PROJECT REPORT

“PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING AND PYTHON” USING IBM TOOLS

MADE BY-NILANJANA HABISYASI

ACADEMIC ROLL NUMBER-65111502817

BHARATI VIDYAPEETH’S COLLEGE OF ENGINEERING

GGSIPU,NEW DELHI

ACKNOLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank Ms.Swathi , for providing me an opportunity to do the project work in Smartbridge and giving us all support and guidance which made me complete the project duly. I am extremely thankful to her for providing such a nice support and guidance, although he had busy schedule managing the corporate affairs.

I owe my deep gratitude to our project guide Charan sir & Prashanth sir, who took keen interest on our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system.

I would not forget to remember Charan sir & Prashanth sir,for their encouragement and more over for their timely support and guidance till the completion of our project work.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of [Smartinternz] which helped us in successfully completing our project work. Also, I would like to extend our sincere esteems to all staff in laboratory for their timely support.

Nilanjana Habisyasi

INTRODUCTION

OVERVIEW-Our project was based on pure regression and machine learning using IBM tools such as IBM Watson,python,IBM cloud,IBM machine learning ,auto ai & Node red (To create the userinterface) to create a model to predict life expectancy of humans using a given data set.

PURPOSE/PROJECT DESCRIPTION-

A typical Regression Machine Learning project leverages historical data to predict insights into the future. This problem statement is aimed at predicting Life Expectancy rate of a country given various features.

Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.

LITERATURE SURVEY

Life expectancy is one of the most important factors in end-of-life decision making. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning improves the quality of the final phase of life by stimulating doctors to explore the preferences for end-of-life care with their patients, and people close to the patients. Physicians, however, tend to overestimate life expectancy, and miss the window of opportunity to initiate Advance Care Planning. This research tests the potential of using machine learning and natural language processing techniques for predicting life expectancy from electronic medical records.

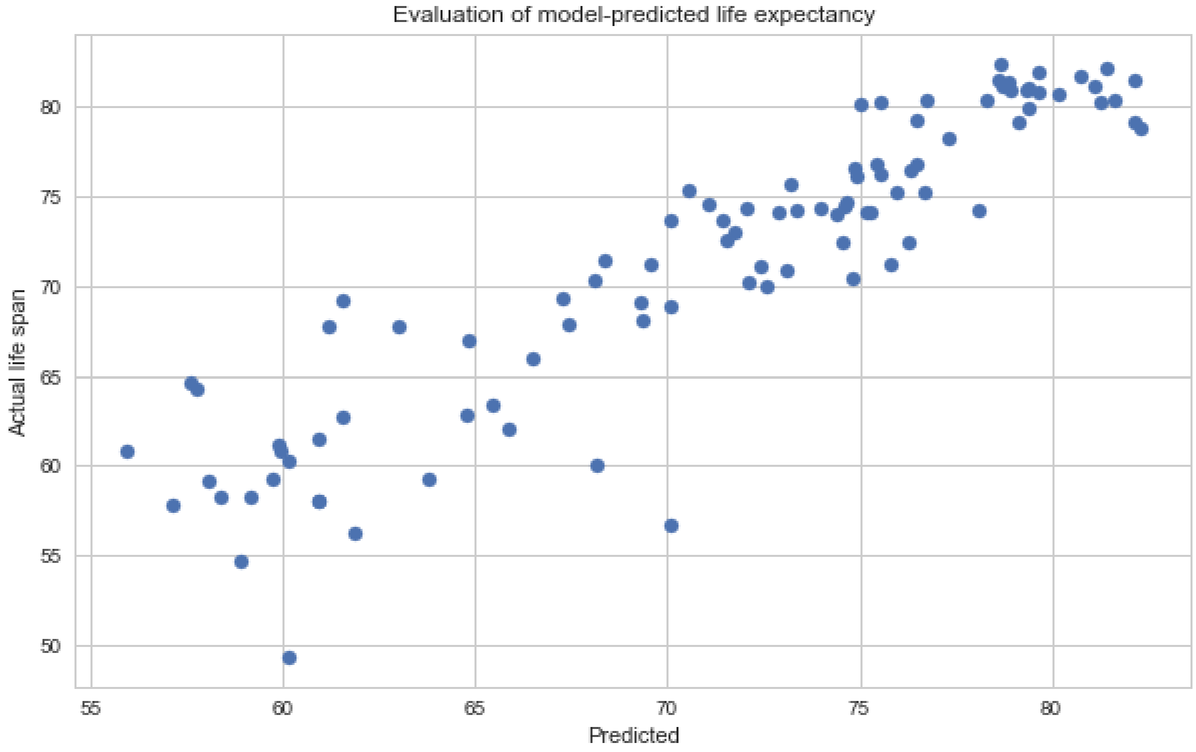
We approached the task of predicting life expectancy as a supervised machine learning task. We trained and tested a long short-term memory recurrent neural network on the medical records of deceased patients. We developed the model with a ten-fold cross-validation procedure, and evaluated its performance on a held-out set of test data. We compared the performance of a model which does not use text features (baseline model) to the performance of a model which uses features extracted from the free texts of the medical records (keyword model), and to doctors' performance on a similar task as described in scientific literature.

