2021).  Figure 1 – Training Set (Schrider and Kem, 2021)   1. **Benefits of Supervised Machine Learning**   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).   1. **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 (Predictive Analytics: What it is and why it matters, 2021). Current applications of such systems are in healthcare, education, artificial intelligence, where deep learning studies are withheld in attempts to understand more knowledge to apply in the sector.   1. **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).   1. **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. |

|  |
| --- |
| **Significance of the Study.** |
| 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. |

|  |
| --- |
| **Hypotheses and/or Research Question/s** |
| Hypothesis: Supervised machine learning, can predict auto dealer car sales.  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? |

|  |
| --- |
| **Target Participants and Research Methods for Data Collection and Analysis** |
| 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, Crosstabulation, T-test, and Anova. Chi squared, which is commonly used to assess tests of independence when using Crosstabulation. 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). |

|  |
| --- |
| **Anticipated Contributions of the Study.**  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, secondhand 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. |

|  |
| --- |
| **Dissertation Project Plan.** |
| |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **RESEARCH PROPOSAL GANTT CHART** | **2021** | | | | | | **2022** | | | | | | | TASKS | **July** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **June** | | 1. RESEARCH AND FIND DATA |  |  |  |  |  |  |  |  |  |  |  |  | | 2. LITERATURE REVIEW |  |  |  |  |  |  |  |  |  |  |  |  | | 3. RESEARCH MACHINE LEARNING |  |  |  |  |  |  |  |  |  |  |  |  | | 4. RESEARCH AND LEARN PYTHON SCRIPTS |  |  |  |  |  |  |  |  |  |  |  |  | | 5. CREATE ARTIFICIAL INTELLIGENCE |  |  |  |  |  |  |  |  |  |  |  |  | | 6. FEED MACHINE LEARNING DATA |  |  |  |  |  |  |  |  |  |  |  |  | | 7. INTERPRET RESULTS AND TEST |  |  |  |  |  |  |  |  |  |  |  |  | | 8. DRAFT RESEARCH FOR PROJECT |  |  |  |  |  |  |  |  |  |  |  |  | | 9. EVALUATION OF FINDINGS |  |  |  |  |  |  |  |  |  |  |  |  | | 10. DISCUSSION OF FINDINGS |  |  |  |  |  |  |  |  |  |  |  |  | | 11. FINALIZE PROJECT |  |  |  |  |  |  |  |  |  |  |  |  | | 12. SUBMIT PROJECT |  |  |  |  |  |  |  |  |  |  |  |  | |

