Solution Title: OnSight Helper

(Early Detection of Tuberculosis in Tanzania)

Table of Contents

[Introduction 2](#_Toc54630621)

[Documentation team 2](#_Toc54630622)

[Solution Owners 2](#_Toc54630623)

[Problem Definition 3](#_Toc54630624)

[Background of the problem 3](#_Toc54630625)

[Prevalence of TB in Tanzania 3](#_Toc54630626)

[What Problem is the solution trying to solve? 4](#_Toc54630627)

[Who are those affected by the problem? 4](#_Toc54630628)

[Solution 5](#_Toc54630629)

[What is this solution about? 5](#_Toc54630630)

[Who Is the targeted user of the solution? 5](#_Toc54630631)

[When was the solution first released? 5](#_Toc54630632)

[When was the solution Lastly released? 5](#_Toc54630633)

[What is the output of the solution? 5](#_Toc54630634)

[What is the expertise types required to develop this solution? 6](#_Toc54630635)

[Have you updated this solution documentation before? 7](#_Toc54630636)

[USAGE 7](#_Toc54630637)

[What is the intended use of the solution output? 7](#_Toc54630638)

[Who are the target users? 7](#_Toc54630639)

[What are the key procedures followed while using the solution? 7](#_Toc54630640)

[DOMAIN AND APPLICATIONS 8](#_Toc54630641)

[DATASET 8](#_Toc54630642)

[Motivation 8](#_Toc54630643)

[Composition 8](#_Toc54630644)

[Collection Process 8](#_Toc54630645)

[Preprocessing 8](#_Toc54630646)

[Uses and maintenance 8](#_Toc54630647)

[Did you use synthetic data? 8](#_Toc54630648)

[Model 9](#_Toc54630649)

[A clear description of the model 9](#_Toc54630650)

[A clear description of the data preparation for the model 9](#_Toc54630651)

[What is the assumption that we made? 10](#_Toc54630652)

[An analysis of the complexity (time, space, sample size) of algorithm used? 10](#_Toc54630653)

[When were the models last updated? 10](#_Toc54630654)

[Evaluation 11](#_Toc54630655)

[Describe the testing methodology 11](#_Toc54630656)

[What is the Performance metrics of the model? 11](#_Toc54630657)

[Describe the test results. 11](#_Toc54630658)

[TESTING BY THIRD PARTIES 12](#_Toc54630659)

[Is there a way to verify the performance metrics? 12](#_Toc54630660)

[In addition to the solution provider, was this service tested by any third party? 12](#_Toc54630661)

[Results 13](#_Toc54630662)

[A clear description of the model result. 13](#_Toc54630663)

[The exact number of training and evaluation runs 13](#_Toc54630664)

[A clear definition of the specific measure or statistics used to report results 14](#_Toc54630665)

[The average runtime for each result, or estimated energy cost 14](#_Toc54630666)

[Environment 15](#_Toc54630667)

[A description of the computing infrastructure used. 15](#_Toc54630668)

[Is the solution deployed? 15](#_Toc54630669)

[STEPS TO REPRODUCE THE SOLUTION 16](#_Toc54630670)

[SAFETY 17](#_Toc54630671)

[Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the solution? 17](#_Toc54630672)

[Do you use data from or make inferences about individuals or groups of individuals? Have you obtained their consent? 17](#_Toc54630673)

[Fairness 17](#_Toc54630674)

[Does the solution implement and perform any bias detection and remediation? 17](#_Toc54630675)

[Concept Drift 17](#_Toc54630676)

[What is the expected performance on unseen data or data with different distributions? 17](#_Toc54630677)

[Does your solution make updates to its behavior based on newly ingested data? 18](#_Toc54630678)

[SECURITY 19](#_Toc54630679)

[How could this solution be attacked or abused? 19](#_Toc54630680)

[Scenario for which the solution is not suitable 19](#_Toc54630681)

[How are you securing user or usage data? 19](#_Toc54630682)

