**HUMAN DEVELOPMENT INDEX USING IBM WATSON**

A MINOR PROJECT REPORT

Submitted to

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY, HYDERABAD**

In partial fulfilment of the requirements for the award of the degree of

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

Submitted by

**JAKKOJU SHRUTHI 18UK1A05D6**

**MUZAFFAR HUSSAIN KHAN 18UK1A0590**

**KOKKIRALA AKSHITHA 18UK1A05E6**

**RUDROJU SATHISH 18UK1A05A4**

Under the esteemed guidance of

**MS. A. SWATHI**

(Assistant professor)

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**VAAGDEVI ENGINEERING COLLEGE**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**VAAGDEVI ENGINEERING COLLEGE**

(Affiliated to JNTU Hyderabad & Approved by AICTE, New Delhi) Bollikunta , Warangal – 506005

**2018-2022**

****

**CERTIFICATE**

This is to certify that the Major Project Report entitled “HUMAN DEVELOPMENT INDEX USING IBM WATSON” is being submitted by J.SHRUTHI (H.NO:18UK1A05D6), MUZAFFARHUSSAINKHAN (H.NO:18UK1A0590) , K.AKSHITHA (H.NO:18UK1A05E6), R.SATISH (H.NO:18UK1A05A4) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering to Jawaharlal Nehru Technological University Hyderabad during the academic year 2021-22, is a record of work carried out by them under the guidance and supervision.

**Project guide Head of the department**

**Ms. A. Swathi Dr. R. Naveen Kumar**

(Asst professor) (professor)

**External**

**ACKNOWLEDGEMENT**

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dr.P.PRASAD RAO, Principal, Vaagdevi Engineering College for making us available all the required assistance and for his support and inspiration to carry out this major project in the institute.

We extend our heartfelt thanks to Dr.R.NAVEEN KUMAR, Head of the Department of CSE, Vaagdevi Engineering College for providing us necessary infrastructure and there by giving us freedom to carry out the major project.

We express heartfelt thanks to Mr.Ch.Jayaprakash, Program Manager, SmartBridge Educational Services Private Limited, for their constant supervision as well as for providing necessary information regarding the major project and for their support in completing the major project ,mini project and internship.

We express heartfelt thanks to the guide, Ms.A.Swathi, Assistant professor, Department of CSE for her constant support and giving necessary guidance for completion of this major project.

Finally, we express our sincere thanks and gratitude to my family members, friends for their encouragement and outpouring their knowledge and experience throughout the thesis.

**JAKKOJU SHRUTHI 18UK1A05D6**

**MUZAFFAR HUSSAIN KHAN 18UK1A0590**

**KOKKIRALA AKSHITHA 18UK1A05E6**

**RUDROJU SATHISH 18UK1A05A4**

**TABLE OF CONTENT**

LIST OF FIGURES……………………………………………………………………………………………………………………….5

LIST OF TABLES………………………………………………………………………………………………………………………….5

LIST OF EQUATIONS…………………………………………………………………………………………………………………..5

ABSTRACT…………………………………………………………………………………………………………………………………6

1 INTRODUCTION………………………………………………………………………………………………………………………7

* 1. PROBLEM…………………………………………………………………………………………………………………8
     1. DATA EXPLORATION……………………………………………………………………………….8
  2. SOLUTION……………………………………………………………………………………………………………….10
  3. DATASET OVERVIEW………………………………………………………………………………………………10
  4. DATASET SOURCE……………………………………………………………………………………………………10

1. METHODOLOGY…………………………………………………………………………………………………………………11
   1. DATASET ATTRIBUTES…………………………………………………………………………………………….11
   2. DATA PREPROCESSING……………………………………………………………………………………………12
      1. ENCODIND AND MAPPING CATEGORICAL DATA……………………………………12
2. APPLICATION OF THE APPROPRIATE LEARNING ALGORITHM……………………………………………13
   1. INTRODUCTION AND BACKROUND OF ALGORITHM………………………………………………13
   2. WHY LINEAR REGRESSION?.........................................................................................13
   3. HYPOTHESIS……………………………………………………………………………………………………………14
   4. COST FUNCTION……………………………………………………………………………………………………14
   5. GRADIENT DESCENT………………………………………………………………………………………………15
   6. APPLICATION OF THE ALGORITHM…………………………………………………………………………15
3. RESULTS……………………………………………………………………………………………………………………………15
   1. OUTPUT SCREENS…………………………………………………………………………………………………17

