A

Project Report On

**Predicting Life Expectancy using Machine Learning**

**Internship under:**

**TheSMARTBRIDGE**

**NAME:** MURALI KRISHNA POLLAI MANDALA

**EMAIL**: [muraliiit3@gmail.com](mailto:muraliiit3@gmail.com)

**PROJECT ID :** SPS\_PRO\_215

**INTERNSHIP TITLE :**Predicting Life Expectancy using

Machine Learning– SB53841

**Category:** Machine Learning

|  |  |
| --- | --- |
| **Table Of Content**  TOPICS PAGENO |  |
| 1. INTRODUCTION 3-5  Overview  Purpose  2. LITERATURE SURVEY 5-6  Existing Problem  Proposed solution  3. THEORITICAL ANALYSIS 7  Block Diagram  Hardware/Software design  4. EXPERIMENTAL INVESTIGATION 8-20  SCREENSHOTS  5. FLOWCHART 20  6. RESULT 21  7. ADVANTAGES & DISADVANTAGES 22-23  8. APPLICATION 23  9. CONCLUSION 24  10. FUTURE SCOPE 24  11. BIBILIOGRAPHY 25  APPENDIX 26  A. SOURCE CODE |  |

**INTRODUCTION**

**Overview**:

This project “Predicting Life Expectancy using Machine Learning” is an web application that predict the expected average life span of people of a given country based on various features. This project is built using IBM services(Watson studio, Node Red, Watson machine learning).

A typical Regression Machine Learning project leverages historical data to

predict insights into the future. This problem statement is aimed at

predicting the Life Expectancy rate of a country given various features.

Life expectancy is a statistical measure of the average time a human being

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 the 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.

**Purpose:**

Life expectancy is the most important factor for 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.

**Project Scope ,Schedule ,Team & Deliverables:**

**Project Team** : MURALIKRISHNA POLLAIMANDALA(single)

**Project Scope:**

* This Project will predict the average time a human being is expected to live based on some factors.
* A country can predict the expected life of their citizens.According to that ,the country can take necessary preventive measures to improve the healthcare and other resources.
* This will serve as an example for countries assess to improve life expectancy for their citizens.
* This also helps in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

**Project Requirements:**

* **Functional Requirement:**

Predicting life expectancy rate of a country.

* **Technical requirements:**

**Software Requirements**: Python,IBM Cloud, IBM Watson,Zoho writer

**Hardware Requirements**: processor-i3 7th generation or higher

**Speed** : 2GHz or more

**Space** : 10GB or Higher

**Ram**  : 2GB or Higher

**Life expectancy with python:**

* Collect dataset required for the project from the extenal sources(kaggle flatform).
* Create necessary IBM cloud services and also create Watson studio which is avalible in the IBM services.
* Configure the watson studio and create Machine Learning service.
* Import dataset in to the Jupyter Notebook avalible in IBM watson.
* code in to the Jupyter Notebook.
* Build Node-RED flow to integrate MLservice.

**Duration** : 28 days

**Deliverable:**

A machine learning model that will predict life expectancy(software).

**Out of Scope:**

In the project, the user will not able to modify or not able to increase the

accuracy of the ML model.

**LITERATURE SURVEY**

**Existing system:**

In our regular prediction system, there are many problems exist, such as :

* whole concept of life expectancy depends on the interpretation given to “full health”.
* Or the factors used to predict the life expectancy of people are based on some associated specific features of particular fields like :
* morbidity and mortality (smoking, alcohol consumption, overweight and obesity, and physical activity)
* Health related disease
* occupational or social class, area level deprivation, geographical area of residence (urban and rural), housing tenure
* Race-based inequalities.

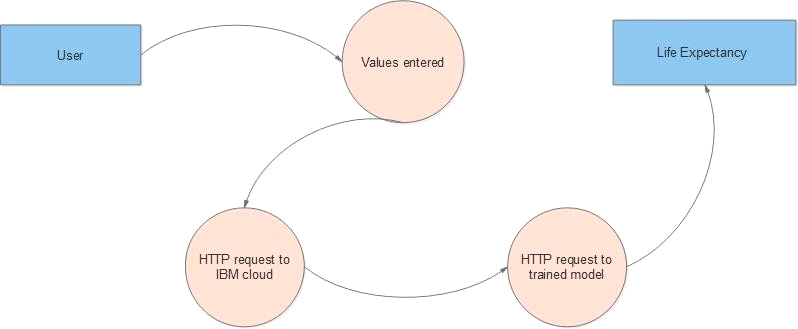
Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries.

**Proposed solution:**

* For the above problem to get solved we have a dataset consist of various factors .In this system we have taken all the correlated features into consideration. So the target output variable i.e expected life span of the people depends upon variety of factors and not factors of particular fields.
* Important immunization like Hepatitis B, Polio and Diphtheria are also considered.
* The data-set related to life expectancy, health factors for 193 countries has been collected from WHO data repository website and its corresponding economic data was collected from the United Nations website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set.
* The project uses immunization factors, mortality factors, economic factors, social factors and other health related factors to predict life expectancy of a country for a given year using a machine learning model.
* Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of its population.

**THEORITICAL ANALYSIS**

**Block/Flow Diagram:**



**Hardware / Software designing**:

* Create necessary IBM Cloud services
* Create Watson studio project
* Configure Watson Studio
* Create IBM Machine Learning instance
* Create machine learning model in Jupyter notebook
* Deploy the machine learning model
* Create flow and configure node
* Integrate node red with machine learning model
* Deploy and run Node Red app.

Input is taken from the user using a “Form” element in Node-Red. Then, an HTTP request is made to the IBM cloud that further makes an HTTP request to the deployed model using model’s instance id. After verification of id, the model sends an HTTP response which is finally parsed by the Node-Red application and the result is displayed on the user screen.

**EXPERIMENTAL INVESTIGATIONS**

**Following factors are taken into account for predicting the life expectancy of a country.**

* Country
* Status: Developed or Developing status of the country.
* Year
* Adult mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population).
* Infant deaths: Number of Infant Deaths per 1000 population.
* Alcohol: Alcohol, recorded per capita (15+) consumption.
* Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
* Hepatitis B: Hepatitis B =immunization coverage among 1-year-olds (%).
* Measles: Measles - number of reported cases per 1000 population.
* BMI: Average Body Mass Index of entire population.
* Under-five deaths: Number of under-five deaths per 1000 population.
* Polio: Polio (Pol3) immunization coverage among 1-year-olds (%).
* Total expenditure: General government expenditure on health as a percentage of total government expenditure (%).
* Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year- olds (%).
* HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years).
* GDP: Gross Domestic Product per capita (in USD).
* Population: Population of the country.
* Thinness 10-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19(% ).
* Thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%).
* Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
* Schooling: Number of years of schooling.