[Was the solution checked for robustness against adversarial attacks? 19](#_Toc54630683)

[Links 20](#_Toc54630684)

# Introduction

## Documentation team

|  |  |  |
| --- | --- | --- |
| S/no | NAME | Institution |
| 1 | Khadija Mzenzi | University of Dodoma |
| 2 | EMMACULATA W. MAGERE | University of Dar es Salaam |
| 3 | Simon Machera (Team Lead) | University of Dar es Salaam |
| 4 | Laurent Deusdedith | University of Dar es Salaam |

## Solution Owners

|  |  |  |
| --- | --- | --- |
| S/no | NAME | Institution |
| 1 | Khadija Mzenzi | Univerdity of Dodoma |
| 2 | EMMACULATA W. MAGERE | University of Dar es Salaam |
| 3 | Simon Machera (Team Lead) | University of Dar es Salaam |
| 4 | Laurent Deusdedith | University of Dar es Salaam |

# Problem Definition

## Background of the problem

Tanzania is among the 30 countries in the world reported to have the highest TB prevalence. According to the 2018 Global report on quality improvement, Tanzania was estimated to have 154,000 people with TB while only 68,273 (44%) were initiated on treatment. Fortunately, of those who had initiated treatment, 90% were successfully treated. Furthermore, while the country has an estimated 1,700 people living with multidrug resistant TB or rifampicin-resistant TB (MDR-TB or RR-TB), only 167 people started MDR-TB/RR-TB treatment. These statistics represent significant gaps in both the drug susceptible TB and MDR-TB treatment cascades (Global Fund Report Tanzania, 2018.).

One significant challenge facing Tanzania and thus leading to the above statistics is misdiagnosis or late diagnosis of tuberculosis, which lead to three major problems; poor treatment outcomes, higher tuberculosis prevalence, and possible multidrug resistant TB. If patients delay seeking healthcare services or they are “lost to diagnostic follow up” (LDFU), they are at a substantially increased risk of tuberculosis morbidity and mortality

## Prevalence of TB in Tanzania

Tanzania is among the countries with the highest TB burden in the world, with approximately 295 TB cases per 100,000 adults. Studies show that for patients who seek healthcare, nearly half of them seek care more than twice before being diagnosed TB. This indicates that a high proportion of TB patients are seen by several different health care providers before receiving a diagnosis, suggesting a low suspicion index among health care providers. As a result, the prevalence of multi-drug resistant tuberculosis (MDRTB) is high in Tanzania

In 2018 a total of 449 RR/MDRTB cases were notified country wide among which 409 (91%) were started on MDR TB treatment. Dar es Salaam region contributed the largest proportion of RR/MDR TB cases detected and enrolled on treatment (28%) followed by Mwanza (11%), Geita (6%), Pwani and Mara (5%) each, Morogoro, Mbeya and Lindi each contributed (4%), Arusha and Mtwara (3%) each, other regions had 27% in total except Katavi which did not detect any case

## What Problem is the solution trying to solve?

The solution is trying to solve the problem of unexperienced healthcare providers that are likely to miss tuberculosis at the first or second visit of a patient. The solutions presented will greatly support healthcare providers by enhancing their ability to identify tuberculosis at first points of care. The solution will help them rise TB suspicion index when diagnosing patients

## Who are those affected by the problem?

Groups that are abnormally affected by these issues are individuals living in urban areas, particularly miners who live in mining towns (miners have a high risk of developing tuberculosis). Additionally, people living with HIV are at high risk for developing TB and are likely to be misdiagnosed.

In 2018 a total of 449 RR/MDRTB cases were notified country wide among which 409 (91%) were started on MDR TB treatment. Dar es Salaam region contributed the largest proportion of RR/MDR TB cases detected and enrolled on treatment (28%) followed by Mwanza (11%), Geita (6%), Pwani and Mara (5%) each, Morogoro, Mbeya and Lindi each contributed (4%), Arusha and Mtwara (3%) each, other regions had 27% in total except Katavi which did not detect any case

# Solution

## What is this solution about?

The solution (OnSight Helper) is a mobile application-based Machine Learning system that provide decision support and symptom assessment to better identify tuberculosis.

