iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31576

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

Human Life Span Prediction using Machine Learning

Ayshwaryaa N[1], Kavipriya R[2], Sobika K[3], Prof. T. Karthikeyan[4]
[1, 2, 3]*UG Scholar, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, Tamil nadu, India.*
[4] *Asst Prof, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, Tamil nadu, India.*

Abstract: Human an incredible creation of god. Every creature in the world has a limited life span, to achieve something in the world. Likewise, we also have a limited life span to survive in the world. To preserve our self from the consequences, even though lot of inventions has been made by human, to prevent from diseases is a major question mark. Life span prediction has a greater impact in our modern society because of our food habits, different types of diseases and environmental conditions. Investigations about the life span of vertebrates have been made, except the human (HOMO SAPIENS). Predicting life span for human being is a vital step. It is an emerging research area that is gaining interest but involved lot of challenges due to the limited amount of resources (i.e., datasets) available. In our proposed system we present an analysis on human to predict the life span. By obtaining the Date of birth, Environmental factors, Food habits, Diseases and Medical history, a lot of investigations will be conducted to predict the sustainability of human. Using Data Analytics and machine learning algorithms, We can analyze and predict the life span of the individuals and we use different classification algorithms for this prediction to achieve higher accuracy.

Keywords: Disease, Machine learning, Classification algorithm, Homo sapiens and Life span

I. INTRODUCTION

Human being has a major role in existing and upcoming technologies to improve the economical growth of the country and they are being in different specialisations and work together for the welfare of the individuals as well as organisations. Health of individuals is as much essential for his/herself and their people who are around as. Now a day, a lot of new diseases has been identified and it affects the people and further they can be a spreader of diseases even without their knowledge by means of direct or indirect contact with the other persons. For Instance, COVID 19 has affected lot of people in different ways and many people lost their lives. Knowing our lifespan is very essential to take care of our self and people around as. Lot of experiment has been made to find the life span of the different species except Homo sapiens. Due to lack of datasets, research on human lifespan is very less. We can identify the lifespan of individuals using certain factors such as environment, food habits, medical records, different stages and types of diseases so on.

II. RELATED WORKS

Jacob B. Hjelmborg had studied the genetic influence on human life span and how it varies with age using the almost extinct cohorts of Danish, Finnish and Swedish twins born. Females and males have similar rates and these are negligible before age 60 for both MZ and DZ pairs. The patterns for females and males are very similar, but with a shift of the female pattern with age that corresponds to the better female survival. We find that genetic influences on lifespan are minimal prior to age 60 but increase thereafter. These findings provide a support for the search for genes affecting longevity in humans, especially at advanced ages. The concept of age-specific genetic influence on the human lifespan is somewhat difficult to formalize, when compared to the question of age-specific genetic influences on a truly age-dependent phenotype, e.g. body mass index (BMI). The two approaches have been made. The first is based on studying the life expectancy of MZ and DZ twins while conditioning on the co-twin lifespan in order to narrow the focus to certain age groups. In the second approach, lifespan variable was used to define an age dependent dichotomous phenotype and analyze it using the traditional methods of analysis for binary traits to investigate the presence of genetic influences on human lifespan [1]. Regression Approach was used to assess the MZ and DZ twin similarity for a lifespan by selection on the independent variable if the overall regression is linear. The mean lifespan of twins increases with co-twin lifespan and this trend occurs more rapidly for MZ than for DZ twins when the co-twin exceeds approximately age 60 for both males and females.

Benjamin Mayne had studied that Ageing is associated with epigenetic changes involving DNA methylation. Furthermore, an analysis of mammals showed that the density of CpG sites in gene promoters, which are targets for DNA methylation, is correlated with lifespan. Using 252 whole genomes and databases of animal age and promoter sequences, we show a pattern across vertebrates and also derive a predictive lifespan clock based on CpG density in a selected set of promoters. The lifespan clock accurately predicts maximum lifespan in vertebrates from the density of CpG sites within only 42 selected promoters [2].

Our lifespan clock provides a wholly new method for accurately estimating lifespan using genome sequences alone and enables estimation of this parameter for both poorly understood and extinct species. The lifespan clock performed well across species from all classes, producing a median absolute error (MAE) of 3.72 years and a maximum relative error of 5.9% [2]. We also found no significant difference between the absolute error rate between the training and testing data sets. Principle component analysis (PCA) was used to visually characterize the variation of CpG density in the different species.