**Finding the most suitable algorithm:** Random forest gives highest accuracy



**Steps**

Create IBM Cloud services

* Watson Studio
* Watson Machine Learning
* Node Red
* Create **Watson Studio** service instance.
* Select **Catalog** found at the top right of the page.
* Click on **Watson** from the menu on the left, which you can find under **Platform** services.
* Select Watson Studio.
* Enter the **Service name** or keep the default value and make sure to select the **US South** as the **region/location** and your desired **organization**, and **space**.
* Select **Lite** for the **Plan**, which you can find under **Pricing Plans** and is already selected. Please note you are only allowed one instance of a Lite plan per service.
* Click on **Create**.
* You will be taken to the main page of the service. Click on **Get Started**.
* Create a New Project
* Add WML service
* Click on the **Settings** in the project view, locate **Associated services** => **Add Service** => **Watson**.
* You should also create a **Access Token** in the project setting. Click on **New token**, give it a name, then click **Create**.
* Create Notebook

Click **Add to project** => **Notebook**

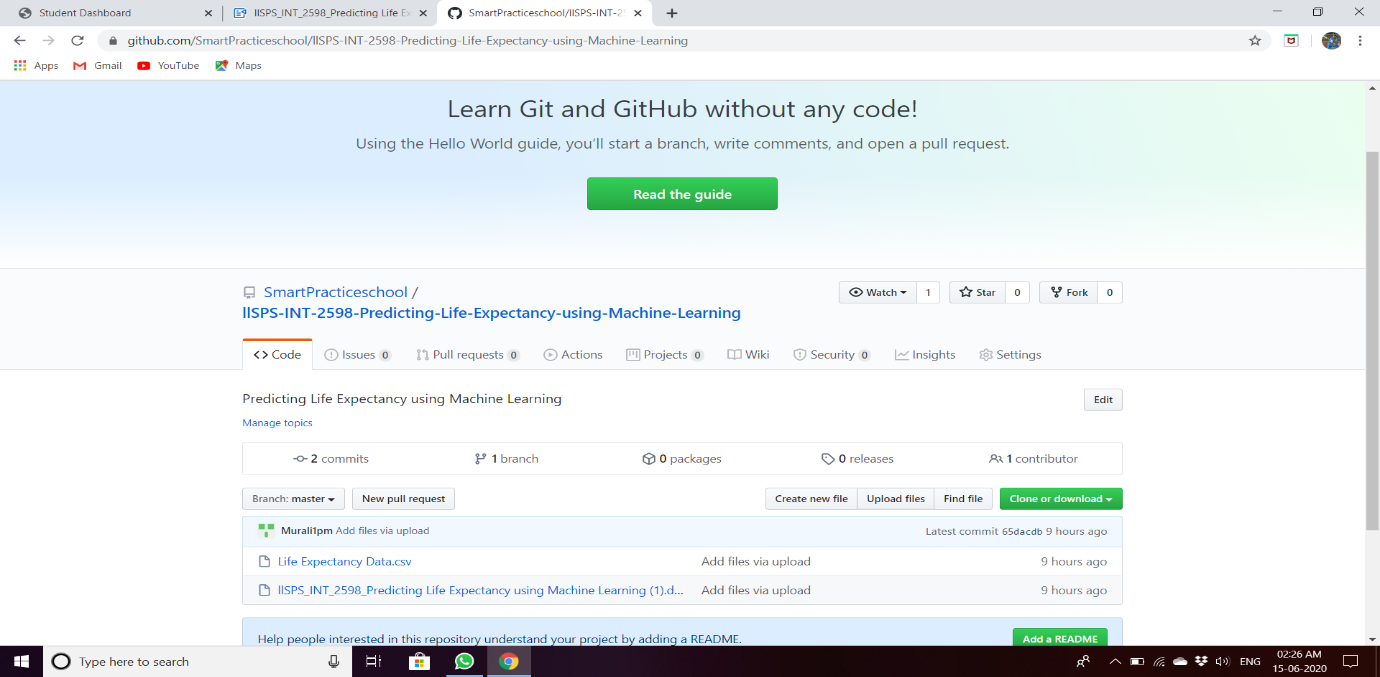
And create your Model here.

* Deploy Model as Web Service
* Build Node-RED Flow To Integrate ML Services

**What is GitHub?**

* Git is a free and open source **distributed version control system** designed to handle everything from small to very large projects with speed and efficiency.
* Git relies on the **basis of distributed development** of a software where more than one developer may have access to the source code of a specific application and can modify changes to it which may be seen by other developers..
* Initially designed and developed by [Linus Torvalds](https://en.wikipedia.org/wiki/Linus_Torvalds) for [Linux kernel](https://www.kernel.org/) development in 2005.
* Every git working directory is **a full-fledged repository** with complete history and full version-tracking capabilities, independent of network access or a central server.

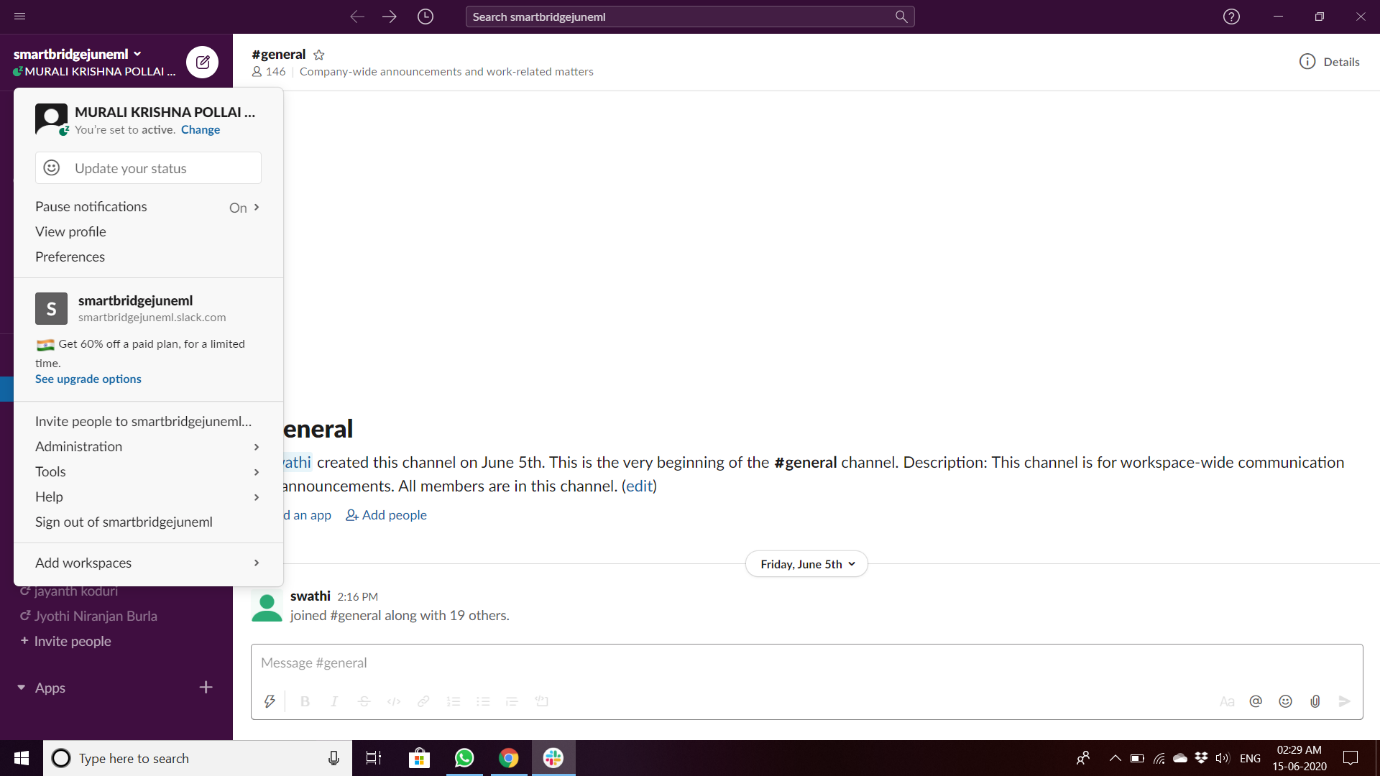
**GitHub Creation:**



**What is Slack?**

* Slack is a communication tool that is created to streamline and simplify conversations.It's similar to other messages apps with the benefit of searchable.
* It is the best flat form where we can share our thoughts with our our team mates.