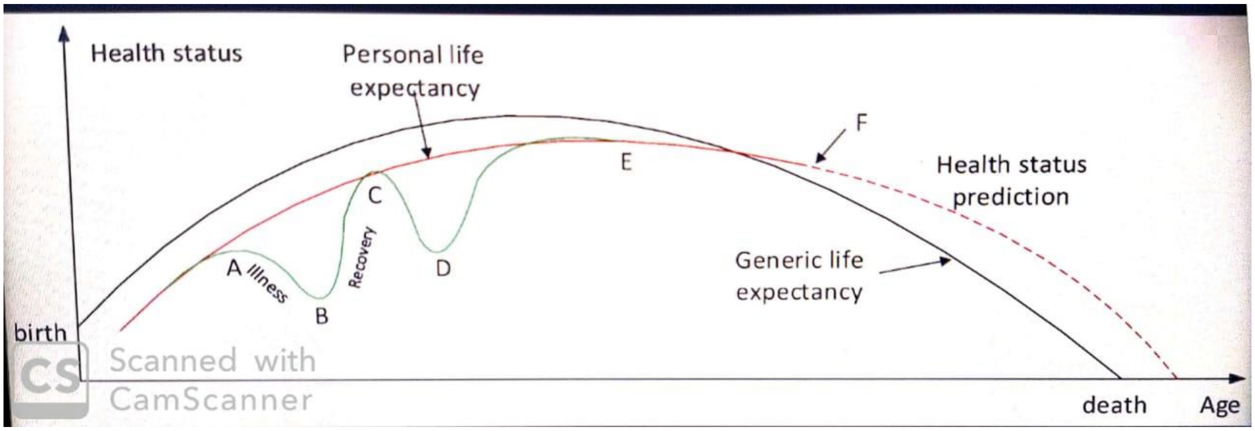
Prognostication of life expectancy is difficult for humans. Our research shows that machine learning and natural language processing techniques offer a feasible and promising approach to predicting life expectancy. The research has potential for real-life applications, such as supporting timely recognition of the right moment to start Advance Care Planning.

THEORITICAL ANALYSIS

The problem of processing datasets such as electronic medical records (EMR), and their integration with genomics, environmental factors, socioeconomic factors and patient behavior variations have posed a problem for researchers in the health industry. Due to the evolution of data science technologies such as big data virtualization and analytics, data wrangling and with the cloud, health workers now have an improved way of processing and developing meaningful information from huge datasets that have been accumulated over many years. For example, big data and machine learning techniques can benefit public health researchers with analyzing thousands of variables to obtain data regarding life expectancy and anxiety disorders. They used the demographics of selected regional areas and multiple behavioral health disorders across regions to find correlations between individual behavior indicators and behavioral health outcomes. Smart environment and wireless network technologies have also been used to improve the monitoring of chronic diseases with the evolutions in the Internet of Things (IoT) and cloud computing by building smart cities and homes, which allowed the rapidly growing elderly population to access healthcare resources in a cost-effective way.