In the research paper, James Jin Kang uses the health related data such as sleep monitoring, heart rate measuring, and calorie expenditure collected and processed by the devices and servers in the cloud. These requirements can be extended to provide a personalized life expectancy (PLE) for the purpose of well being and encouraging lifestyle improvement. It is based on the concurrent models and methodologies to calculate and predict life expectancy (LE) and proposes an idea of using multi-phased approaches [3]. They used the demographics of selected regional areas and multiple behavioural health disorders across regions to find correlations between individual behaviour indicators and behavioural health outcomes. Smart environment and wireless network technologies have also been used to improve the monitoring of chronic diseases with the evolution in the Internet of Things (IoT) and cloud computing by building smart cities and homes, which allowed the rapidly growing elderly population to access health care resources in a cost-effective way. depicts the life expectancy predicted by an inference system, which transmits health data over wireless sensor networks. As a result, the current prediction of LE, which was found to be up to a maximum of five years, could potentially be extended to a lifetime prediction by utilizing generic health data.

Death is an inevitable part of life and while it cannot be delayed indefinitely it is possible to predict with some certainty when the health of a person is going to deteriorate. Greg McKelvey predicted risk of mortality for patients from two large hospital systems in the Pacific Northwest. Using medical claims and electronic medical records (EMR) data we greatly improve prediction for risk of mortality and explore machine learning models with explanations for end of life predictions. The insights that are derived from the predictions can then be used to improve the quality of patient care towards the end of life. The machine learning model is deployed as a single layer binary classification layer that can be accessed by a cloud based app from any browser. This system uses data from the hospital systems' claims live feed or Electronic Health Record (EHR) data feed. New data can be continuously pulled into the cloud which is then transformed into a standardized schema. A confusion matrix for health care organization prediction can be constructed [4]. The output from the model is a scaled risk score between zero and one. We use a threshold function such that if the score is above the threshold then it is flagged as prediction for end of life otherwise it is flagged as surviving.

III. PROBLEM IDENTIFICATION

There are certain approaches which are used for predicting life expectancy by using a very few attributes like Heart beat monitoring, Sleep monitoring and cholesterol identification and it does not guarantee in predicting the accurate life expectancy because we need certain more parameters such as diseases they have and so on are necessary for predicting the lifespan of the individuals. Furthermore, by observing the human body periodically using different types of sensors which are embedded in the human body does not provide accurate results and provide inconvenience to the person. The hindrance is found in the data processing which involves a complex chain to take place like collecting the data from sensor, transferring the data it to the mobile or web application, storing in the database then converting it into the valid data set and finally using certain models to predict the accurate results. In certain cases, predicting the lifespan of human being is not taken into account because of the lot of variants found in the human life. Many complex techniques methods have been used so that it includes attaching the IOT devices and sensors in the human body and which may harm the body due to the emission of radiation, disturbs the pattern of sleep cycle and so on.

IV. PROPOSED MODEL

In our proposed system it helps to predict the life span of individuals using the certain factors such as data of birth, gender, food habits, environmental condition, average life expectancy of the country, the type and stages of diseases the individual person have, their alcoholic and smoking behaviours and so on are used for this prediction by training the certain classification models to provide the accurate results. Thus the earlier prediction may be helpful to the doctors and also patients to get cured and medicated. The processing of the data is carried out by four modules namely,

1) Data Cleaning
2) Exploratory Data Analysis
3) Feature Extraction
4) Correlation
5) Predicting with ML algorithms

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

The detailed architecture Fig.1 described the processing of the data which is fetched from the patients report.

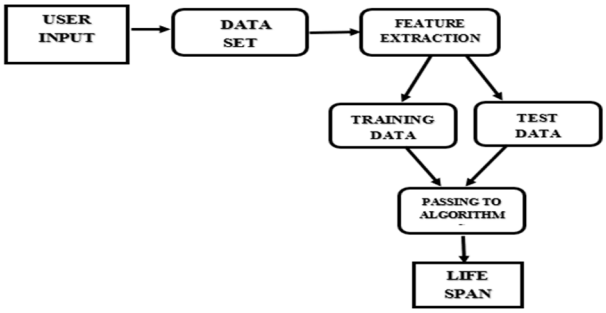


Fig. 1 Proposed Architecture of predicting human life

A. Data Cleaning

Data cleaning plays a significant role while processing a large number of datasets from the health survey. When the data is used with the invalid or null data the generating of the final results becomes inexact thus all the irrelevant, inaccurate data is removed. Data cleaning may be performed interactively with data wrangling tools, and as batch processing through scripting. The data sets are cleansed to get high quality of data from the available data sets. The pseudo code of the missing data is given as below,

```
df.isnull().sum()
df.describe()
df.info()
```