This solution will help to rise the TB suspicion index of the healthcare provider when diagnosing patients

## Who Is the targeted user of the solution?

The targeted users are the healthcare providers, especially those dealing with diagnostics of TB

## When was the solution first released?

The solution was first released on 18th October 2020. The first released version was only returning the status of the patient, either Positive or Negative TB

## When was the solution Lastly released?

The solution was last released on 25th October 2020. The last released version returns the probability in percentage of the patient being positive TB

Link to the solution!

Link!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

## What is the output of the solution?

The output of the solution is the probability of the patient being Positive TB (expressed in percentage). Based on the percentage, the healthcare provider may suggest further diagnostics. If for instance the solution shows that a percent is 90% Positive TB, the healthcare provider may recommend further confirmatory tests.

## What is the expertise types required to develop this solution?

The following table will elaborate the type of expert required to develop the solution

|  |  |  |
| --- | --- | --- |
| **Expert Required** | **Role** | **Required skilled** |
| TB expert | To highlight the main features of TB patients | Proficient in TB disease theory |
| Data scientist | To use the features highlighted by the TB expert to generate synthetic but accurate data  To use the synthetic data to train a model | Statistics, probability and distribution theory knowledge  Know how to use python tools for Exploratory data analysis(EDA)  Python knowledge |
| Backed Developer | To create an API for the model | Flask framework  Heroku  Github |
| Mobile App developer | To develop the mobile app that will act as an interface between the doctor and the model | Flutter Framework  Sqlite database |
| Research team | To extract information from TB experts and structure them in well order | Communication skills  Basic Statistics theory |

## Have you updated this solution documentation before?

Yes, we have updated the solution documentation. The initial documentation was about the solution version 1 that was based on returning only the status of the patient. But the current Documentation Is based on the solution that return probability of the patient having TB within the range of 0 to 1 ( it is represented as percentage in the mobile app)

# Usage

## What is the intended use of the solution output?

The solution’s output is intended to raise the TB suspicion index of healthcare providers when conducting TB diagnosis to patients. The intention is to eliminate the chances of missing TB patient at first sight

## Who are the target users?

The targeted users are the healthcare providers when diagnosing patients with a like hood of being with TB infections

## What are the key procedures followed while using the solution?

The following are the key procedures when using the solution

Step1: The healthcare provider will login in his/her account in the OnSight Helper Mobile application

Step2: On the Homescreen, select Questionnaire

Step3: under the consent page, the patient will have to accept the declaration of allowing the system app to collect personal data and use the data in prediction of the health status plus storing the data for future use.

Step4: The questions will display in a good order and the doctor will have to ask the patient to answer the question. The doctor will fill all the questions in the questionnaire

Step5: Finally, after filling all the questions, the doctor will submit the question. The answers will be sent to the model for prediction and the results will be displayed in the mobile app inform of percentage of being Positive TB. Furthermore, the patient data will be saved in Diagnostic reports activity for any future use.

# DOMAIN AND APPLICATIONS

The domains that the application is to be tested on are health centers, Dispensaries and Hospitals.

But due to time limitation, We were not able to directly test the system to its areas of application

# DATASET

## Motivation

Due to lack of public dataset that qualify our expectation, we decided to use a data generative model to generate synthetic data with heavy influence from field experts (TB field experts)

## Composition

The dataset is composed of signs, symptoms, risk factors, diseases & infections and habits of patients associated with TB disease plus their corresponding TB status.

## Collection Process

The information that we used to generate the dataset was collected from TB experts and online literatures and magazines that concern TB in Tanzania.