5. CONCLUSION………………………………………………………………………………………………………………………19

6. APPENDIX……………………………………………………………………………………………………………………………19

7. REFERENCCES………………………………………………………………………………………………………………………27

**LIST OF FIGURES**

Figure 1: HDI distribution of Sri Lanka………………………………………………………………………………………..7

Figure 2: HDI versus Country………………………………………………………………………………………………………8

Figure 3: HDI versus Mean years of schooling…………………………………………………………………………….8

Figure 4: HDI versus Life expectancy………………………………………………………………………………………….9

Figure 5: HDI versus Gross national income (GNI) per capita……………………………………………………..9

Figure 6: HDI versus GNI per capita rank minus HDI rank……………………………………………………………9

Figure 7: architecture……………………………………………………………………………………………………………….10

Figure 8: dataset overview……………………………………………………………………………………………………….10

Figure 9: Label encoding…………………………………………………………………………………………………………..12

Figure 10: Mapping of string datatype to numeric datatype…………………………………………………….12

Figure 11: Dataset after mapping…………………………………………………………………………………………….13

figure 12: distribution of human development index……………………………………………………………….15

Figure 13: Correlation matrix……………………………………………………………………………………………………16

Figure 14: predicted y versus Testing y…………………………………………………………………………………….16

Figure 15: Accuracy………………………………………………………………………………………………………………….17

Figure 16: output screen – 1…………………………………………………………………………………………………….17

Figure 17: output screen-2……………………………………………………………………………………………………….18

Figure 18: final output …………………………………………………………………………………………………………….18

**LIST OF TABLES**

Table 1 : Dataset attributes……………………………………………………………………………………………………..11

**LIST OF EQUATIONS**

Equation 1: Hypothesis of MLR………………………………………………………………………………………………..14

Equation 2 : Cost function of MLR……………………………………………………………………………………………14

Equation 3: Gradient descent of MLR………………………………………………………………………………………15

**ABSTRACT**

The Human Development Index (HDI) is a statistical composite index of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers (very high, high, medium & low) of human development. A country scores a higher HDI when the lifespan is higher, the education level is higher, and the gross national income GNI (PPP) per capita is higher. The HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The HDI can also be used to question national policy choices, asking how two countries with the same level of GNI per capita can end up with different human development outcomes. In this project  we will be building a machine learning model to predict the Human Development Index of a country by taking a few important aspects as inputs. Our model will at last predict the HDI score of a country and will also tell under which category it falls into  (very high, high, medium or low).

1. **INTRODUCTION**

The Human Development Index (HDI) is a statistical tool used to measure a country's overall achievement in its social and economic dimensions. The social and economic dimensions of a country are based on the health of people, their level of education attainment and their standard of living.

**Calculation of HDI**

* Health - Life expectancy at birth
* Education - expected years schooling for school-age children and average years of schooling in the adult population
* Income - measured by Gross National Income (GNI) per capita

Above three dimensions combined to calculate the Human Development Index (HDI). The value of the Human Development Index (HDI) is between zero and one. Very high, high, medium and low are four main tiers based on the HDI. A country is in the very high tier if its HDI is in the top quartile and the low tier if it’s HDI in the bottom quartile.

Chart, line chart

Description automatically generated

Figure 1: HDI distribution of Sri Lanka

Above is example, figure shows the distribution of the Human Development Index in Sri Lanka. This shows distribution of 1980 – 2014, referring this diagram, we can identify distribution of the HDI with the year. In 1980 HDI was very low value and in 2014 it increased to near value of 0.76. Therefore we can assume the other three dimensions are increased in continuously.