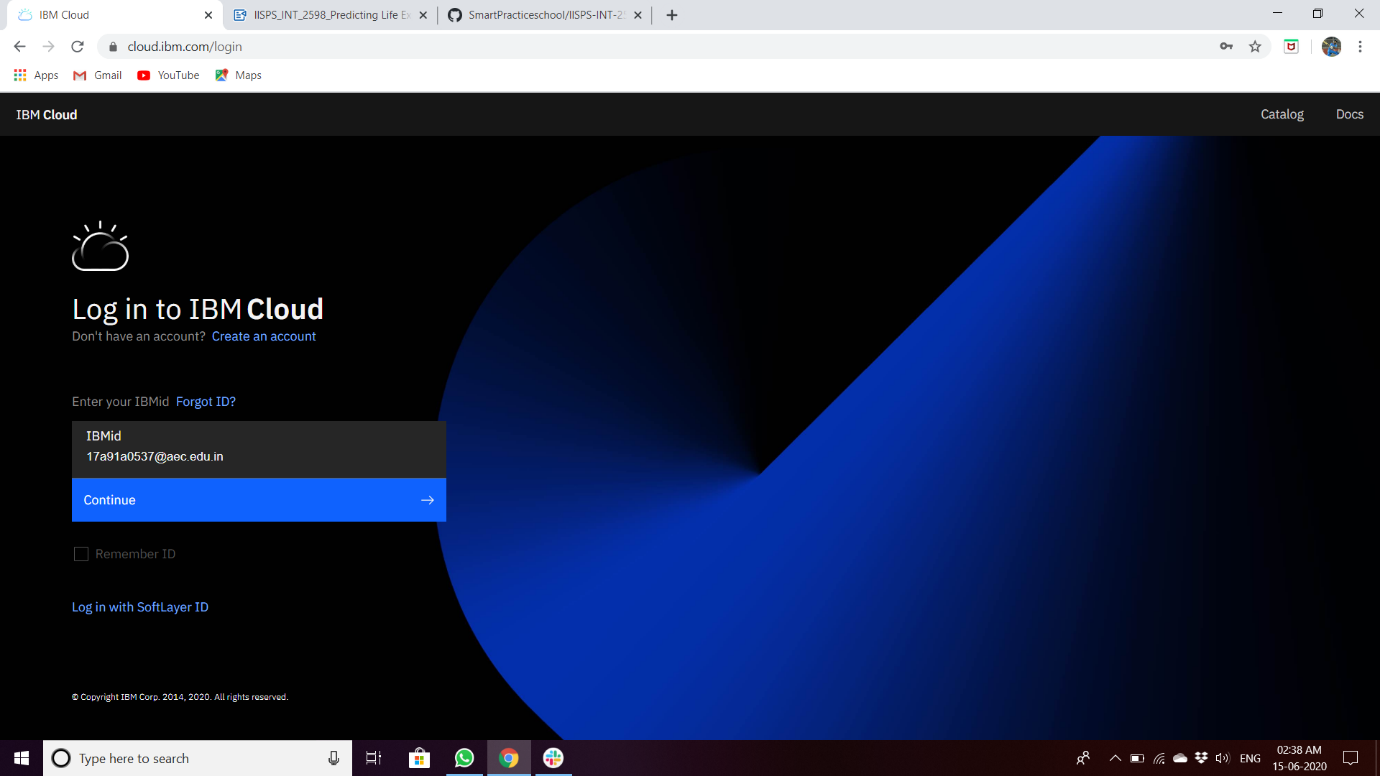
**Slack app:**



**IBM Cloud:**

* IBM cloud computing is a set of cloud computing  services for business offered by the  information technology company IBM.
* IBM Cloud  includes infrastructure as a service  (IaaS),  software as a service  (SaaS) and platform as a service  (PaaS) offered through public, private and hybrid cloud delivery models, in addition to the components that make up those clouds.