## Preprocessing

After analyzing the information, we selected important variables, and these variables were fed in the data generative model to develop for us 10,000 patients with different characteristics and TB status

## Uses and maintenance

The synthetic dataset that we created was used to train a prediction machine learning model of our solution

## Did you use synthetic data?

Yes, our model was developed based on synthetic patient data. We prepared a data generative model in python to generate for us 10,000 patients by giving it the features of TB patients. The generative models use statistics and distribution theory to create the dataset

Here is the link to the data generative model:

<https://drive.google.com/drive/folders/1E38UYVUSwf3HRCMfrAmL-10PnoSI1M7i?usp=sharing>

# Model

## A clear description of the model

Date created: 19th October 2020

Model Version: Version 2.0

Model type: Joblib

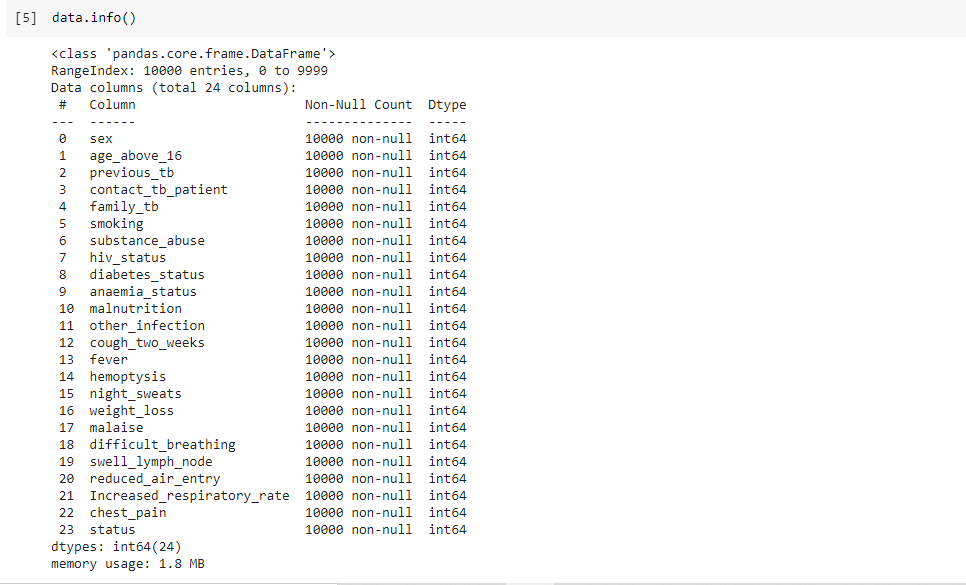
Training Algorithm: RandomForestClassifier

|  |  |
| --- | --- |
| Algorithm | RandomForestClassifier |
| n\_estiamtors | 20 |
| Random\_state | 0 |

## A clear description of the data preparation for the model

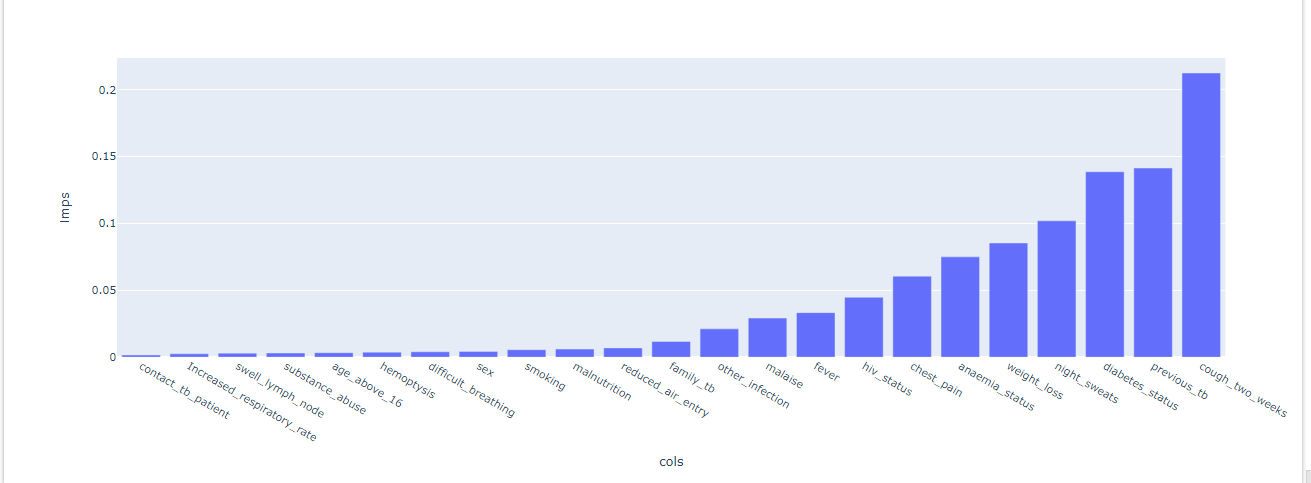
Google Colab and JupyterNotebook were used in preparing the data for the model.