* 1. **PROBLEM**

While exploring this dataset it is clearly visible the relationship between Human Development Index (HDI) with the other dependent variables. (Life Expectancy at Birth, Expected years schooling for school-age children and average years of schooling in the adult population, Gross National Income (GNI) per capita (PPP US$)) As shows in the below graphs these dependent variables takes a high value in the countries where Happiness score is a high value.

* + 1. **DATA EXPLORATION**

**Chart, scatter chart

Description automatically generated**

Figure 2 : HDI versus Country

Chart, scatter chart

Description automatically generated

Figure 3 : HDI versus Mean years of schooling

Chart, scatter chart

Description automatically generated

Figure 4 : HDI versus Life expectancy

Chart, scatter chart

Description automatically generated

Figure 5 : HDI versus Gross national income (GNI) per capita

Chart, scatter chart

Description automatically generated

Figure 6 : HDI versus GNI per capita rank minus HDI rank

* 1. **SOLUTION**

In the perspective of Machine learning, for a real world problem like this we can use Linear regression algorithm to predict Human Development. Prediction of Human Development will help the countries to analyze and identify dependent variables and do required steps to develop their Human Development Index (HDI).

* + 1. **TECHNICAL ARCHITECTURE**

**Diagram

Description automatically generated**

Figure 7 : architecture

* 1. **DATASET OVERVIEW**

**A picture containing graphical user interface

Description automatically generated**

Figure 8: dataset overview

* 1. **DATA SOURCE**

URL: <https://www.kaggle.com/undp/human-development>

1. **METHODOLOGY**
   1. **DATASET ATTRIBUTES**

|  |  |  |
| --- | --- | --- |
| **Column name** | **Data type** | **Description** |
| Country | String | Name of the Country |
| HDI rank | Numeric | Rank of the Country based on the Human Development Index |
| Human Development Index (HDI) | Numeric | The Human Development Index (HDI) is a statistical tool used to measure a country's overall achievement in its social and economic dimensions. The social and economic dimensions of a country are based on the health of people, their level of education attainment and their standard of living. |
| Life expectancy | Numeric | Average number of years that a newborn is expected to live if current mortality rates continue to apply. |
| Mean years of schooling | Numeric | Average number of completed years of education of a country's Page 8 of 23 population aged 25 years and older, excluding years spent repeating individual grades. |
| Gross National Income (GNI) per Capita | Numeric | GNI per capita is gross national income divided by midyear population. |
| Internet users | Numeric | Number of internet users. |

Table 1 : List of dataset attributes

**2.2 DATA PREPROCESSING**

**2.2.1 ENCODING CATEGORICAL DATA**

Since this is a linear regression model, there can be only numerical data as variables. But in the dataset there is a categorical variable ‘Country’. So, values in this variable has to be encoded to use in a linear regression model. LabelEncoder() and Mapping methods are used to do this task. Country column’s data passed into these methods to encode these data.



Figure 9 : Label encoding

Following is the mapping of countries( 195 unique) to the numeric values :

Text, letter

Description automatically generated*Figure 10 : Mapping of string datatype to numeric datatype*

After mapping, the values in the country column will be changed to their respective numeric values. Following figure shows that :

A picture containing table

Description automatically generated

*Figure 11 : Dataset after mapping*

1. **APPLICATION OF THE APPROPRIATE LEARNING ALGORITHM**

**3.1 INTRODUCTION AND BACKROUND OF THE ALGORITHM**

Linear Regression is a supervised Machine learning algorithm which is used for regression problems (Numerical). This is used to predict the relationship between independent and dependent variables. Multiple linear regression means, linear regression with multiple variables. Some of the applications for multiple linear regression are prediction of economic growth of a country, prediction of product prices with the time, estimation of housing sales etc.

**3.2 WHY LINEAR REGRESSION?**

In this problem the HDI is based on the four main dimensions. Therefore Human development index will depend on Life expectancy, Mean years of schooling, Gross National Income (GNI) per Capita, Internet users. Increasing and decreasing of mentioned four dimensions are continuously cause for the value of the Human Development Index. We are used the Linear Regression algorithm for this type of problems. Therefore we have to use Linear Regression algorithm to go through this.