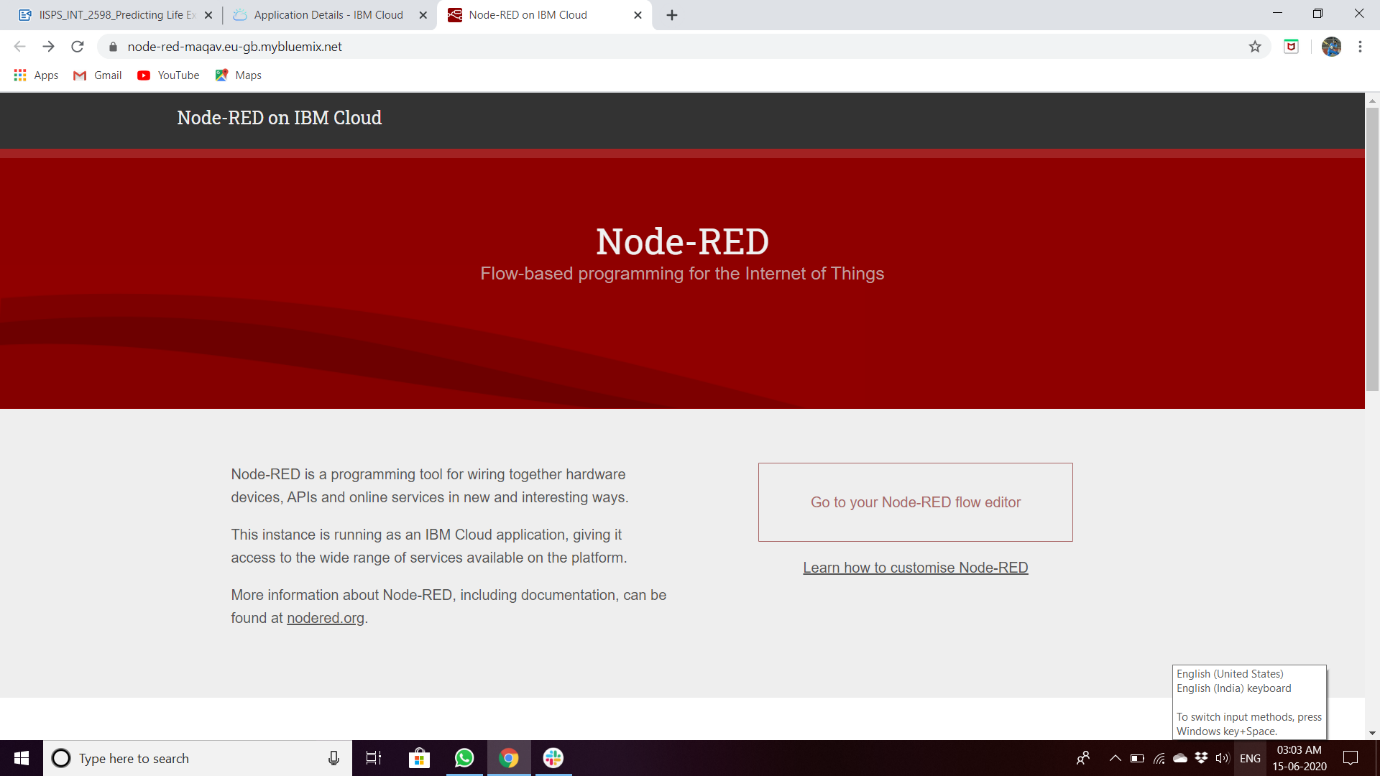
**IBM account creation:**

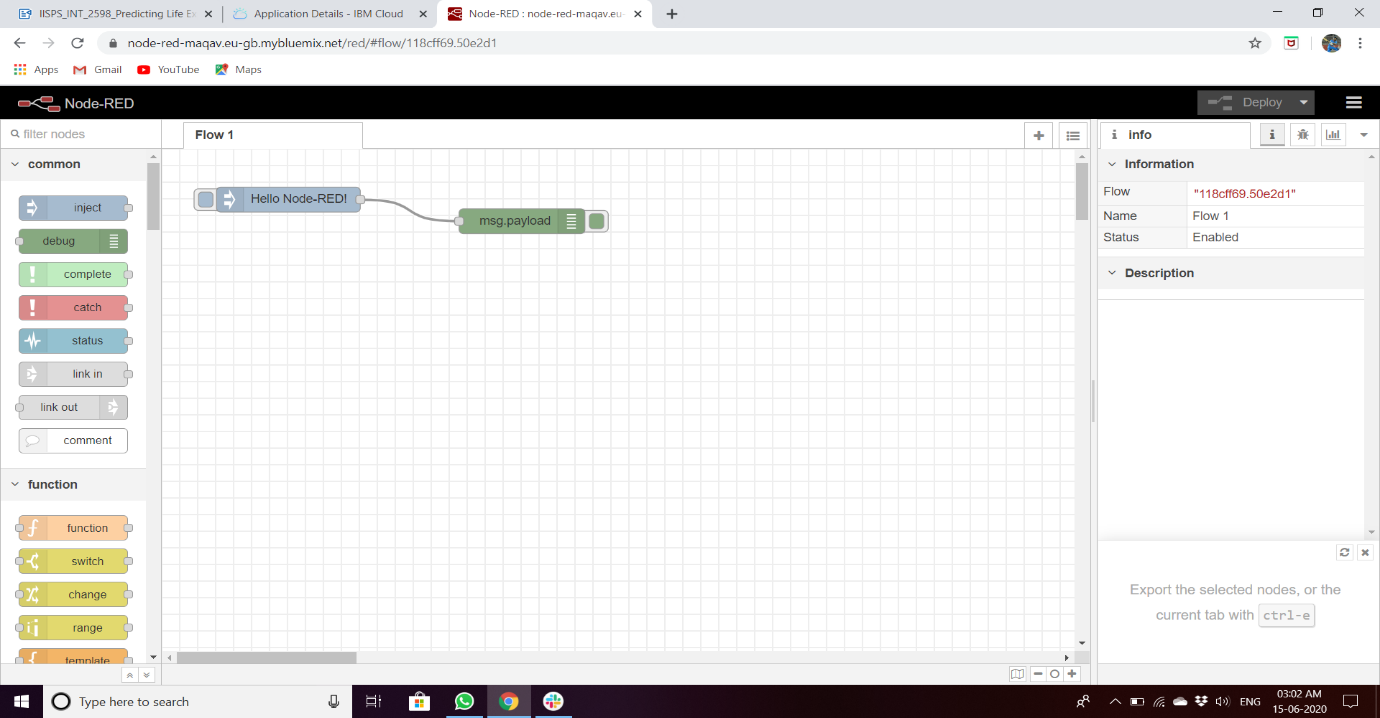


**Node-RED:**

Node-RED is built on Node.js, taking full advantage of its event-driven, non-blocking model. This makes it ideal to run at the edge of the network on low-cost hardware such as the Raspberry Pi as well as in the cloud.

**Node-RED Starter App:**





**IBM Watson Use Cases:**

The interviewed organizations had similar, though somewhat varied Watson Assistant deployments. The artificial intelligence research and innovation manager of a financial services organization was an early adopter of Watson Assistant. He shared with Forrester: “The conversations began with IBM to understand Watson and see if we could find a use case. And because artificial intelligence was such a new technology, we didn’t want to have a use case that would be exposed directly to customers. So, we found the internal use case for employees.” Other interviewees focused their initial deployments on externally facing use cases. The three main categories of use cases covered in this study are as follows:

* **Agent assist:** In the report “Stop Trying To Replace Your Agents With Chatbots,” Forrester highlights agent assist as a preferred method for blending customer service automation and humans: “Using chatbots internally first is a good starting point for many firms just setting out on their chatbot journey. Your agents make an ideal and captive test bed for a bot before you expose it to your customers.
* **Customer self-service:** This use case deploys a customer-facing chatbot that can respond and contain simple queries, search for complex answers from content or a knowledge base, and properly route to a human.
* **Employee self-service:** This use case is also an internally facing utilization of Watson and is aimed at answering employee questions. The organizations interviewed for this study used Watson to augment HR and IT help desks.

**some examples:**

⦁ using Watson, Autodesk built AVA,a virtual agent designed to resolve the most  common  support issues

⦁ **Staples Easy Button i**s one of the service by pressing the button we can get what we  need.



⦁ **Self Driving Cars** one such service which is based on watson IOT for automative.

 ⦁ USTA(United States Tennis Association) uses IBM watson to enhance player performance.

**Watson at Work**:

<https://www.youtube.com/watch?v=W3iPbFTAAds&feature=youtu.be>

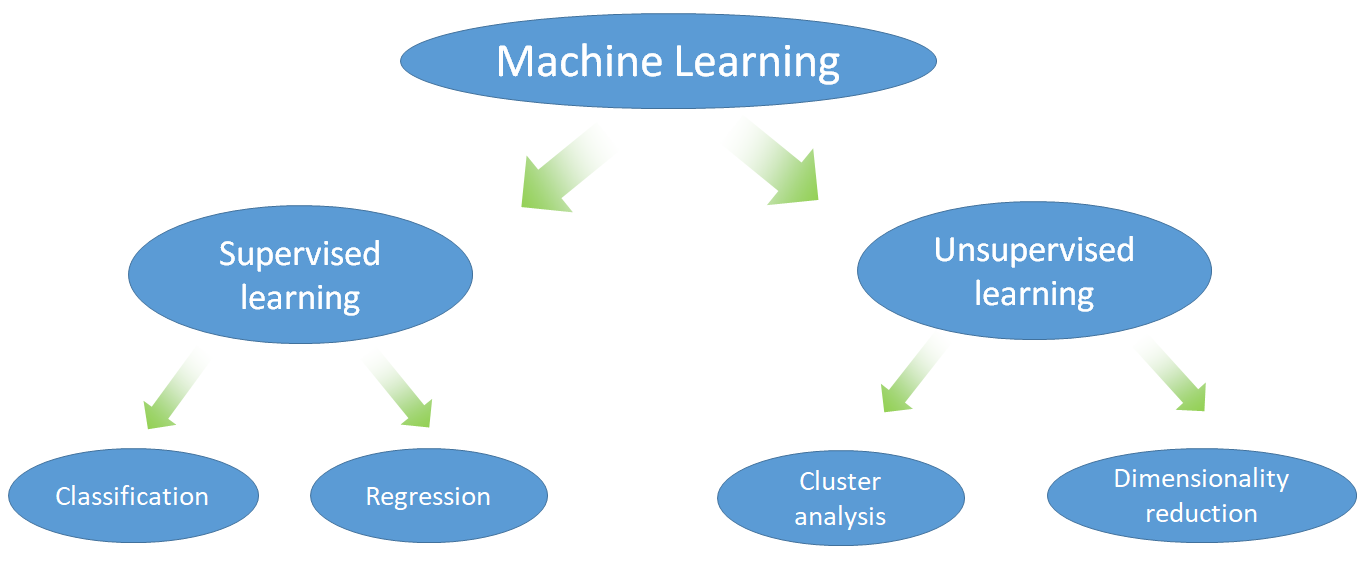
**IBM Watson Machine Learning**:

**Introduction to Machine Learning**:

**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: ***The ability to learn***. Machine learning is actively being used today, perhaps in many more places than one would expect.

**Classification of Machine Learning**

Machine learning implementations are classified into three major categories, depending on the nature of the learning “signal” or “response” available to a learning system which are as follows:



**Supervised learning :** When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning.

* **Classification :** When inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.
* **Regression :** Which is also a supervised problem, A case when the outputs are continuous rather than discrete.