Below is a brief information of the dataset



It is seen that the dataset does not contain any null value

Assessing the feature Importance of the to the models, the graph below shows the feature importances



## What is the assumption that we made?

The assumption we made is that the data that we have generated and used it to train our model will work almost similar to the data that we could have obtained directly from health centers

## An analysis of the complexity (time, space, sample size) of algorithm used?

To train the model, I used google colab

Time to train: 5 to 10 secs

Sample size: The model has 364KB

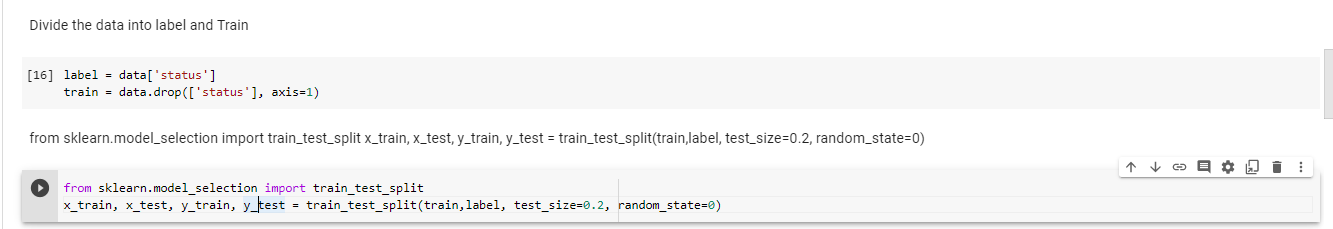
## When were the models last updated?

The model was lastly update on 19th October 2020 to provide the output as range of probabilities between 0 and 1

# Evaluation

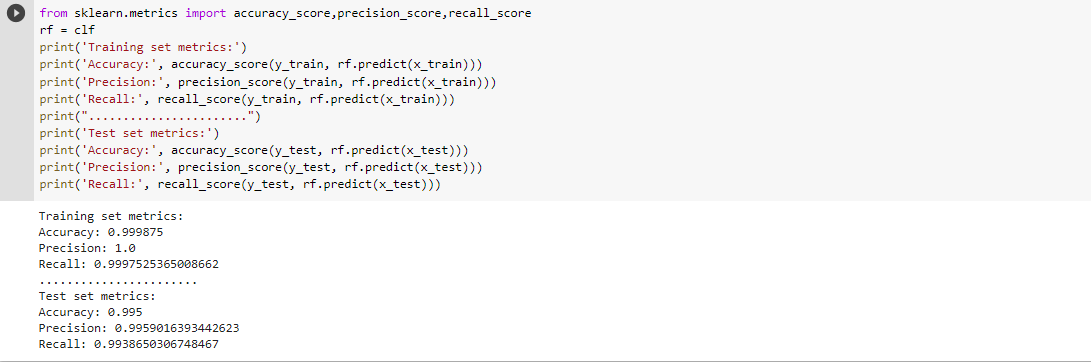
## Describe the testing methodology

The dataset was plit into training data set and testing dataset using the train\_test\_split from sklearn.model\_selection