|  |
| --- |
| **Ethical Considerations.**  **Refer to *guidance points below. You are also additionally required to read MCAST Document 074 ‘Research Ethics Policy and Procedure’ that is available on the College website via link*** [***http://www.mcast.edu.mt/179***](http://www.mcast.edu.mt/179)  *Research shall be conducted in such a manner so as to avoid any psychological and physical harm to humans and animals and financial damage to organizations:*   1. *Only the supervisor and examiners will have access to any data gathered.* 2. *Participants will remain free to withdraw from the study at any time without having to provide any reason. In the case of withdrawal, all the records and information collection will be deleted.* 3. *The participant, who is the sole proprietor of the data provided, is granting that such data would be processed for this study purposes only.* 4. *The data collection process will be a transparent process.* 5. *All transcriptions and/or electronic recordings reflecting the data collected, once exhausted, are to be deleted* 6. *Confidentiality, anonymity and data protection procedures are to be ethically abided by.* 7. *The researcher would provide a soft copy of the study to the participant, if required.* |
| *Enter details here regarding possibility of issues regarding confidential personal data:*  The researcher will ensure that no personal data or confidential data is divulged in this study. The data gathered from the study will be stored on the researcher’s computer and will be accessible only to the researcher.  The researcher will ensure that the participants will be kept anonymous. No participant’s demographics as name and address are necessary for the study therefore personal data will not be asked and stored.  Data will not be kept longer than necessary. The maximum amount of time it will be kept is for one year, and afterwards it will be shredded.  The researcher will emphasize that the data will be used solely for the purpose of this study only, and that the identity of the participants will be kept anonymous. |
| Enter details here regarding possibility of physical harm:  The researcher will ensure that no physical harm will occur to the participants during the study. The researcher will provide the participants with materials that will be used for the study and are safe and have no way in harming the participants physically.  The environment in which the target participants will take part in the study will be safe and will not endanger the participants physically.    Participants are required to make use of a PC to conduct the experiments for the practical tasks. This should not take more than an hour but if more time is required a break might be considered to avoid any physical strain on the students. |
| *Enter details here regarding possibility of moral harm:*  The researcher will ensure that the participants will not suffer from any moral harm. The researcher will provide a clear and understandable explanation to the participant about the conducted study, by using simple words that the participants are knowledgeable of and highlighting their part in the study.  The researcher will ask participants if they have any questions regarding the study and their part in it.  The researcher will make sure to answer all and any questions related to the conducted study.  The researcher will ensure that he asks simple and straight to the point questions to the participants so that they will understand better what is being asked from them.  The researcher will take into consideration the sensitivity and privacy of the participants. What will be gathered from the participants will only be used for the sole purpose of the study and what happens in the session will remain private.  In the case of disabled persons and other social groups, if they will be within the target population, they would still be taking part in the study and they will receive special care.  The participants are provided with the right to withdraw from participating in this study, without the need to provide any reasons to the researcher. |
| *Enter details here regarding possibility of business harm:*  The researcher will ensure that there will be no business harm as a result of the conductive study.  The researcher will provide any businesses who may have a query about business harm with a soft copy of the study.  The researcher will make sure that only relevant data will be gathered. The data will remain private and used only for the sole purpose for feedback purposes and improvement of the solution.  The researcher will make sure to abide by the new legislation (GDPR) regarding business and data protection. |

**List of Key References:**

|  |
| --- |
| 1. Sharma, P., Khater, S. and Vashisht, V., 2021. Sales Forecast of Manufacturing Companies using Machine Learning navigating the Pandemic like Covid-19. *2nd International Conference on Computation, Automation and Knowledge Management*, pp.2-3. 2. Jerenz, A., 2008. *Survival analysis: Estimation of the price-response function*. [online] Revenue Management and Survival Analysis in the Automobile Industry. Available at: <http://dx.doi.org/10.1007/978-3-8349-9840-84> [Accessed 28 June 2021]. 3. Mamgain, S., Kumar, S., Nayak, K. and Vipsita, S., 2018. Car Popularity Prediction: A Machine Learning Approach. *2018 Fourth International Conference on Computing Communication Control and Automation*, p.1. 4. Sas.com. 2021. Machine Learning: What it is and why it matters. [online] Available at: <https://www.sas.com/en\_us/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human%20intervention.> [Accessed 1 July 2021]. 5. Statistics Solutions. 2021. *Using Chi-Square Statistic in Research - Statistics Solutions*. [online] Available at: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/using-chi-square-statistic-in-research/> [Accessed 30 June 2021]. 6. Investopedia. 2021. *T-Test Definition*. [online] Available at: <https://www.investopedia.com/terms/t/t-test.asp#:~:text=A%20t-test%20is%20a,be%20related%20in%20certain%20features.&text=A%20t-test%20looks%20at,to%20determine%20the%20statistical%20significance.> [Accessed 30 June 2021]. 7. Investopedia. 2021. How Analysis of Variance (ANOVA) Works. [online] Available at: <https://www.investopedia.com/terms/a/anova.asp#:~:text=Analysis%20of%20variance%2C%20or%20ANOVA,the%20dependent%20and%20independent%20variables.> [Accessed 30 June 2021]. 8. Kotsiantis, S., 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31, pp.249 - 268. 9. Yellow.com.mt. 2021. Car Dealers in Malta & Gozo. [online] Available at: <https://www.yellow.com.mt/auto-dealers/> [Accessed 1 July 2021]. 10. Raosoft.com. 2021. Sample Size Calculator by Raosoft, Inc.. [online] Available at: <http://www.raosoft.com/samplesize.html> [Accessed 1 July 2021]. 11. Schrider, D. and Kern, A., 2021. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics, 34(4/), pp.303 - 304. 12. Sas.com. 2021. Predictive Analytics: What it is and why it matters. [online] Available at: <https://www.sas.com/en\_us/insights/analytics/predictive-analytics.html> [Accessed 1 July 2021]. |