**Unsupervised learning :**Whereas when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms.

* **Clustering :** When a set of inputs is to be divided into groups. Unlike in    classification, the groups are not known beforehand, making this typically an unsupervised task.

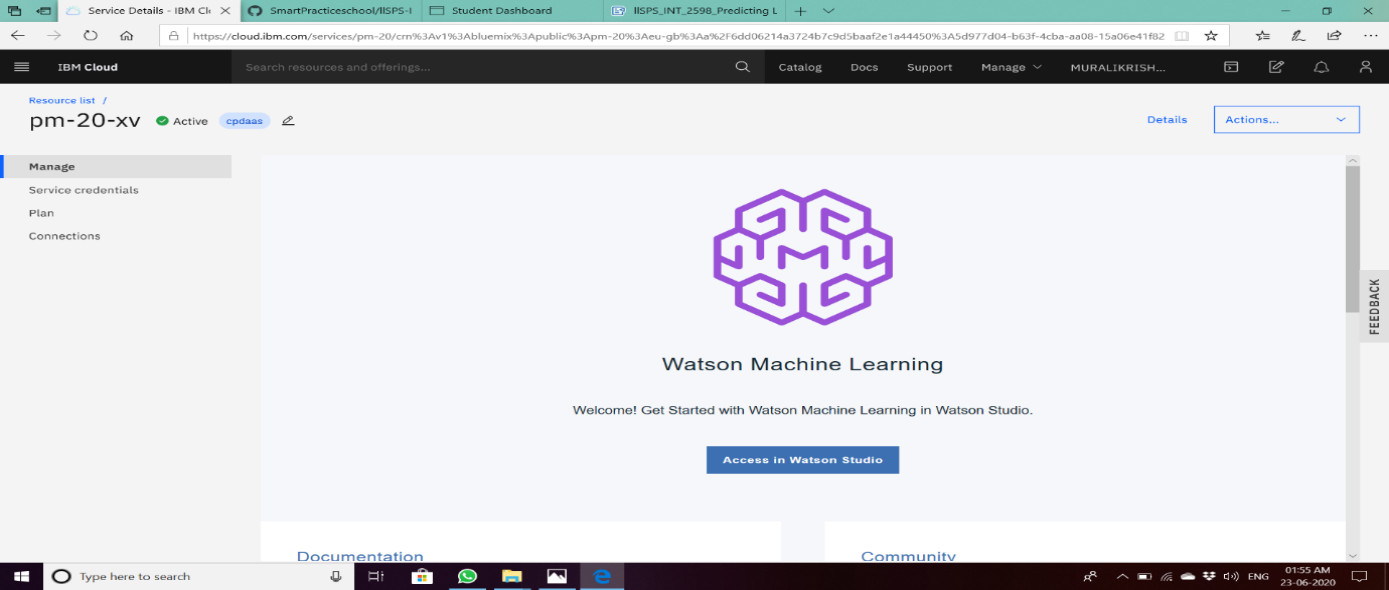
**Reinforcement learning :** When you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback according to the solution the algorithm proposes comes under the category of Reinforcement learning, which is connected to applications for which the algorithm must make decisions (so the product is prescriptive, not just descriptive, as in unsupervised learning), and the decisions bear consequences. In the human world, it is just like learning by trial and error.

**IBM Watson Machine Learning:**

<https://www.youtube.com/watch?v=NmdjtezQMSM>

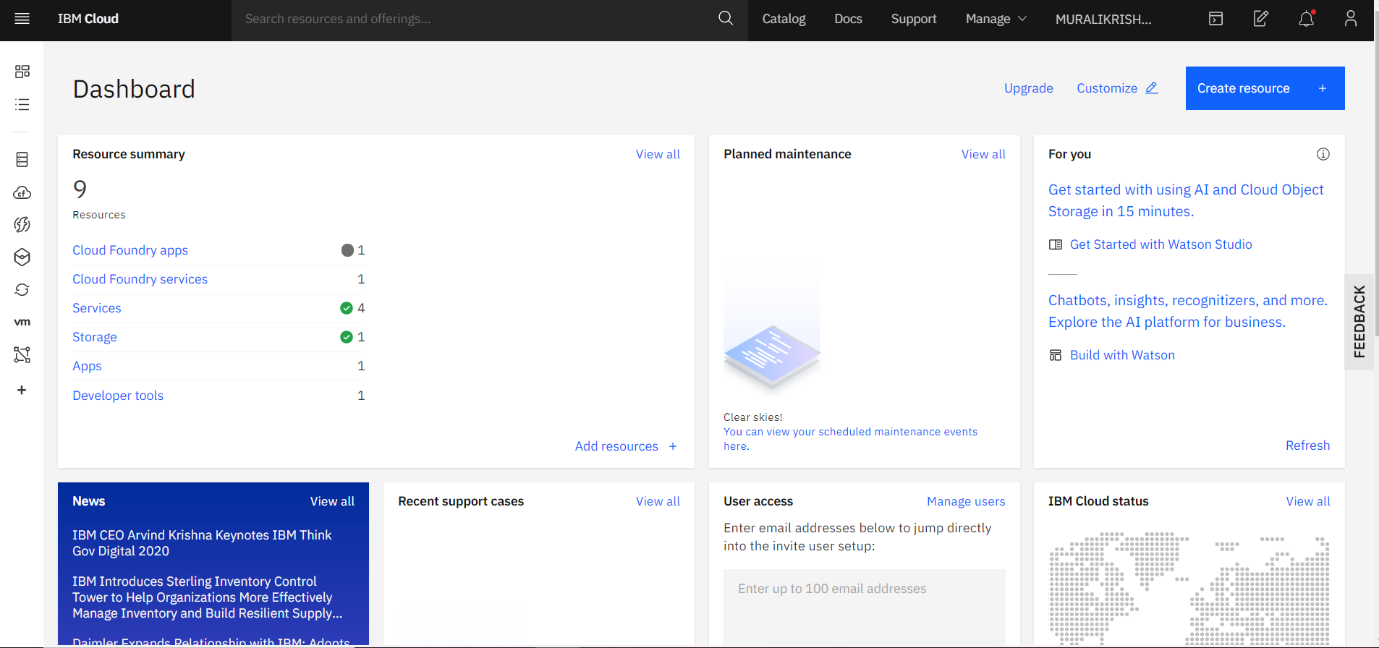
 Using IBM Watson Machine Learning, you can build analytical models and neural networks, trained with your own data, that you can deploy for use in applications.

Watson Machine Learning provides a full range of tools and services so you can build, train, and deploy Machine Learning models. Choose from tools that fully automate the training process for rapid prototyping to tools that give you complete control to create a model that matches your needs.

