**INITIAL RESEARCH PROPOSAL FORM**

(also referred to as ‘Statement of Intent Form’)

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

|  |  |
| --- | --- |
| **Research Title:**  Using Predictive Data Mining Techniques to Diagnose Conditions and Condition Related Symptoms | |
| **Institute name**  MCAST IICT | |
| **Course / Programme:**  MCAST Bachelors of Science in Software Development | |
| **Level and year of study**  Level 6 – 2016/2017 | |
| **Main area of study being proposed:**  This study aims to find the best data mining tools and techniques in order to make the most accurate predictions based on patient medical data. These predictions can help to diagnose conditions such as diabetes by recognizing patterns in their medical history or their periodic entries using an application which will be used to gather the data. | |
| **Name of Researcher:**  Dorielle Spiteri | **Researcher’s I.D. Number:**  0206997M |
| **Signature of Researcher** | **Date of submission of Form**  29/06/2017 |
| **Name of Tutor (or Recommended Tutor):**  Luke Vella Critien / Diane Desira | |

|  |
| --- |
| **Personal Motivation for the Choice of Research Theme.** |
| The motivation behind this are of research comes from my own experience of being diabetic. Since being diagnosed, a year and a 5 months ago, I have learned a lot about diabetes and how there are different types beyond type 1 and type 2. I have also learned how if a person with diabetes is not careful this condition can be the cause of other conditions since diabetics are more prone to develop intolerances to specific food areas, kidney failures, eye problems and heart diseases amongst others. These are all terrifying things which keep me motivated to track my blood glucose regularly and to watch what I eat. Having said that I quickly discovered that unless you feel dizzy or generally unwell, it is easy to forget to check your blood glucose.  The current system at the hospital, provides the patient with a diary to write down the glucose readings so that at every appointment they can check the readings and determine whether or not the patient needs more, less or a different medication. The initial research idea was to create a database where diabetic patients can log their readings through an applications since the pen and paper system is not as reliable. The application would produce weekly and monthly reports to point out and calculate any anomalies. The idea later progressed into a more generic health system which everyone can use in order to track their health and hopeful help them in recognizing and hopefully predicting any conditions from an early stage. |

