**Interim Report**

**Disease Prediction**

Keyurkumar Kankotiya (0781500)

1. **Executive summary**

Currently, humans face many health conditions due to the current environmental condition and their living habits. The identification of similar conditions at their earlier stages is important, to help the extremity of it. As per the current situation, it is very hard to get instant appointments in Windsor and it is possible that it can expand to a critical level, so the goal of this project is to identify and predict the patients with appropriate precautions in the meantime. This could be achieved by using the machine learning method to ensure that this categorization reliably identifies persons with conditions. The examination of conditions is also a difficult task. Hence, data mining plays a critical part in disease prediction. This project proposes disease identification grounded on the case’s symptoms using machine learning algorithms like convolutional neural network (CNN) and K- nearest neighbor( KNN).

The project for disease analysis is a research project that seeks to identify and analyze patterns in disease data. The project will involve collecting and analyzing data from various sources, such as medical records, public health databases, and surveys. The data will be used to identify trends in disease prevalence, risk factors, and other factors that may influence the spread of disease. The project will also seek to develop predictive models to help predict the spread of disease and inform public health interventions.

The main objective of our project is to make a platform where everyone, especially students, can check on their health by putting their symptoms into the chatbot and getting the right disease information, so they can take further precautions and treatment if needed.

Students are our focus as more than 7 thousand students come to Windsor for studies every year, and they are new to the climate and conditions, which will help them with their health issues.

1. **A progress**

To accomplish this project the main task is to create an ML model with good accuracy.

**COMPLETED TASKS:**

***Find dataset:***

Firstly, we have a good time finding a perfect dataset that has all details, symptoms, precautions, and disease details, and we found the dataset for which we were looking. To assure the data's integrity and accuracy, it has been cleansed and preprocessed.

***Data preparation:***

After getting the dataset, the next step was to prepare the data for further analysis. Visualization was done after the data was cleaned.

***Feature Engineering:***

To extract significant characteristics from the acquired data, the team used substantial feature engineering. This entails identifying the most useful variables, converting, and normalizing data, and developing new features with predictive value.

***Model Development:***

On the acquired dataset, many machine learning models, including decision trees, and K-Nearest Neighbours, were trained and assessed. Model performance has been evaluated using a variety of criteria, including accuracy, precision, recall, and F1-score. The team also performed model interpretability analysis to better understand the elements that influence the model's predictions.

***Model Selection:***

The team chose the top-performing model for further improvement and deployment after extensive examination and comparison. The chosen model has shown promising prediction performance and the potential for real-world applications.

***Create UI:***

As we are creating a chatbot no doubt we’ll need a good user interface for communication. We started working on our UI, and for that, we used the best tool which is known as FIGMA.

**WORK CURRENLTY UNDERWAY:**

***Model Refinement:***

The chosen model is being fine-tuned and optimized to increase its predicted performance even more. This involves optimizing feature selection, tweaking hyperparameters, and correcting any overfitting concerns.

***Validation and Testing:***

To investigate generalizability and robustness, the improved model is being verified on an independent test set. This entails testing the model's performance using previously unknown data to confirm its dependability in real-world circumstances.

**UPCOMING TASKS:**

***Deployment and Integration:***

Following refinement and validation, the model will be implemented in a safe and scalable environment, evaluated in a real-world clinical context.

***Evaluation and Performance Monitoring:***

The team will continually monitor and assess the model's prediction accuracy, dependability, and usefulness in the production environment. This will assist in identifying any flaws or restrictions and will drive future developments.

1. **Key progress**

We know that this is not a small project, and we are working in a group and that’s why it makes it easy for us to work smoothly on this project.

I worked on a few tasks like creating a model and creating UI for the project. For the model, we used Python language and Jupiter Notebook as an editor. For me, it was a hard task to rearrange the data so it will be used to create a model. I spent almost a week rearranging the data.

The second difficult work was to create a UI. I never worked on UI, so it was hard for me to adapt to a new tool and work on it. I watched a few YouTube videos and whenever I found it difficult, I took help from Google, and I tried to create a UI in a way that I can make changes easily if needed.

1. **Data and Metrics**

***Dataset:***

A complete dataset was compiled, and it included records of 41 disease with varied symptoms, precautions, descriptions, and severity.

***Data Quality*:**

Data cleaning techniques were used, yielding a cleaned dataset with few missing values and inconsistencies. After data preparation, over 99% of the data was maintained, assuring data quality and dependability.

***Feature Engineering:***

Extensive feature engineering was carried out, resulting in the extraction of 50 meaningful feature. These characteristics were carefully chosen for their clinical significance and possible predictive value.