Pseudo code for irrelevant data removal

1) Data Encoding: Encoding Data also plays a vital role i.e., the Fig 2, with the help of this we can achieve data into an equivalent numerical format. Label Encoder is used to perform the encoding task; the label encoder class from the sklearn library will be helpful to transform the data into a new encoded data.

```
GENE DESCRIPTION    0
ACCESSION NO        0
DOB                 0
COUNTRY             0
AVC                 0
GENDER              0
MOBILE NUMBER       0
DISEASES            0
STAGES              0
START AGE           0
DISLIFEEXP          0
INHERITED           0
AN                  0
SN                  0
SMAGE               0
AVSMEXP             0
HEALTHY LIFE        0
AGE                 1
dtype: int64
```

Fig. 2 Data Encoding

B. Exploratory Data Analysis

It is one of the approaches to analyzing data sets to outline the important characteristics of data. It is mainly promoted to encourage statisticians to examine the data that could lead to data collection and experiments. Support the selection of appropriate statistical tools and techniques. EDA is performed using two methods based on visualization i.e., univariate, bivariate and multivariate. Helps to ensuring the best outcomes for the project. Such level of reliability can be achieved only after data is validated and ensuring the dataset was collected without errors. It allows to starting leveraging data science and some technologies.

1) *Seaborn:* The Seaborn is Python data matplotlib based data visualization. . It is data set-oriented plotting function which operates on data frames and arrays

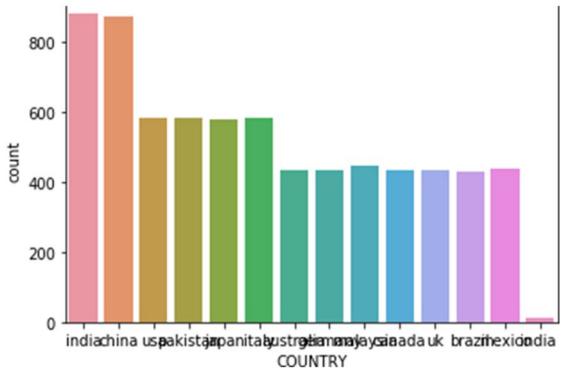


Fig. 3 Seaborn

C. Feature Extraction

Feature extraction is proposed to increases the accuracy of learned models by extracting features from the input data. This framework reduces the dimension of data by removing the redundant data and increase training and inference speed. Detecting the interesting key points and areas in the medical image by a bag of technique is more important in feature extraction phase. It is very important to concentrate on the methods that work efficiently with multilabel datasets. Feature extraction mechanism computes numeric or symbolic information from the data, which are referred to as the features. It is to compress the data set into a lower dimensional data vectors so that classification can be achieved.

D. Correlation

Correlation is variables within a dataset can be related for lots of reasons. The statistical relationship between two variables is referred to as their correlation. Correlated variables should be removed to improve the skills of the model. It may also interest in the correlation between input variables with the output variable in order provide insight into which variables may or may not be relevant as input for developing a model. percentage of correct predictions

E. Predicting Using Ml Algorithms

Machine learning algorithm is a way of identifying patterns in data and it is used to automatically make predictions or decisions. The two main methods of machine learning are regression and classification. In this we used some of the algorithms under regression and classification. Predicting of lifespan can be completed by using these algorithms.

1) *Linear Regression:* Linear Regression is a machine learning algorithm based on Supervised Learning. It is an attractive model because the representation is so simple and it is useful for finding linear relationship between target and one or more predictors. Linear Regression algorithm performs the task to predict a dependent variable value based on a given independent variable. This method is mostly used for forecasting and finding out cause and effect relationship between variables. This algorithm gives efficiency of 46% .
2) *Logisitc Regression:* Logistic regression is a technique used in machine learning from the field of statistics[6]. This method is used for binary classification problems. This algorithm is based on predictive analysis which is used to describe data and also it explains the relationship between one dependent binary variable and one or more nominal or ratio-level independent variables. This algorithm gives efficiency of 4.5% .
3) *Support Vector Machine:* Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression[7]. But generally, they are used in classification problems. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyper plane (MMH)[2]. To conduct the first support

vector machine (SVM)-based study comparing the diagnostic accuracy of T1-weighted magnetic resonance imaging (T1-MRI). Support vector classifier gives an accuracy of 4.5%.