* 1. **HYPOTHESIS**

**Text

Description automatically generated**

*Equation 1: Hypothesis of MLR*

Actually there are no relationship between X variables and the Y variables. In here we can use this hypothesis for the linear regression algorithm.

* 1. **COST FUNCTION**

**Text

Description automatically generated**

*Equation 2 : Cost function of MLR*

There are more than one independent variable, we used below equation as the cost function of the algorithm. If there has more than one independent variable we have to select multiple linear regression to solve the problem. This multiple linear regression algorithm generalized version of the linear regression.

* 1. **GRADIENT DESCENT**

**Text

Description automatically generated**

*Equation 3: Gradient descent of MLR*

Gradient Descent is the process of minimizing a function by following the gradient of the cost function.

* 1. **APPLICATION OF THE ALGORITHM**

In this dataset, we can identify four main independent variables and one dependent variables. Independent variables are Life expectancy, Mean years schooling, Gross National Income (GNI) per capita (PPP US$), Internet users. Dependent variable is the Human Development Index. Here dependent variables are denoted by X and independent variable is denoted by y. This dataset is split into two parts, training dataset and test dataset. 70% of the data is divided as training data and 30% of the data is divided into test data. Training dataset is used to train the machine learning model and test dataset is used to test and evaluate the accuracy of trained model. An instance of LinearRegression() class is called as the model to train data. X\_train and y\_train parameters are passed into this model with the help of fit() method.

**4. RESULTS**

**Chart, histogram

Description automatically generated**

*figure 12: distribution of human development index*

Chart

Description automatically generated

*Figure 13: Correlation matrix*

After training the Linear Regression model can evaluate the accuracy of the model with test data. To do that predict () method will be used and X\_test is passed as a parameter.

Chart, scatter chart

Description automatically generated

*Figure 14 : predicted y versus Testing y*

Calculating the R squared value of y\_test and y\_pred.

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.



*Figure 15 : Accuracy*

The R squared value: 0.9540989333484

Finally ‘0. 9540989333484’ is the R squared value in this prediction. Referring this value we can get idea about used algorithm. Finally got more accurate R squared value by using the Linear Regression algorithm.

**4.1 OUTPUT SCREENS**

A screenshot of a computer

Description automatically generated with medium confidence

*Figure 16 : output screen – 1*

By choosing the predict button results in the following output screen – 2:

**A screenshot of a computer

Description automatically generated**

*Figure 17 : output screen-2*

By choosing all the values and clicking on the predict button results in the following final output screen:

*Graphical user interface

Description automatically generated*

*Figure 18 : final output*

**5.CONCLUSION**

In conclusion, we proposed a machine learning model to predict the Human Development Index of the country using the machine learning tools. This system can be used to predict the index of human development of a country and also know into which category it falls into like low, very low, medium, high and very high. From the result, when tested among different attributes like life expectancy, mean years of schooling, GNI per capita and etc in dataset we got accuracy of 95%.

**6. APPENDIX**

**Home.html**

<!DOCTYPE html>

<html>

<head>

<title>Home</title>

<style>

.navbar

{

margin: 0px;

padding:20px;

background-color:;

opacity:0.6;

color:black;

font-family:'Roboto',sans-serif;

font-style: italic;

border-radius:20px;

font-size:25px;

}

a

{

color:grey;

float:right;

text-decoration:none;

font-style:normal;

padding-right:20px;

}

a:hover{

background-color:black;

color:white;

border-radius:15px;0

font-size:30px;

padding-left:10px;

}

p

{

color:white;

font-style:italic;

font-size:30px;

}

body

{

background: linear-gradient(rgba(0,0,0,0.8),rgba(0,0,0,0.8)),url("https://upload.wikimedia.org/wikipedia/commons/thumb/b/b4/2020\_Inequality-Adjusted\_Human\_Development\_Index\_Map.png/430px-2020\_Inequality-Adjusted\_Human\_Development\_Index\_Map.png");

height: 125vh;