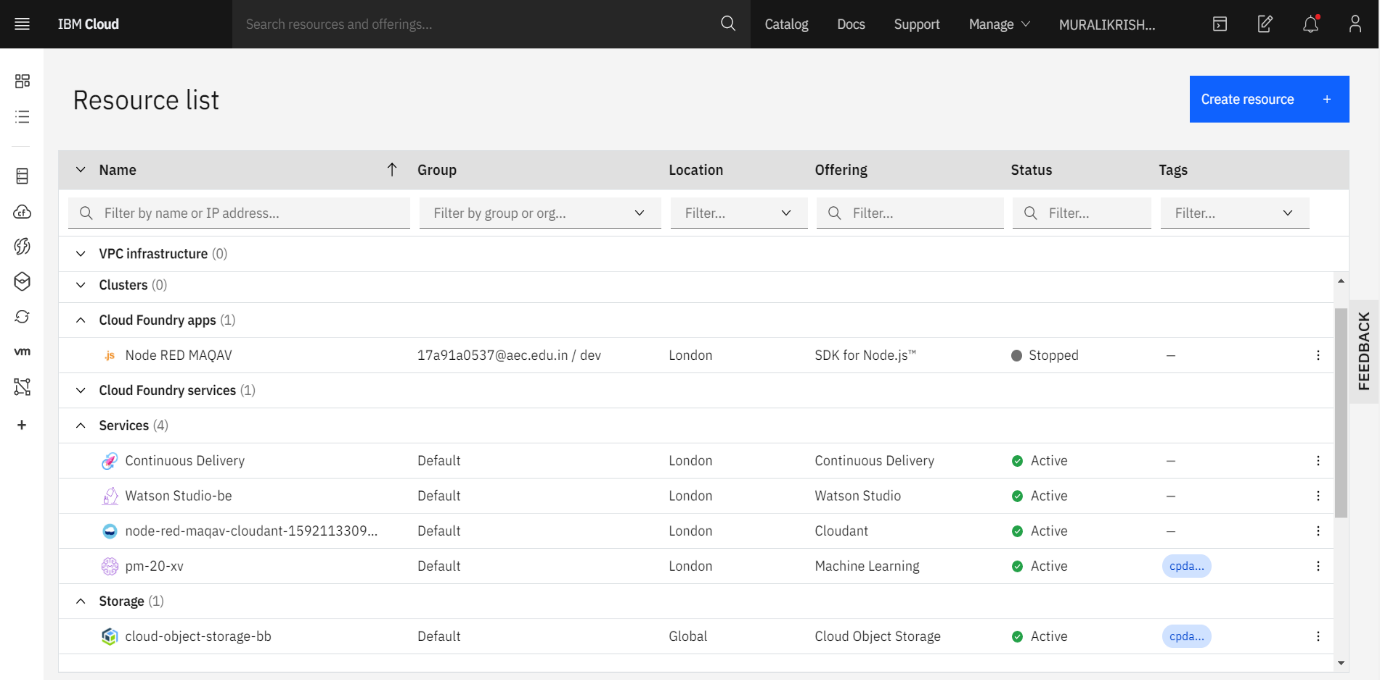


**SCREENSHOTS**

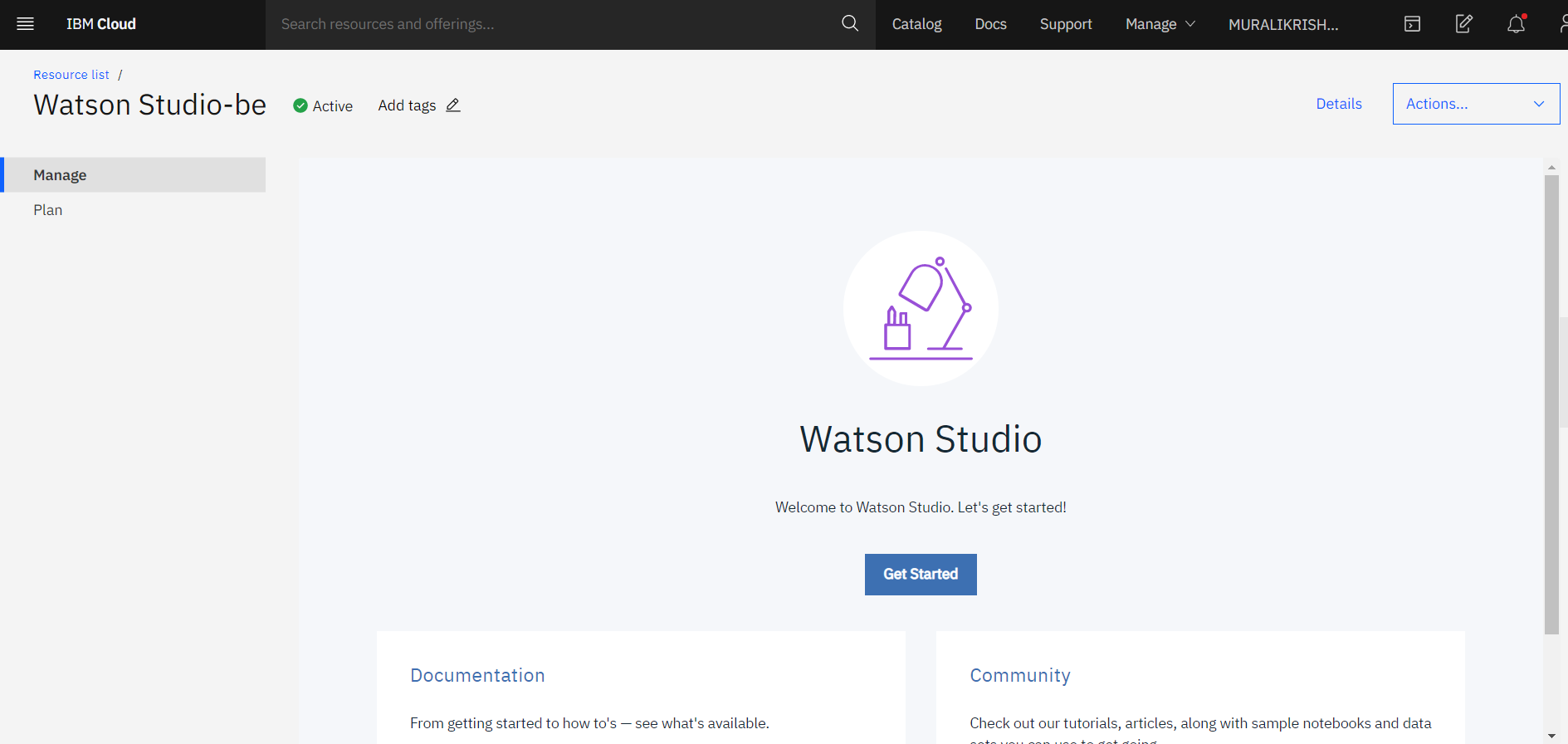
**IBM CLOUD DASHBOARD:**

****

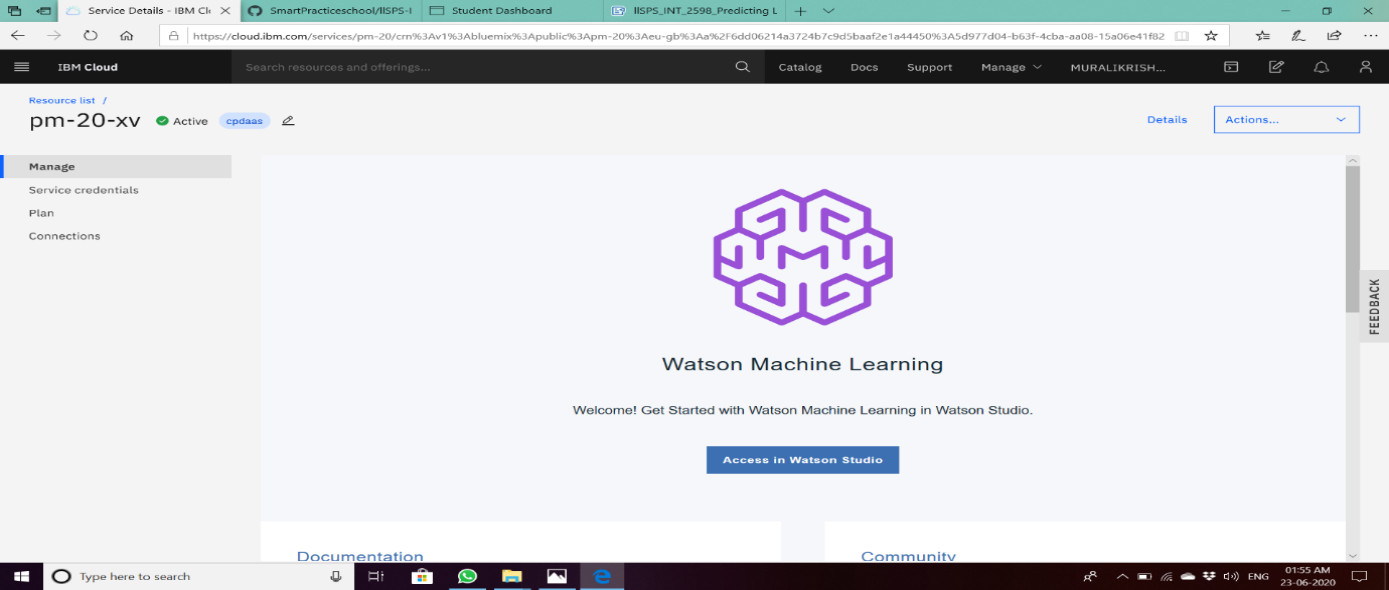