4) *Guassian Naive Bayes:* Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class[8]. A frequency table for each attribute is created and the likelihood of each feature is calculated. Based on the likelihood, the conditional probabilities for each class is determined, and the class with the maximum conditional probability is considered. The Gaussian Naive Bayes gives an accuracy of 12.20%.

5) *Decision Tree:* Decision tree is a type of supervised machine learning where the data is continuously split according to a certain parameter[9]. The tree can be explained by two entities, namely decision nodes and leaves[3]. The leaves are the decisions or the final outcomes. Decision tree classifier gives an accuracy of 87.4%.

F. Predicting With Random Forest

Random Forest is a supervised learning algorithm. Like you can already see from it is name, it creates a forest and makes it somehow random. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. A lot of benefits to using Random Forest Algorithm, but one of the main advantages are that it reduces the risk of overfitting and the required training time. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing. I will talk about random forest in classification, since classification is sometimes considered the building block of machine learning. This algorithm gives efficiency of 90%.

V. RESULT AND DISCUSSION

The implementation of the proposed solution begins with installation of anaconda software. This process is followed by launching Jupyter notebook which helps to import the certain necessary packages i.e., pandas, numpy, sklearn etc. After importing all the packages, various machine learning is implemented for identifying an algorithm with high accuracy. The algorithm which is found to be more accurate is embedded with UI backend for database connectivity.

In this proposed system, we analyzed the lifespan among human beings based on some of the health and environmental factors. By gathering data through survey among people, we studied the correlation between people diseases and their environmental factors. In this work, we also analyze the life expectation of individual people. The result can be shown using the interface and it provides lifespan expectancy for each human being by analyzes the given data.

In our proposed system we have obtained a better accuracy with the help of random forest algorithms through which better result will be obtained comparatively with other algorithms.