It may be possible to create a prediction of personal life expectancy, which can be further used to calculate health indexes on a generic level for which the individualized expectancies may be compared against.

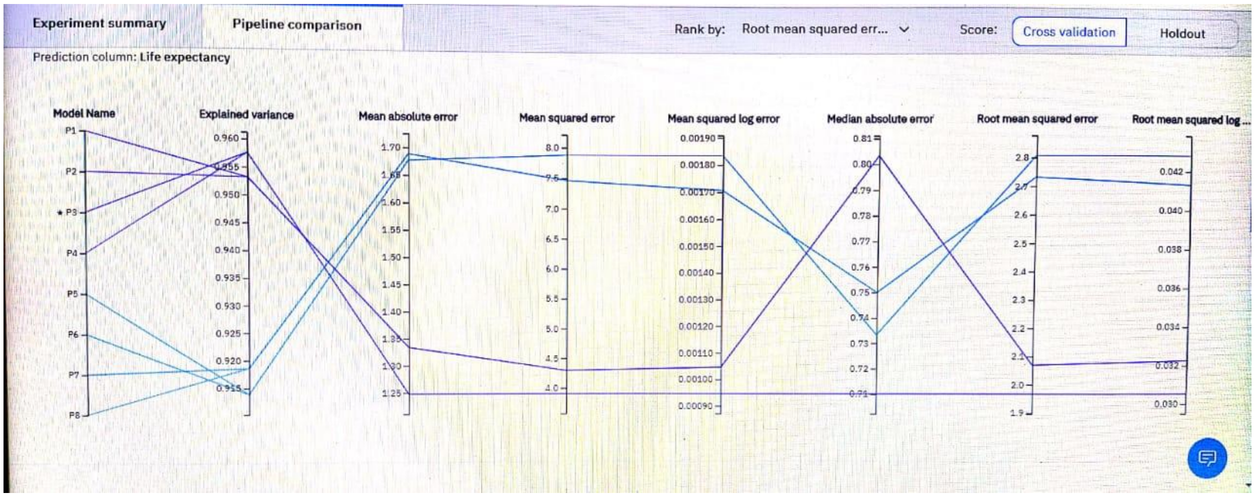




This model can also depict the life expectancy predicted by an inference system, which transmits
health data over wireless sensor networks.
Generic life expectancies may be calculated from
multiple sources obtained and analyzed by big data. These values can be used to create a personalized
graph that most resembles the individual in question, with consideration for personal characteristics
such as their age, gender, ethnicity, living environment and current comorbidities or lifestyle habits.
There may be an enormous number of variables to consider, of which increasing the number may
obviously increase the accuracy of a LE prediction. This personalized graph can then be compared with
other individuals who may be living similar lifestyles and share similar traits to provide an idea of the
generic life expectancy. This concept is described in Figure with the red and black lines superimposed
on each other. During point A, the graph shows that the user was ill and there was a decrease in
the overall health status until point C when the user recovered, followed by another case of illness
at point C until they recovered finally at point E. Point F describes the present moment in time, at
which point any values following this point would be an inferred prediction of the individual’s health
status for the future. This prediction would be an inference made of physiological data analyzed by

the cloud computation. Whilst the red line shows a trend line of the history of the user’s health status
and a rough estimate of his or her future health, the graph may also show fine trends as depicted by
the green line. This would be used to inform short term information about whether the user’s health is
improving or declining on a more detailed level.