***Model Development:***

Several machine learning models were trained and assessed on the dataset, including logistic regression, decision trees, and K-Nearest Neighbors. The models predicted the target disease with an average accuracy of 100%, showing high overall performance.

***Model Refinement:***

The chosen model is presently being refined to improve its performance. To increase the model's accuracy and generalization capabilities, hyperparameter tweaking and feature selection strategies are being used. Preliminary findings show that the new models outperform the old ones.

***Interpretability:***

Model interpretability approaches such as feature importance analysis and partial dependency plots were utilized to give meaningful explanations for the model's predictions, assuring the model's interpretability and explain ability to clinicians and stakeholders.

1. **Changes**

Everything worked perfectly till the Model creation phase, we thought it will be easy but actually, when we tried to create the model we found that we have to rearrange our data in a way that can easy to fit the model. It took around a week to rearrange all things.

We planned a week to complete our UI, but for us, it was a new technology and new tool to create UI and for that, we had to make changes in our design *timeline.*

1. **Challenges**

***Modelling methodologies were changed:***

Initially, we intended to utilize solely logistic regression models for disease prediction. However, we discovered that alternative machine learning models, such as decision trees and K-Nearest Neighbors, performed better during model building. As a result, we implemented these concepts into our strategy.

1. **Result**

We have made tremendous progress in the disease prediction project and have met the majority of our goals. We successfully collected a good dataset, cleaned the data, did feature engineering, and created a machine learning model for disease prediction. The model performed well overall, with an accuracy of 100%.

Although we have met our major aims, our work has certain limits and flaws. First, because our dataset is small, the model's generalizability to different populations may be limited. Furthermore, there may be missing data or inconsistencies in the data that were not found during data cleaning, which might influence the model's accuracy.

To overcome these constraints, we intend to acquire a larger and more varied dataset in order to increase the model's generalizability. We will also investigate various modelling approaches and algorithms in order to increase the model's performance and resilience. We will also continue to evaluate and improve the model to ensure its real-world applicability and efficacy.

1. **Timeline**

We're now in the middle of the disease prediction project, and we've made great strides so far. Based on our present efforts, we anticipate finishing the project within the next three months. Here is a revised project timeline:

* 1. **Below are tasks we did as of now.**
     1. Problem statement
     2. Data visualization
     3. Data cleaning
     4. Data transformation
     5. Models finalize.
  2. **We are working on the below tasks.**
     1. Model modifications for input
     2. UI for the remaining screen.
  3. **Budget**
     1. Domain which will cost around $250.
     2. Good system which will cost around $3000.
     3. Hosting which will cost around $500.
     4. UI/UX cost $4000.
     5. Website creation cost $5000.

1. **Future work**

Based on our preliminary assessment of the illness prediction project, we have recommended a number of next measures to increase the initiative's efficacy, dependability, and real-world impact. Here is our strategy for the project's next phase:

***Collect a larger and more varied dataset:***

We intend to collect a larger and more diverse dataset from different sources, including more hospitals and healthcare providers, to increase the model's generalizability and performance.

***Investigate more modelling approaches and algorithms:***

While our present machine learning model worked well, we would like to investigate other modelling approaches and algorithms in order to increase the model's accuracy and durability.

***Validate and develop the model further:***

We will continue to validate and enhance the model to verify its real-world applicability and efficacy, including testing it on fresh data and comparing its performance to other illness prediction models.

***Create a user-friendly interface:***

To make the model more accessible to doctors and healthcare providers, we want to create a user-friendly interface that clearly presents the model's predictions and suggestions.

***Conduct pilot testing and get feedback:***

Before implementing the model in the real world, we will conduct pilot testing and gather feedback from clinicians and healthcare providers to confirm that the model is fulfilling their needs and giving relevant insights.

1. **Conclusion**

To summarize, we have achieved substantial progress so far in the illness prediction effort. We collected and cleaned a large dataset, built a machine learning model with high overall performance, and used model interpretability approaches to assure the model's explain ability to clinicians and stakeholders. However, we recognize that our work has significant limits and obstacles, such as the limited sample size and probable missing data.

Moving forward, we want to solve these constraints by amassing a larger and more diversified dataset and experimenting with different modelling methodologies and algorithms. In addition, we want to continue validating and refining the model to ensure its real-world applicability and efficacy.

Based on our present work, we expect to finish the project in six months for a total cost of roughly $45,000. At this time, we do not expect any substantial modifications to the project's timeframe or budget.

Overall, we feel that our illness prediction initiative has the potential to have a substantial impact in the healthcare profession by assisting physicians in more successfully predicting and preventing diseases. We will continue to work hard to guarantee the success of the project and will keep stakeholders updated on our progress.