-webkit-background-size: cover;

background-size:cover;

background-position: center center;

position: relative;

}

</style>

</head>

<body>

<div class="navbar">

<a href="/Prediction" >Predict</a>

<a href="/Home">Home</a>

<br>

</div>

<br>

<center><b><font color="white" size="15" font-family="Comic Sans MS" >Human Development Index</font></b></center>

<div>

<br>

<center>

<p>The Human Development Index (HDI) is a statistic composite index of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers (very high, high, medium & low) of human development. A country scores a higher HDI when the lifespan is higher, the education level is higher, and the gross national income GNI (PPP) per capita is higher. The HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The HDI can also be used to question national policy choices, asking how two countries with the same level of GNI per capita can end up with different human development outcomes. These contrasts can stimulate debate about government policy priorities. In order to do this we will be building a machine learning model to predict the Human DeveIopment Index of a country by taking few important aspects as inputs. Our model will at last predict the HDI score of a country and will also tell under which category will it be falling (very high, high, medium or low)

</p>

</center>

</div>

</body>

</html>

**Indexnew.html**

<html>

<style>

div.header1{

top:20;

position: fixed;

padding-left: 490px;

}

\*{

margin:0;

padding:0;

border:0;

outline:0;

text-decoration:none;

font-family:montserrat;

}

body

{

background-image:url('https://image.freepik.com/free-vector/binary-world-map\_46706-715.jpg');

background-position: center;

font-family:sans-serif;

background-size:cover;

margin-top:40px;}

.main{

background-color:rgb(0,0,0,0.6);

width:800px;

height:590px;

margin:auto;

position:center;

border-top-left-radius:100px;

border-bottom-right-radius:100px;

}

.main input[type="text"],.main input[type="text"],.main input[type="text"],.main input[type="text"],.main input[type="text"],.main input[type="text"],.main input[type="text"]{

border:0;

background:none;

display:block;

margin:20px auto;

text-align:center;

border:2px solid #3498db;

padding:10px 3px;

width:280px;

outline:none;

color:white;

border-radius:24px;

transition:0.25s;}

.bor{

border:0;

background:none;

display:block;

margin:20px auto;

text-align:center;

border:2px solid #8e44ad;

padding:10px 3px;

width:500px;

outline:none;

color:white;

transition:0.25s;}

.main input[type="text"]:focus,.main input[type="text"]:focus,.main input[type="text"]:focus,.main input[type="text"]:focus,.main input[type="text"]:focus,.main input[type="text"]:focus,.main input[type="text"]:focus{

width:280px;

border-color:#8e44ad;

}

.logbtn{

display:block;

width:35%;

height:50px;

border:none;

border-radius:24px;

background:linear-gradient(120deg,#3498db,#8e44ad,#3498db,#8e44ad);

background-size:200%;

color:#fff;

outline:none;

cursor:pointer;

transition:.5s;

font-size:25;

}

.logbtn:hover{

background-position:right;

}

input::placeholder{

color:#F5FFFA;

}

.bottom-text{

margin-top:60px;

text-align:center;

font-size:13px;

}

</style>

<body>

<center><div class="header1"><font color="#FF0000" font-family="Fascinate Inline" size=7 ><b>Human Development Index </b></font></div></center>

<br><br><br><br><br>

<form class="main" action="/predict" method="post">

<br>

<center><select id="Country" name="Country">

<option value="">Select the name of the Country</option>

<option value="0">Afghanistan</option>

<option value="8">Australia</option>

<option value="13">Bangladesh</option>

<option value="31">Canada</option>

<option value="76">India</option>

<option value="138">Poland</option>

<option value="179">Turkey</option>

</select></center>

<input class="form-input" type="text" name='Life expectancy' placeholder="Enter life expectancy rate(range 50-89)"><br>

<input class="form-input" type="text" name='Mean years of schooling' placeholder="Enter Mean years of schooling(range 1-15)"><br>

<input class="form-input" type="text" name='Gross national income (GNI) per capita' placeholder="Enter GNI(range 290-129000)"><br>