|  |
| --- |
| **Outline of Key Literature and Theoretical Framework or Propositions.** |
| As described by Micheline Kamber et Al (2012) in their book called “Data Mining: Concepts and Techniques” data mining (also known as knowledge discovery from data of KDD) is ‘a young and fast growing field’. In today’s world data is being gather daily from every industry and every corner of the world and it is becoming increasingly important to analyse it. Thanks to the computerization of everything and powerful data collection and storage tools data mining and predicting patterns has become easier. By using data mining we can turn large collections of meaningless data into actual knowledge which can help any business or industry with decision making. In this book the authors also describe the how data mining is a step in the process of knowledge discovery. To begin with that from a database is cleaned and transferred into a data warehouse, from there one can make data selections that are relevant to the and then step into data mining from which patterns can be recognised through evaluation and presentation and this eventually turns into knowledge. This knowledge is the key to helping human beings in taking important decisions. Any can of data can be used for data mining as long as the data is meaningful for the study taking place.  Like many fields medicine and healthcare in general is also venturing into data mining following the success it has in fields like marketing and e-business. In their paper “Predictive Data Mining for Medical Diagnosis: an Overview of Heart Disease Prediction”, Dipesh Sharma et Al (2011) describe how the healthcare environment is ‘information rich but knowledge poor’. This means that there is an enormous is not overwhelming amount of data available but there is also a huge lack of effective analysis tools to discover ‘hidden relationships and trends’ within the data. In their paper they prove how Decision Trees outperform and sometimes even Bayesian classification have similar accuracy to decision tree while other predictive methods like KNN, Neural Networks and Classifications based on clustering are not performing well. They go on explaining how they performed analysis on a data set to predict heart disease. The highest level of accuracy was obtained using decision trees with results of 89% and 99.2%.  In another paper called “Data Mining Applications in Healthcare” by Hian Chye Koh and Gerald Tan we see highlighted the importance of data mining in healthcare, “In healthcare, data mining is becoming increasingly popular, if not increasingly essential”. Data mining is a relatively new methodology and technology, coming into prominence only in 1994. It aims to identify valid, and new and potentially useful, and understandable correlations and patterns in data by combing data sets to recognize patterns that are too subtle or complex for humans to detect. Koh and Tan talk about the Cross-Industry Standard Process for Data Mining, or CRISPDM methodology for data mining: business understanding, data understanding and preparation, modelling, evaluation, and employment. Data mining Applications in healthcare have great potential. These applications can range from the evaluation of treatment effectiveness to the management of healthcare to customer relationship management as well as detection of fraud and abuse. More specialized medical data mining, such as predictive medicine and analysis of DNA micro-arrays are other possibilities that were not explore in their paper as it lies outside the scope of this paper. In their paper they also use a data set to predict diabetes using a decision tree as just like in previous papers the results are very good with accuracy percentages as high as 95.59%.  An interesting paper about applications of data mining for the analysis of blood glucose and diabetes mellitus data is called “Data Mining Technologies for Blood Glucose and Diabetes Management” written by Riccardo Bellazzi and Ameen Abu-Hanna (2009). Since diabetes is a lifelong disease, data available for an individual patient can be massive and useful in predictive analysis. Interpreting blood glucose readings is important not only in diabetes monitoring but also when monitoring patients in intensive care units. In their paper they explore blood glucose home monitoring data of diabetes mellitus patients and monitoring data from hospitalized ITU patients. They use data mining techniques to do analysis applied to the blood glucose time series over 60 days of home monitoring for an 11-year-old with type 1 diabetes. The data is divided into five levels, including hypoglycaemias, normal values, and three levels of hyperglycaemia, from moderate to severe. Pattern that emerged are that measurements collected around 8 p.m. are hypoglycaemias. Histograms show that dinner time is critical for hypoglycaemias, whereas higher values are collected at breakfast. |

|  |
| --- |
| **Significance of the Study.** |
| The aim of this study is to gather data in order to predict medical conditions. The goal is to create a system which will be able to recognize patterns in the user’s entries that can be linked to a specific condition or condition related symptoms. To check the correctness and accuracy of the results actual diagnoses are needed as a comparison. Recognizing a conditions at an early stage is important since letting things progress can lead to many complications but in today’s busy world it is easy to take out health for granted.  This research aims to prove that predicting such conditions can be achieved by gathering and analysing the right data. |

|  |
| --- |
| **Hypotheses and/or Research Question/s** |
| * Is it possible to predict medical conditions by analysing the user’s data? * What is the best predictive data mining technique to get the most accurate results? * Which attributes should be used to make predictions? * How much data needs to be gathered before the analysis and predictions process? |

|  |
| --- |
| **Target Participants and Research Methods for Data Collection and Analysis** |
| * Collect or find an existing a dataset related to the area of research. * Analyse and clean the data so that useful information is kept and used for data mining such as;   + Sex, Age, Weight, Height, Blood Pressure, Glycosolated haemoglobin, etc. * Analyse actual diagnoses to determine what was the cause or what where the symptoms the patient was having before being diagnosed * Conduct some exploratory analysis on the data to discover patterns * Create decision trees (or other predictive techniques) to make predictions and compare the results with actual results * Evaluate the results using a confusion matrix |