## What is the Performance metrics of the model?

Since we used synthetic data, the data was not biased (the ratio of positive TB patient to no TB was 50/50) but our model was overfitting. But it’s not too bad in our scenario since it works pretty fine. Below diagram shows the metrics on training



## Describe the test results.

To test the model, we created an ideal patient (patient\_a), passed in the variables and gave it to the model for prediction. Based on the variables we entered, the model was 35% sure that the person had TB

See the Diagram below



# TESTING BY THIRD PARTIES

## Is there a way to verify the performance metrics?

Yes, here is the link to where the model is located

<https://drive.google.com/drive/folders/1E38UYVUSwf3HRCMfrAmL-10PnoSI1M7i?usp=sharing>

To use it, you have to import it to python environment like in jupyter notebook or vscode and then you can proceed to testing the metrics

## In addition to the solution provider, was this service tested by any third party?

No, till the date of documentation, the solution was only tested by solution providers

# Results

## A clear description of the model result.

The result of the model is a probability in a range between 0 and 1. This probability represents the percentage that the module is sure the person is TB positive

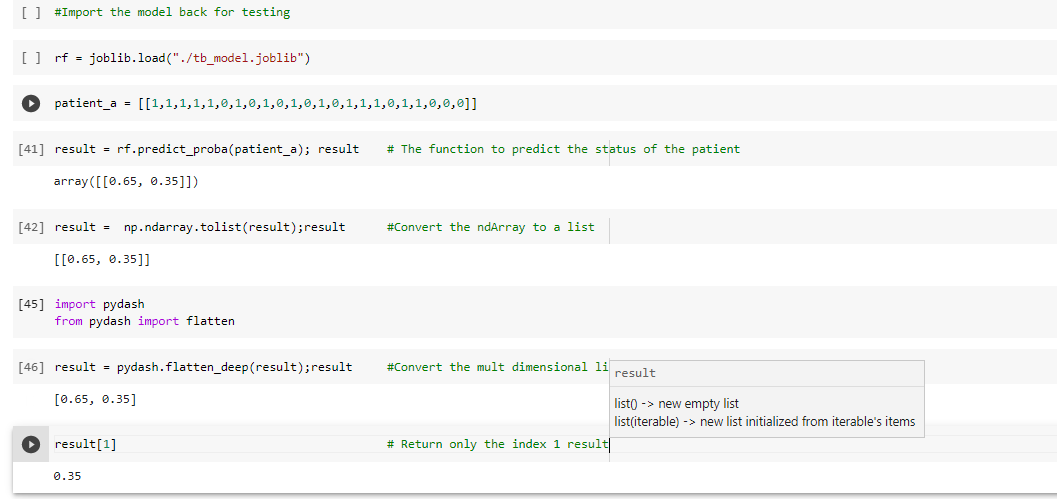
This is how we get the result:

-The predict\_proba() function will return a ndArray of format [[x,y]] x is the probabily of No TB, and Y is the probability of positive TB.

-Convert the ndArray to multi – dimensional list

-convert the mult – dimensional list to one dimension

-Take only th index 1 result (For our case this is the probability of the person being TB positive)