EXPERIMENTAL INVESTIGATIONS

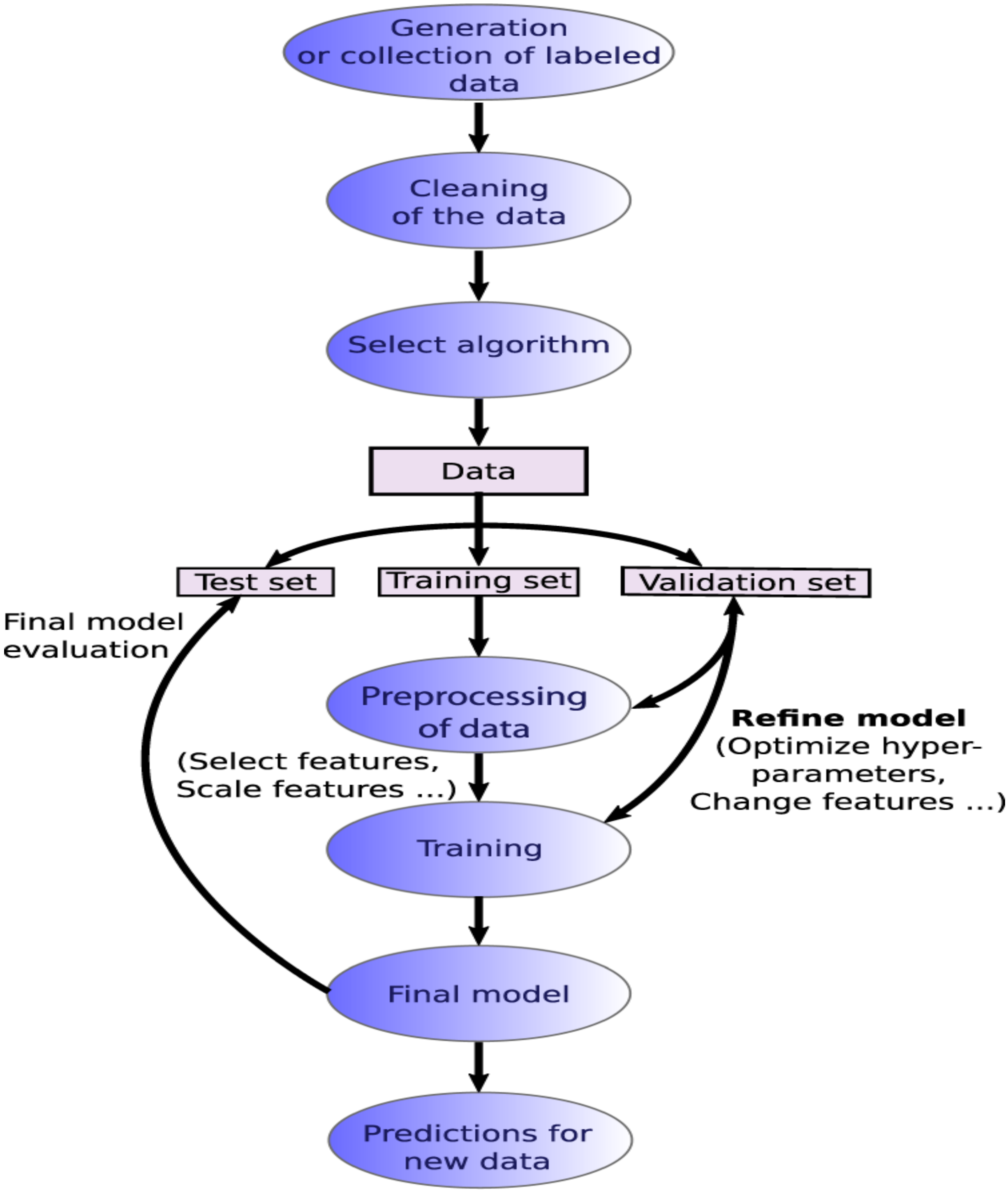

Prediction column: Life expectancy

Rank by: Root mean squared err...

Score: Cross validation

THIS SHOWS THE PIPELINING COMPARISION.