<input class="form-input" type="text" name='Internet users' placeholder="Enter the number of internet users"><br>

<center><input type="submit" class="logbtn" value="Predict"></center>

<div class="bor"><center><b><font color="white" size=5>{{showcase}}</font></b></center></div>

</form>

</body>

</html>

**Resultnew.html**

<html>

<style>

.idiv{

width:60%;

margin:auto;

background-color:black;

text-align:center;

margin-top:2%;

border-radius:10px;

background-image:url("");

background-repeat: no-repeat;

margin-top:2%;

}

body{

background-color:black;

font-family:segoe ui;

background: linear-gradient(rgba(0,0,0,0.8),rgba(0,0,0,0.8)),url(https://www.nationsonline.org/gallery/World/Human-development.jpg);

height: 100vh;

-webkit-background-size: cover;

background-size: cover;

position: relative;

}

input{

font-size:1.3em;

width:80%;

text-align:center;

}

input placeholder{

text-align:center;

}

button{

outline:0;

border:0;

background-color:darkred;

color:white;

width:100px;

height:40px;

}

button:hover{

background-color:brown;

border:solid 1px black;

}

h1{

color:red;

}

h2{

color:olive;

font-size:16;

}

h3{

color:olive;

font-size:22;

}

</style>

<head>

<title>-- Human Development Index Score -- </title>

</head>

<body>

<div class='idiv'>

<br/>

<h1>Human Development Index</h1>

<h2><i>A machine learning web application using flask</i><h2>

<br/>

<h3>{{prediction\_text}} </h3>

<br/>

<br/>

<br/>

</div>

</body>

</html>

**app.py**

import numpy as np

import pandas as pd

from flask import Flask, render\_template, request

import pickle

app = Flask(\_\_name\_\_)

model = pickle.load(open('HDI.pkl','rb'))

@app.route('/')

def home():

return render\_template('home.html')

@app.route('/Prediction',methods=['POST','GET'])

def prediction():

return render\_template('indexnew.html')

@app.route('/Home',methods=['POST','GET'])

def my\_home():

return render\_template('home.html')

@app.route('/predict',methods=['POST','GET'])

def predict():

input\_features = [float(x) for x in request.form.values()]

features\_value = [np.array(input\_features)]

features\_name = ['Country','Life expectancy','Mean years of schooling','Gross national income (GNI) per capita','Internet Users']

df = pd.DataFrame(features\_value, columns=features\_name)

#predictions using the loaded model file

output = model.predict(df)

print(round(output[0][0],2))

print(type(output))

y\_pred = round(output[0][0],2)

if(y\_pred >= 0.3 and y\_pred <= 0.4) :

return render\_template("resultnew.html",prediction\_text = 'Low HDI' + str(y\_pred))

elif(y\_pred >= 0.4 and y\_pred <= 0.7):

return render\_template("resultnew.html",prediction\_text = 'Medium HDI' + str(y\_pred))

elif(y\_pred >= 0.7 and y\_pred <= 0.8):

return render\_template("resultnew.html",prediction\_text = 'High HDI' + str(y\_pred))

elif(y\_pred >= 0.8 and y\_pred <= 0.94):

return render\_template("resultnew.html",prediction\_text = 'Very High HDI' + str(y\_pred))

else :

return render\_template("resultnew.html",prediction\_text = 'The given values do not match the range of values of the model. Try giving the values in the mentioned range'+str(y\_pred))

return render\_template('result.html', prediction\_text=output)

if \_\_name\_\_ == '\_\_main\_\_':

app.run(debug=True,port=5000)

**7. REFERENCES**

[1] "Human Development Index HDI," 26 March 2009. [Online]. Available: http://wikiprogress.org/articles/initiatives/human-development-index/.

[2] "Human Developments Reports," UNITED NATIONS DEVELOPMENT PROGRAMME, [Online]. Available: http://hdr.undp.org/en/data.

[3] "Kaggle," 25 January 2017. [Online]. Available: https://www.kaggle.com/undp/humandevelopment#human\_development.csv .