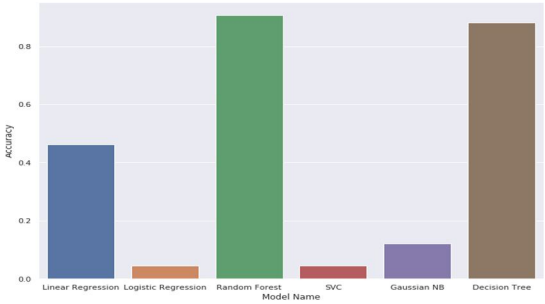


Fig. 4 Comparative Analysis of Accuracy

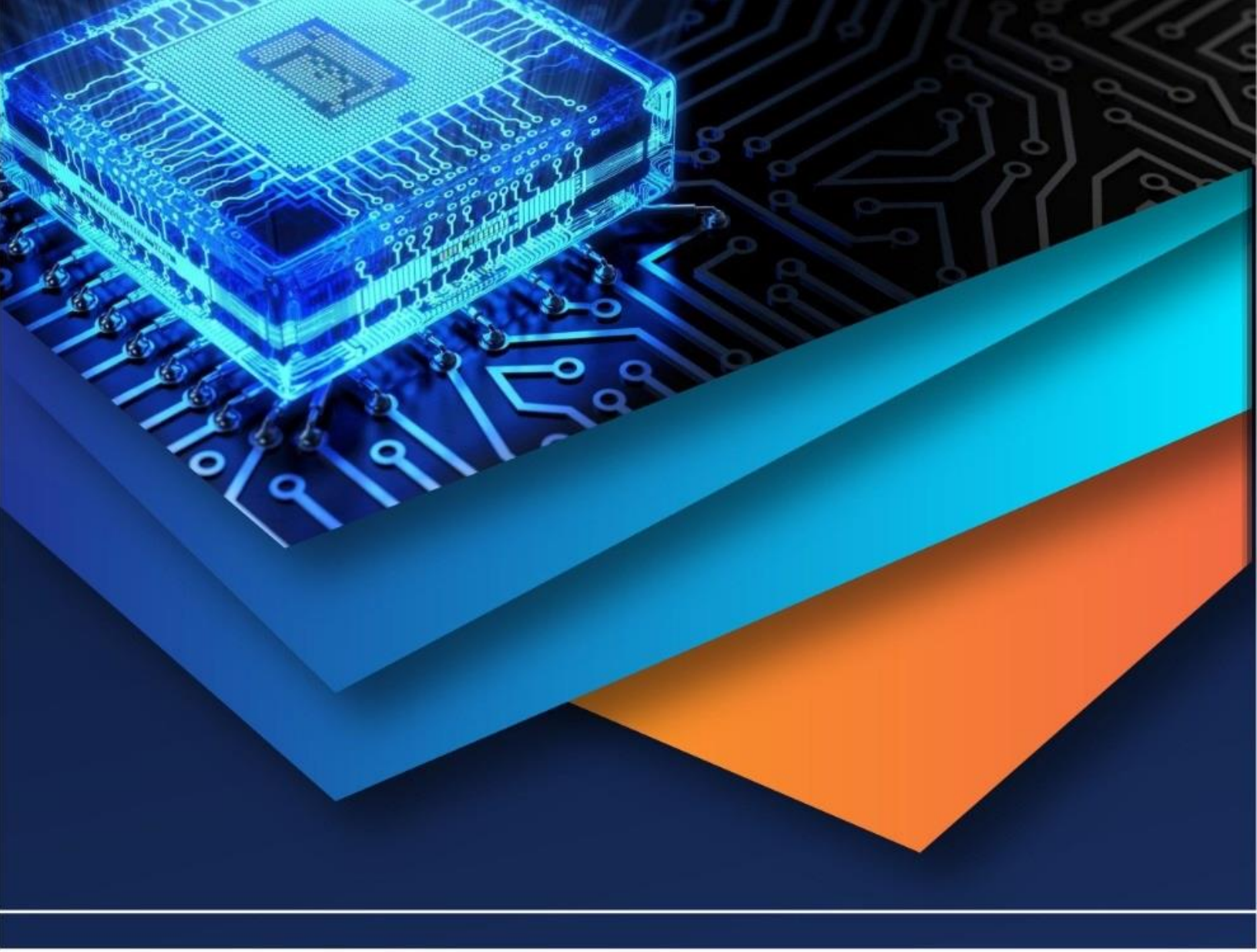
VI. CONCLUSION AND FUTURE ENHANCEMEN

To conclude, the analysis of the human lifespan can be predicted earlier. By employing data through datasets, the correlation between attributes like diseases, gender, ages and environmental factor are monitored. Through this paper, the Random forest algorithm is discovered in order to predict the human lifespan with more accuracy. The advantage of Random forest algorithm, gives more flexibility without obtaining the processed data and accurate. Furthermore, the future enhancement can be made by using deep learning algorithm which may give better solution and also use DNA to get genetic information to provide more accurate analyzes.

REFERENCES

[1] Jacob B. Hjelmborg, James W. Vaupel, Nancy L. Pederson, Mac McGue, Karre Christensen, Marku Koskenvuo, Axel Skytthe [2019] "Genetic influence on human lifespan and longevity".
[2] Oliver Berry, Campbell Davies, Jessica Farley, Simon Jarman [2019] "A Genomic Predictor of lifespan in vertebrates Benjamin Mayne".
[3] Linda Mary, John Ashima Sharma, Siddhant Gujarathi [2019],"Detector and Predictor System for lifeaspan using Naives Bayes and Decision Tree Algorithm".
[4] Barros, Rodrigo C, Basgalupp, Carvalho, Freitas [2012], A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42.
[5] The Prediction of heart disease using Naïve bayes classifier International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 03 | Mar 2019
[6] Developing Machine Learning Algorithms for the Prediction of Early Death in Elderly Cancer Patients: Usability Study. JMIR Cancer. 2019 Jul-Dec; 5(2): e12163. Published online 2019 Sep 26. doi: 10.2196/12163
[7] S.R.Bhagya Shree, H.S.Seshadri, [2019]," Prediction using Naïve Bayesian Classifier", NeuralComput & ApplicDOI 10.1007/s00521-061-2416-3.
[8] James Jin Kang, Sasan Adibi (2019) "Systematic Predictive Analysis of personalized life expectancy using smart devices".
[9] A. Ahmad, Carly Eckert, Greg McKelvey, Kiyana Zolfagar, Ankur Teredesai (2018) "Death versus Data Science: Predicting end of Muhammad".

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY