PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

Life expectancy is a statistical measure of the average time a human being is expected to live. It depends on many factors like regional variations, education, economic circumstances, sex difference, mental illness, year of birth, physical illness and many more factors. But, in a country it can be narrowed to some of the specific factors.

The factors that effects life expectancy in a country

GDP	Education	Alcohol Intake	Expenditure on health care system.	Specific disease related deaths.

This project could be completed task wise as assigned.

TASK-1) Project planning and kickoff
TASK-2) Explore IBM cloud platform
TASK-3) Explore IBM Watson services
TASK-4) Introduction to Watson studio
TASK-5) Predicting life expectancy with python
TASK-6) Predicting life expectancy without python.

PROJECT SUMMARY:- to predict life expectancy with machine learning.
PROJECT REQUIREMENTS:- IBM cloud, python, github.
PROJECT SCHEDULE:- to complete within 30 days.

SCOPE

The problem of processing datasets such as electronic medical records(EMR) and their integration with genomics, environmental factors, socioeconomic factor and patient behavior variations have posed a problem for researchers the health industry. Due to rapid innovations in machine learning field such as big data, analytics,visualization, deep learning, health workers now have improved way of processing, and developing meaningful information from huge datasets that have been accumulated over many years .

Big data and machine learning can benefit public health researchers with analyzing thousands of variables to obtain data regarding life expectancy. We can use demographics of selected regional areas and multiple behavioral health disorders across regions to find correlation between individual behavior indicators and behavioral health outcomes.

Schedule and strategies

Dataset preparation and preprocessing

Data is the foundation for any machine learning project. The second stage of project implementation is complex and involves data collection, selection, preprocessing, and transformation. Each of these phases can be split into several steps.

Data collection

This is the first step in a machine learning project.We have to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.The type of data depends on what you want to predict. There is no exact answer to the question “How much data is needed?” because each machine learning problem is unique. In turn, the number of attributes data scientists will use when building a predictive model depends on the attributes’ predictive value.

Data visualization

A large amount of information represented in graphic form is easier to understand and analyze. Some companies specify that a data analyst must know how to create slides, diagrams, charts, and templates.Most of the times visualization helps us in finding correlations and outliers which are not visible when we look at the raw data.

Labeling

Supervised machine learning, entails training a predictive model on historical data with predefined target answers. An algorithm must be shown which target answers or attributes to look for. Mapping these target attributes in a dataset is called labeling.Data selectionAfter having collected all information, we choose a subgroup of data to solve the defined problem.

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

Data preprocessing

The purpose of preprocessing is to convert raw data into a form that fits the required model . Structured and clean data helps in getting more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

DATA TRANSFORMATION

In this final preprocessing phase, we transform or consolidate data into a form appropriate for machine learning. Data can be transformed through scaling, normalization, attribute decompositions, and attribute aggregations. This phase is also called feature engineering.

Dataset splitting

Any dataset for predictive analysis should be partitioned into three subsets — training, valiidation and test sets

Training set

We create a training set to train a model and define its optimal parameters known as hyperparameters which helps in increasing the accuracy of the model in case of classification or decreasing the loss in case of regression task.

Validation set

The validation set is used to evaluate a given model, but this is for frequent evaluation. We use this data to fine-tune the model hyperparameters. Hence the model occasionally sees this data, but never does it “Learn” from this. We use the validation set results and update higher level hyperparameters. So the validation set in a way affects a model, but indirectly. A small portion of data is separated from training set and used as validation dataset.

Test set.

The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained(using the train and validation sets). The test set is generally what is used to evaluate competing models . Many a times the validation set is used as the test set, but it is not good practice. The test set is generally well curated. It contains carefully sampled data that spans the various classes that the model would face, when used in the real world.

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

Model training

After we have preprocessed the collected data and split it into three subsets,we can proceed with a model training. This process entails “feeding” the algorithm with training data.

Modeling

During this stage, we train numerous models to see which one of them provides the most accurate predictions. We can use cross validation to find the most suitable hyperparameters. In this stage we observe the loss from our model and introduce new parameters like l1,l2 regularization ,weight decay to avoid overfitting.

Deployment

The wml_credentials (created during watson studio instantiation phase) are used to save the model and create a scoring endpoint for our model which will be used in node red application.A flow is constructed using different components of nod red like forms, https requests, text fields, functions. Input is given to the application through a form and the functions are supplied with API keys, Instance IDs and scoring endpoint to connect to the model and create an output. The output is displayed through a text field.

Technical Requirements

A basic knowledge of machine learning algorithms and practices along with mathematics knowledge is required to perform the regression task.In-depth knowledge of python and different machine learning libraries.Knowledge and practice is required to use different IBM Cloud services like watson studio and node red.

Software Requirements

Any environment which can run python. For example Anaconda distribution, which is excellent for all sorts of data science purposes.

Access to IBM Cloud services for deployment.

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

Project Deliverables

Python notebook containing all the code.
A node red application which can input data and outputs a prediction for life expectancy.
A json file containing the architecture of node red project.
Notebook created using Autoai.
URL of the node red application .