FLOWCHART


Generation or collection of labeled data
Cleaning of the data
Select algorithm
Data
Test set
Training set
Validation set
Final model evaluation
Preprocessing of data
Refine model
(Optimize hyper-parameters,
Change features ...)
(Select features,
Scale features ...)
Training
Final model
Predictions for new data

RESULT-WE HAVE SUCCESSFULLY MADE THE MACHINE LEARNING MODEL TO PREDICT LIFE EXPECTANCY USING PYTHON AND GIVEN DATASET**.**

ADVANTAGES/DISADVANTAGES:

Comparison to human performance

To put the reported results in perspective, we provided a comparison of the model's performance to human performance as described by [15]. To make a truly valid comparison, our study design should include judgments about life expectancy from GPs about the actual patients that the medical records used for this research correspond to. Making this comparison was however impossible within the scope of this research, and with the use of this dataset.

Data limitations

One of the main challenges we faced during this research was the amount of available data. Our dataset consisted of roughly 1200 patients which is a fair amount of data according to clinical standards, but is not considered to be a lot of data for training neural networks. We partially addressed this problem by splitting each medical record into fifty time slices, thereby increasing the number of cases with a factor of fifty. However, more data would have been desirable for training the model, in order to increase the accuracy and reduce overfitting.

Interpretation of the output

We choose to return a probability distribution for a large range of months, rather than producing a single-value prediction or a classification with few classes. While such output indeed delivers very interesting results, we also needed a way to operationalize these probability distributions in order to evaluate the model's performance.

Transparency

When it comes to incorrect predictions, both the baseline and the keyword model tend to make overly pessimistic predictions. It would be interesting to investigate *why* the models have a tendency toward overly pessimistic predictions, despite being trained with and tested on balanced data.

APPLICATIONS- Probing into sanitation: To dig deeper into how sanitation specifically, I evaluated access to sanitation against statistics on child mortality rates. Not surprisingly, there's a strong negative correlation between sanitation and mortality: as sanitation improves, mortality rates decrease for neonates, infants, and children under 5.

According to the World Health Organization, diarrheal disease, the leading cause of death for children under five, is spread by poor sanitation conditions:

'Diarrhoea is a symptom of infections caused by a host of bacterial, viral and parasitic organisms, most of which are spread by faeces-contaminated water.'

FUTURE SCOPE- One possibility is that the pace of age-specific mortality improvement over the next half century will be similar to the pace of improvement over the last 50 or 100 years.

A second possibility is that the pace of life-expectancy increase over the next half century will be similar to the rate of increase over past decades.

Finally, the third possibility is that mortality improvements will accelerate in the future. Biology and biomedicine may be on the verge of unprecedented breakthroughs in knowledge about specific diseases and about the aging process itself – many knowledgeable scientists are of this opinion. Specifically, instead of increasing by 2.5 years per decade, life expectancy may increase by 3, then 4, and then 5 years per decade over the next three decades and perhaps by 6, 8, or even 10 years per decade in the 2030s and 2040s.

CONCLUSION- Even though the model's performance is far from perfect, we consider this work to be among the first steps in a line of research that has much potential for clinical applications, for several reasons: good prognostication has the potential to contribute significantly to end-of-life decision making, therefore we believe that any increase in prognostic accuracy is worth persuing. Additionally, human prognostication is costly, time-consuming, requires medical expertise, and is a subjective task. Without compromising prediction accuracy, the model is able to make predictions quickly, automatically and systematically, while it does not depend on human medical expertise. Even though the model reaches only 29% accuracy, we consider 9% point improvement to be promising, considering that the model is trained on a relatively small data sample.

BIBLIOGRAPHY

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0775-2#Sec24

https://link.springer.com/chapter/10.1007/978-3-030-05075-7_6

https://cloud.ibm.com/docs/overview?topic=overview-whatis-platform

https://www.kaggle.com/kumarajarshi/life-expectancy-who

(DATA SET REFERENCE)