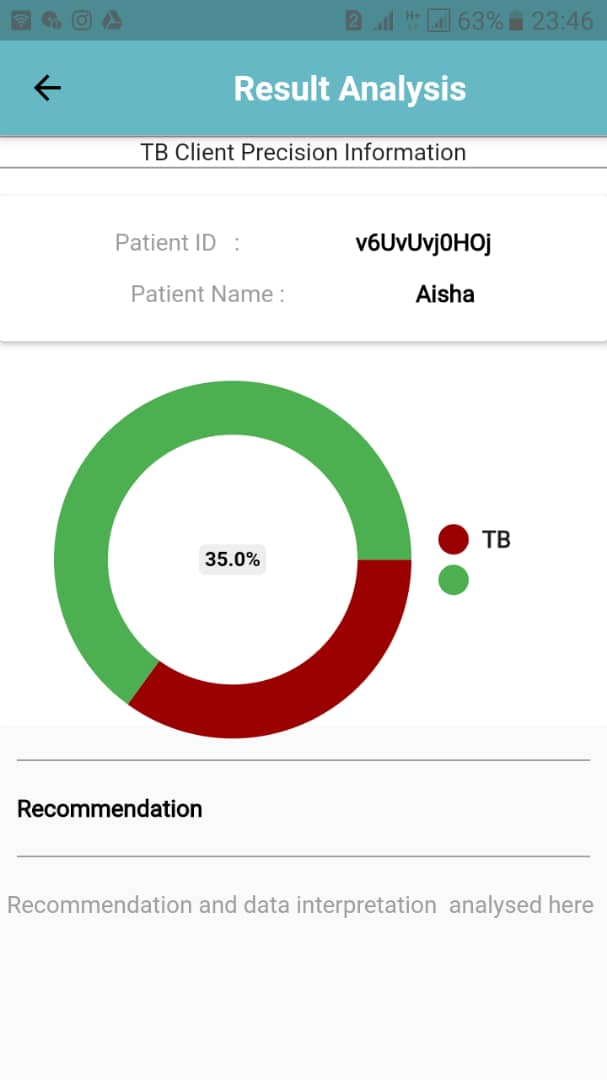
## The exact number of training and evaluation runs

Number of training: 4

Number of Evaluations run: 7

A clear definition of the specific measure or statistics used to report results.

The result will appear as a graph showing the percentage of patient being TB positive. A sample result is as below



## The average runtime for each result, or estimated energy cost

Since the model is hosted online, the runtime will vary depending on the internet speed. But for an area with good internet speed, the average run time is 4 to 10 seconds

# Environment

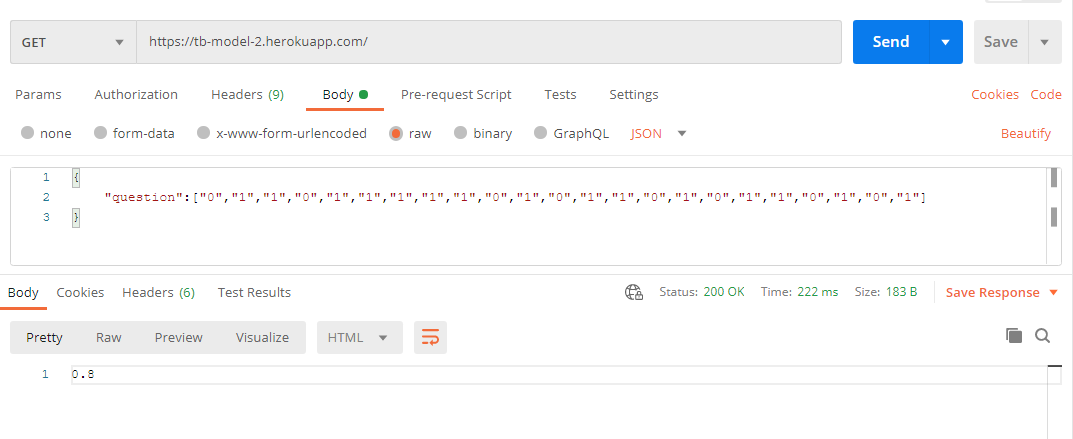
## A description of the computing infrastructure used.

|  |  |
| --- | --- |
| Operating system used | Windows 10 professional |
| Programming language | Python 3.6.4  Flutter 1.22.2 (mobile app) |

## Is the solution deployed?

Yes, our solution is deployed to heroku. Here is the link to the solution <https://tb-model-2.herokuapp.com/>

The solution can be tested in Postman by sending a json array of variables as in the below diagram



The required number of features to be put when testing is 23

# STEPS TO REPRODUCE THE SOLUTION

The following steps were taken when developing the solution

Step1: Creating a data generative model to create the dataset of 10,000 patients.