**RESOURCE LIST:**

****

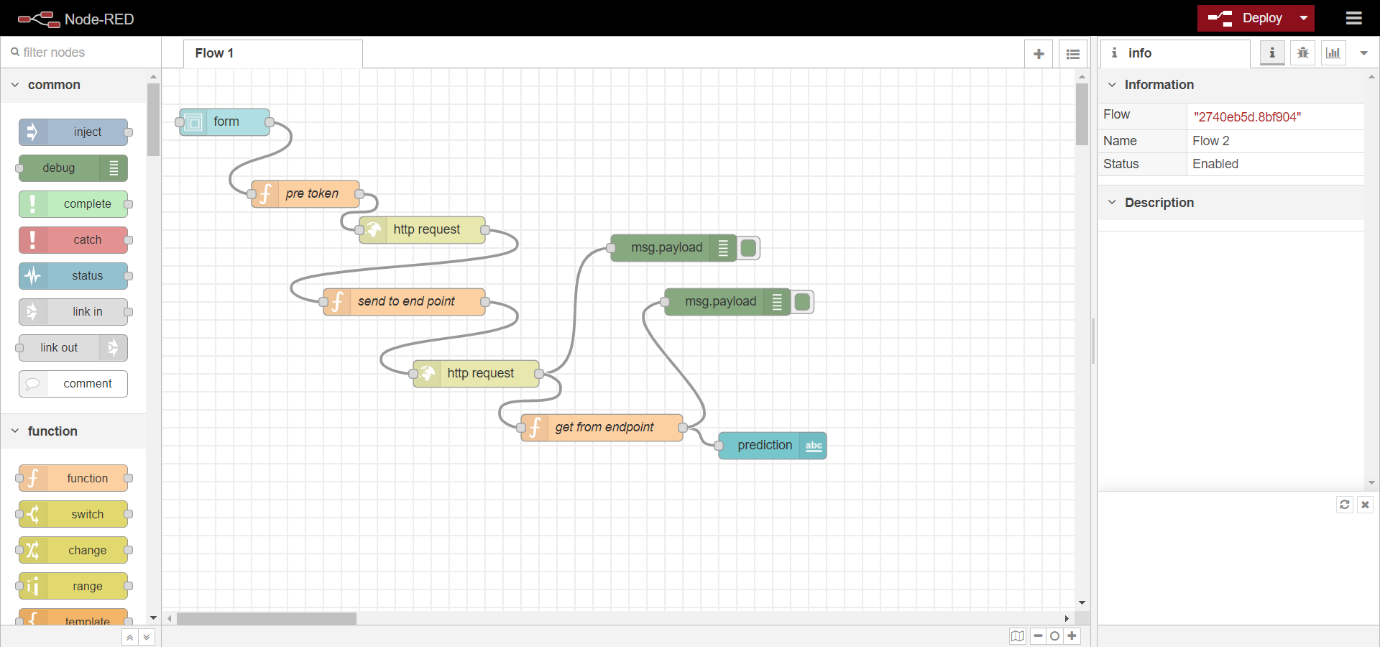
**WATSON STUDIO :**

****

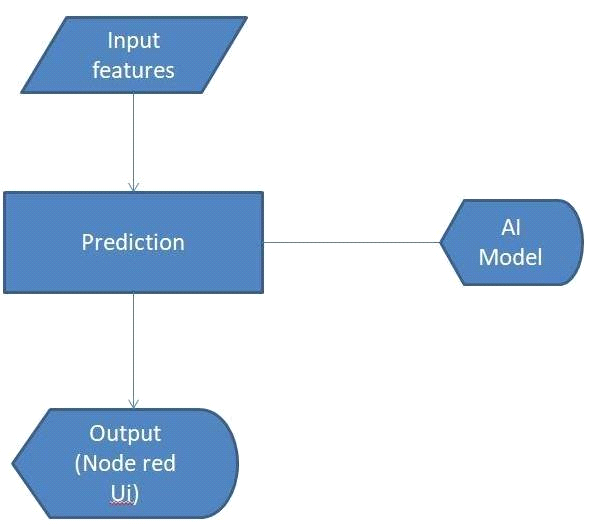
**WATSON MACHINE LEARNING SERVICE**:



**NODE RED FLOW:**

****

**FLOWCHART**

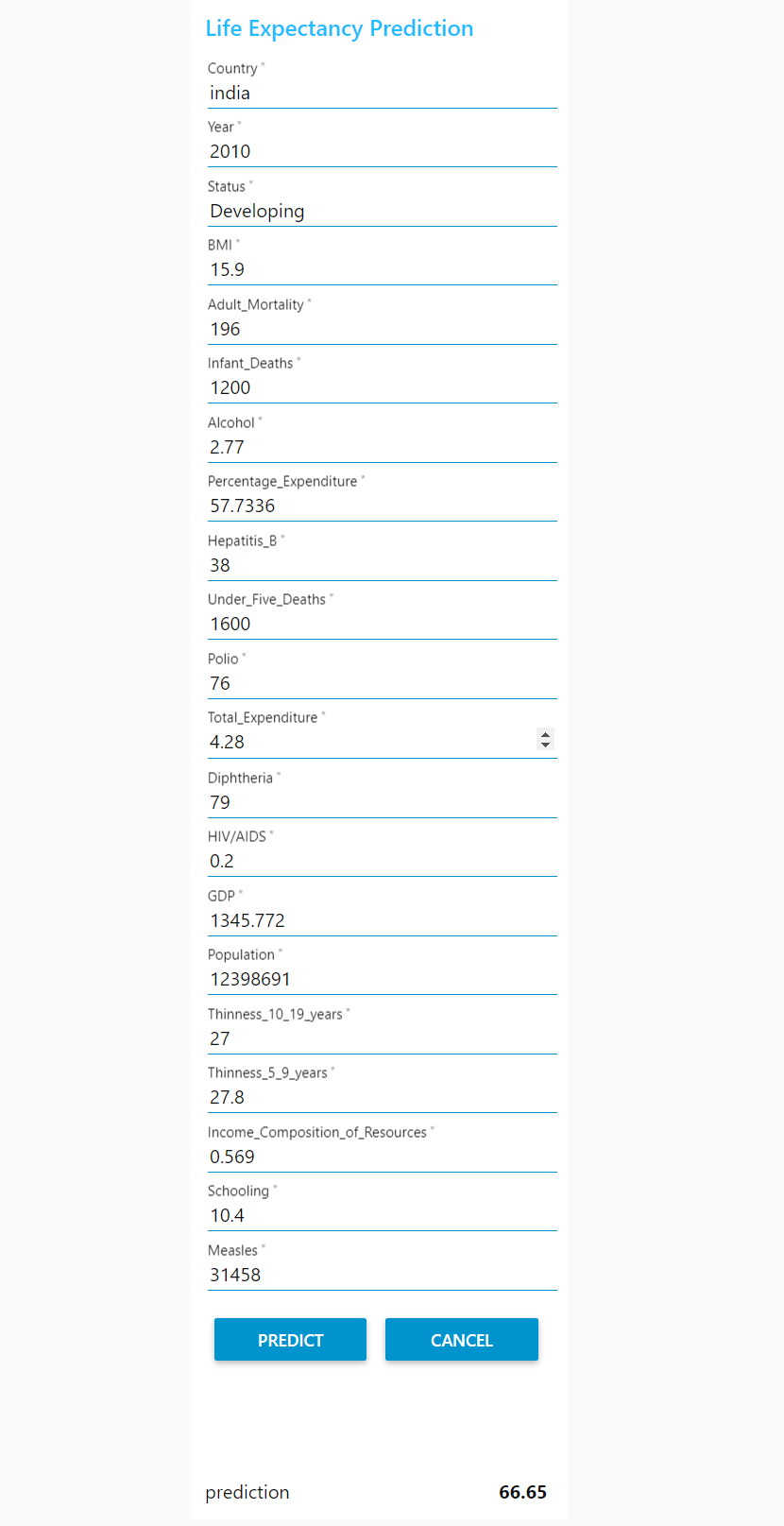


**RESULTS**

Finally our Node-RED dash board integrates all the components and displayed in the Dashboard UI by typing URL-

* [https://node-red-maqav.eu-gb.mybluemix.net/red/#flow/7fd17189.8ebe5](https://node-red-maqav.eu-gb.mybluemix.net/red/" \l "flow/7fd17189.8ebe5)
* <https://node-red-maqav.eu-gb.mybluemix.net/ui/#!/0?socketid=JjAswIY-IecTBU4vAAAM>

in browser.



**ADVANTAGES & DISADVANTAGES**

**Advantages :**

* Since the observations this dataset are based on different countries, it will be easier for a country todetermine the predicting factor which is contributing to lower value of life expectancy.
* The data-sets are made available to public for the purpose of health data analysis..
* The observations in the dataset used are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of its population.
* Some of the past research was done considering multiple linear regression based on data set of one year for all the countries. But the dataset used for training the model contained data of past 15 years to give a fairly better prediction.
* The application is easy and simple to use.
* The machine learning algorithm used in the project is Random Forest regression which is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces over fitting problem in decision trees and also reduces the variance and therefore improves the accuracy.
* Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.

**Disadvantages :**

* Can be only used by the people having the knowledge of data analysis.
* As the model is deployed on cloud, so one requires good internet connection to use the application.
* The model used is Random Forest regression and Random Forest creates a lot of trees (unlike only one tree in case of decision tree) and combines their outputs. By default, it creates 100 trees in Python sklearn library. To do so, this algorithm requires much more computational power and resources.
* Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes.
* The Node-Red application needs to make HTTP request to IBM cloud and then another HTTP request to the model before providing the prediction. That makes the application a bit slow.

**APPLICATIONS**

* This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.
* It will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy and can be used in various organization to improve the quality of service.
* The project can be used as a basis to develop personalized health applications.
* The governments can plan and develop their health infrastructures by keeping the most correlated factors in mind.
* The project can help governments to keep track of their country’s health status so they can plan for the future accordingly.

**CONCLUSION**

By doing the above procedure and all we successfully created Life expectancy prediction system using IBM Watson studio, Watson machine learning and Node-RED service. The potential use of project is not limited to health care in practice, but could also be useful in other clinical applications such as clinical trials. The project makes a good use of machine learning in predicting life expectancy of a country that can help respective government in making policies that will serve for the benefit of the nation and entire humankind.

**FUTURE SCOPE**

* Look at class within a particular country and see if these same factors are same in determining life expectancy for an individual.
* Use the Twitter API to incorporate NLP analysis for a country to see how it relates to Life Expectancy.
* Increase the dataset size with continuing UN and Global Data to incorporate new added features like population, GDP, environmental, and etc in order to test and clarify country groupings.
* As more data comes, that can be fed to the model for more accurate predictions.
* Currently, the project is just a web application. It can be developed to support other platforms like Android, IOS and Windows Mobile.
* Other regression models can also be used for prediction and later the best among them should be chosen

**BIBILOGRAPHY**

* A Systematic Literature Review of Studies Analyzing Inequalities in Health Expectancy among the Older Population (Benedetta Pongiglione, Bianca L. De Stavola,George B. Ploubidis)
* <https://cloud.ibm.com/>
* IBM Developer, “IBM Watson Studio: Create a project”, 2019. [Online]. Available: [https://www.youtube.com/watch?v=- CUi8GezG1I&list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L&index=2](https://www.youtube.com/watch?v=-%20CUi8GezG1I&list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L&index=2)
* IBM Developer, “IBM Watson Studio: Jupyter notebook basics”, 2019 [Online]. Available: <https://www.youtube.com/watch?v=Jtej3Y6uUng>
* IBM Cloud setup [Online]. Available: <https://www.ibm.com/cloud/get-started>.
* IBM Developer, “Node-RED starter tutorial” [Online]
* Available[: https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/](:%20https:/developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/%20) .
* “Node-RED labs on the use of the Watson Developer Cloud services - watsondeveloper-cloud/node-red-labs.” [Online].
* Available: <https://github.com/watsondeveloper-cloud/node-red-labs>
* “Infuse AI into your applications with Watson AI to make more accurate predictions”. [Online]. Available:

<https://www.ibm.com/watson/products-services%20>

* IBM Watson, “Intro to IBM Watson”, 2018 [Online]. Available: [https://www.youtube.com/watch?v=W3iPbFTAAds&feature=youtu.be .](https://www.youtube.com/watch?v=W3iPbFTAAds&feature=youtu.be%20.)
* “Get an understanding of the principles of machine learning.” [Online]. Available: [https://developer.ibm.com/technologies/machine-learning/series/learning-pathmachine-learning-for-developers .](https://developer.ibm.com/technologies/machine-learning/series/learning-pathmachine-learning-for-developers%20.)
* IBM Developer, “IBM Watson Machine Learning: Get Started in IBM Cloud”, 2020 [Online]. Available: [https://www.youtube.com/watch?v=NmdjtezQMSM .](https://www.youtube.com/watch?v=NmdjtezQMSM%20.%20)
* Watson Studio Workshop, “Chapter 4 Build and Deploy models in Jupyter Notebooks” [Online]. Available: [https://bookdown.org/caoying4work/watsonstudioworkshop/jn.html .](https://bookdown.org/caoying4work/watsonstudioworkshop/jn.html%20.)
* Kumar Rajarshi, “Life Expectancy (WHO) Statistical Analysis on factors influencing Life Expectancy”, 2018. [Online].
* Available: [https://www.kaggle.com/kumarajarshi/life-expectancy-who](https://www.kaggle.com/kumarajarshi/life-expectancy-who%20)
* IBM Developer, “IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog”, 2019. [Online].
* Available: [https://www.youtube.com/watch?v=DBRGlAHdj48&list=PLzpeuWUENMK2PYtas CaKK4bZjaYzhW23L](https://www.youtube.com/watch?v=DBRGlAHdj48&list=PLzpeuWUENMK2PYtas%20CaKK4bZjaYzhW23L%20) .

**APPENDIX**

**A. Source code**

* **Watson Studio**

**Life\_expectancy\_prediction.ipynb :**

<https://github.com/SmartPracticeschool/llSPS-INT-2598-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Life_expectancy_prediction.ipynb>

**Node Red App Link :**

<https://node-red-maqav.eu-gb.mybluemix.net/red/#flow/7fd17189.8ebe5>

**Github Link :**

<https://github.com/SmartPracticeschool/llSPS-INT-2598- Predicting-Life-Expectancy-using-Machine-Learning>

**Thank you**