|  |
| --- |
| **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/MainMenu/Full-TimeCourses/Rules,PoliciesandRegulations.aspx***](http://www.mcast.edu.mt/MainMenu/Full-TimeCourses/Rules,PoliciesandRegulations.aspx)   1. *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* 2. *Only the supervisor and examiners will have access to any data gathered.* 3. *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.* 4. *The participant, who is the sole proprietor of the data provided, is granting that such data would be processed for this study purposes only.* 5. *The data collection process will be a transparent process.* 6. *All transcriptions and/or electronic recordings reflecting the data collected, once exhausted, are to be deleted* 7. *Confidentiality, anonymity and data protection procedures are to be ethically abided by.* 8. *The researcher would provide a soft copy of the study to the participant, if required.* |
| *Details regarding possibility of issues regarding confidential personal data:*  Ideal an existing datasets from internet and or other publicly available sources, would be used. Otherwise I plan to use my own data or fictional data so no other participants will be needed.  In such case no personal data other than age, gender and numbers related to attributes needed for the research would be collected throughout the process. |
| *Details regarding possibility of physical harm:*  For this research, no physical harm should be done as there would be no physical activities since the data will be either fictional or pre-existing. |
| *Details regarding possibility of moral harm:*  The purpose if this research is not to inflict any moral harm on the contrary, if successful it can predict medical conditions and save the users further damage if the condition is left undiscovered. However users must make responsible use of the application and insert reliable data. |
| *Details regarding possibility of business harm:*  The data and results produced should be of no harm to any Business. Many companies are exploring medical data mining as it is a growing field with a lot of potential so if anything this research could help them if not turn into a project they’re interested in. |

|  |
| --- |
| **Anticipated Contributions of the Study.** |
| Medical Data Mining is a growing field that is being targeted by many companies. This study can be of contribution towards making people more aware of their health by not taking it for granted as well as help medical professionals with diagnosing conditions or determining their causes and the symptoms associated. |

|  |
| --- |
| **Dissertation Project Plan.** |
| * Find the best language and platform in order to create the mobile application which will be the data gathering source as well as the ultimate place where reports are produced and displayed to the user. I intend to familiarize myself with the best language I can find (or learn a new language such as ionic) over the summer and at the beginning of the scholastic year. * Continue building on the Literature Review between October and December. * Refine the data base (which I already created for the second year project) * Build the prototype: November to January/Beginning of February * Find the best data set; February * Analyse the data set; Exploratory analysis, Decision Tress, Predictions and Confusion Matrix: February and March * Analyse results and create graphs: March to April * Presentation: Go through the project and make changes as needed changes and write the documentation: April * Conclusion and Abstract: May to June. * Completion; Submit assignment: sometime in June. |

**List of Key References:**

Riccardo Bellazzi and Ameen Abu-Hanna (2009). “Data Mining Technologies for Blood Glucose and Diabetes Management”. *Journal of Diabetes Scienceand Technology Volume 3. Issue 3*, 603–612.

Micheline Kamber, Jiawei Han Jian Pei (2012). Data Mining: Concepts and Techniques. *Third Edition. Morgan Kaufmann* ISBN: 9780123814791

Hian Chye Koh and Gerald Tan. “Data Mining Applications in Healthcare”. *Journal of Healthcare Information Management Volume 19 No 2 (),* 64–72.

Dr John Schorling. Diabetes data. Retrieved from

<http://192.38.117.59/_tag/Teaching/share/data/Diabetes.html>.

Dipesh Sharma, Jyoti Soni Ujma Ansari and Sunita Soni (2011). “Predictive Data Mining for Medical Diagnosis an Overview of Heart Disease Prediction”. *International Journal of Computer Applications Volume 17 No 8*, 43–48.

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

|  |  |  |
| --- | --- | --- |
| ***Nature of ethical consideration*** | ***Outcome (\*)*** | ***Comments*** |
| *Consideration of possibility of issues regarding confidential personal data:* |  |  |
| *Consideration of possibility of physical harm* |  |  |
| *Consideration of possibility of moral harm* |  |  |
| *Consideration of possibility of business harm* |  |  |

***(\*) Legend to record outcome by Institute Research Sub Committee:***

***A***  *– Ethical considerations have been* ***addressed appropriately*** *by Researcher;*

***B*** *– No (****Nil****) relevant ethical considerations are applicable under purpose of study as described by Researcher.*

***C*** *– Ethical consideration have* ***not been addressed appropriately*** *by Researcher;*

***D*** *– Applicable ethical consideration have* ***not been considered*** *by Researcher.*

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