***This section is to be filled in by the representative of the Institute Research Sub-Committee (IRC) prior to forwarding of this Form to the ‘MCAST Research Ethics Committee’ for final ethics approval:***

|  |  |  |
| --- | --- | --- |
| ***Nature of Ethical Consideration*** | ***Outcome (Tick)*** | ***Comments/Advice*** |
| All ethical issues have been adequately tackled. |  |  |
| Possibility of issues regarding misuse of data or some form of harm. |  |  |

|  |  |
| --- | --- |
| **Details of Representative to the Institute Research Sub-Committee.** | |
| Name | Signature |
| Designation | Date |

****Annex 1: Participant Information Letter**

**Title of Research:**  **Applying Supervised Machine Learning Techniques to Predict Auto Dealer Car Sales**

You are being invited to take part in a research study. Before you decide to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

**What is the purpose of the study?**

This research is being undertaken on machine learning via supervised learning techniques. This study will help build a machine learning artificial intelligence capable of accurately predicting the sales figures of new cars sold by taking previous data and applying this to the algorithm embedded in the program.

**Why have I been chosen?**

You have been chosen because you make part of the population of the industry this project will be based upon.

**Do I have to take part?**

It is up to you to decide whether your take part. If you decide to take part, you will be given this information sheet to keep and be asked to sign a corresponding consent form.

**What will happen to me if I take part?**

You will then be given a questionnaire on the data of the business to be used in the study. Our data will be used to feed a machine learning artificial intelligence that will be able to analyze past scenarios to predict future outcomes. Your image will be used only for the use of the study and will not promote any negative/positive advertisement.

**What are the possible disadvantages and risks of taking part?**

There are no disadvantages or risks foreseen in taking part in the study.

**What are the possible benefits of taking part?**

By taking part you will be contributing to the development of a set of recommendations for the study that might work in favor of auto dealers by accurately predicting sales figures of the upcoming projects.

**What if something goes wrong?**

If you wish to complain or have any concerns about any aspect of the way in which you have been approached or treated during the course of this study, please contact **Clive.Smith.e25590@mcast.edu.mt**

**Will my details be kept confidential?**

All information which is collected about you during the research will be kept strictly confidential so that only the researcher carrying out the research will have access to such information and will not be shared with any other individuals. Participants should note that data/images collected from this project may be retained and published in an anonymized form. By agreeing to participate in this project, you are consenting to the retention and publication of data.

**What will happen to the results of the research study?**

The results will be written up into a dissertation for my final project of my Bachelor of Science (Honours) in Business

Analytics.

**Who is organizing the research?**

The research is conducted as part of a degree in Business Analytics.

**Who may I contact for further information?**

If you would like more information about the research before you decide **Clive.Smith.e25590@mcast.edu.mt**

*Thank you for your interest in this research…*

****Annex 2: Participant (or Guardian) Consent Form**

**Title of Research: Applying Supervised Machine Learning Technique to Predict Auto Dealer Car Sales**

**Name of Researcher: \_Clive Smith\_\_**

# Please initial box

1. I confirm that I have read and understand the Information Letter

for the above study and have had the opportunity to ask questions.

1. I understand that my/my charge’s participation is voluntary and that I/my

charge am/are free to withdraw at any time without giving any reason.

Name of Participant/

Guardian Date Signature

**Diagram

Description automatically generated with medium confidence**

**Clive Smith 28th June 2021**

Researcher Date Signature

*1 for participant; 1 for researcher*