Here is the link to the Generative model: <https://drive.google.com/drive/folders/1E38UYVUSwf3HRCMfrAmL-10PnoSI1M7i?usp=sharing>

Step2: Writing a python script to train the model using the dataset from step 1

Here Is the link to the python script:

<https://drive.google.com/drive/folders/1DSH-I8n6BN6s6sq9O2kUVn92BkRA7vtP?usp=sharing>

Step3: Write a flask application server for the model

Here is the code for flask server

<https://drive.google.com/drive/folders/1LkRiw0yfwousXzW8Gx21SOhl5OUiqLyT?usp=sharing>

Step4: Hosting the Flask server to heroku.

Here is the link to the heroku API :

<https://tb-model-2.herokuapp.com/>

Step5: Developing a mobile application that will be used by healthcare providers and link it with the heroku API from step4

Here is the link to mobile app:

<https://github.com/CymonMachera/TB_ml_hackathon.git>

# SAFETY

## Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the solution?

Yes, this solution is expected to be used by healthcare providers. It is greatly discouraged to be used directly by patients or other nonprofessional people without the guidance of healthcare providers. Directly usage of the solution by patient may mislead them. The solution is not a confirmatory test of TB, rather its main focus is to raise the suspicious index of healthcare provider

Another great risk of this solution is using it as a confirmatory tool. It is to our expectation that healthcare providers will go beyond the results provided by the solution, either by suggesting further diagnosis to the patients.

To minimize the risk, our solution will be deployed as a mobile app, and only authorized institutions and people will be able to use our solution

## Do you use data from or make inferences about individuals or groups of individuals? Have you obtained their consent?

Our solution has used synthetic data that we have generated using the information obtained from

TB experts and other online literatures

# Fairness

## Does the solution implement and perform any bias detection and remediation?

No, it doesn’t

# Concept Drift

## What is the expected performance on unseen data or data with different distributions?

Since or model takes in only fixed variables, if the distribution changes then retraining the model to fit the new distribution is inevitable

## Does your solution make updates to its behavior based on newly ingested data?

Currently our solution doesn’t make any updates based on newly ingested data, but it is a feature to be added in the future.

One of the functionalities of our solution is to collect new patient history data, so in future these data can be used to retrain the model.

## How is the solution tested and monitored for model or performance drift over time?

Our solution is able to store the prediction status of patients. So, after doing the confirmatory tests, the results may be compared and assessment may be done as to how much our model is precise in prediction

## Does the solution allow for checking for differences between training and usage data?

No, it doesn’t. In order to see the training data, we must give you direct access to them

## Do you test the solution periodically?

Yes, the performance of our solution is tested periodically with different patient data

# SECURITY

## How could this solution be attacked or abused?

The solution could be attacked or abused when it is used by individuals who are not healthcare providers. This solution is built specifically for healthcare providers and should only be used for the purpose of diagnosis supports.

## Scenario for which the solution is not suitable

The solution is not suitable for any non-medical activities

## How are you securing user or usage data?

The usage data are stored in the database but with the consent of the patient. The data will be used to create a TB dataset that can be open sourced for use

## Was the solution checked for robustness against adversarial attacks?

Yes, our solution is hosted on a very secure platform (heroku) and accessed via a mobile app

# Links

Link to training and evaluation code

<https://drive.google.com/drive/folders/1DSH-I8n6BN6s6sq9O2kUVn92BkRA7vtP?usp=sharing>

Link to pre-trained model(s)

<https://drive.google.com/drive/folders/1LkRiw0yfwousXzW8Gx21SOhl5OUiqLyT?usp=sharing>

Link to deployment code

<https://drive.google.com/drive/folders/1LkRiw0yfwousXzW8Gx21SOhl5OUiqLyT?usp=sharing>

Link to dataset(s)

<https://drive.google.com/drive/folders/1E38UYVUSwf3HRCMfrAmL-10PnoSI1M7i?usp=sharing>