DISEASE PREDICTION

USING MACHINE LEARNING

PROJECT SYNOPSIS

OF MAJOR PROJECT

Submitted By

Kunal Raj

University Roll no - 10300220051

Gourab Ghosh

University Roll no -10300220041

Aman Kumar

University Roll no -10300220013

Chandra Kant

University Roll no -10300220037

Under the Supervision
Of

Dr. Ramkrishna Ghosh

Associate Professor

Department of Information Technology

Haldia Institute of Technology

ICARE Complex, Hatiberia, Haldia, Dist - Purba Medinipur

West Bengal, India, PIN - 72165

CONTENTS

INTRODUCTION

Machine Learning model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services. [10]

To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data.

For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm. Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience.

The machine learning algorithm has two phases: 1) Training & 2) Testing.

To predict the disease from a patient's symptoms and from the history of the patient, machine learning technology is struggling from past decades. [9]

Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying complete machine learning concepts to keep the track of patient's health. ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. [6]

By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services.

To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm. [5]

Due to an increased amount of data growth in the medical and healthcare field the accurate analysis on medical data which has been benefited from early patient care.

With the help of disease data, data mining finds hidden pattern information in the huge medical data of the data set. [1]

We proposed a disease prediction platform, based on the vitals of the patient. Our disease prediction ML model predicts the occurrence of kidney diseases.

Here are some examples of how useful ML can be in field of medicine -

- **Spotting diseases:** Machine learning can look at X-rays and scans to help find problems early.
- **Medicine magic:** It can help design new drugs and figure out which ones might work best or you.
- **Personalized care:** Like a super-smart assistant, it can suggest the best treatment plan based on your health.
- **Predicting health bumps:** Machine learning can warn doctors if you might get sick, so they can help prevent it.

Machine learning is not meant to replace doctors; rather, it serves as a powerful assistant, augmenting their expertise with data-driven insights and capabilities. [2]

This collaboration between human intelligence and machine learning holds immense promise for the future of healthcare.

By leveraging machine learning's ability to analyze vast amounts of data and identify complex patterns, doctors can gain a deeper understanding of their patients' conditions and make more informed treatment decisions. [2]

Machine learning can also automate routine tasks, freeing up doctors' time to focus on what they do best: building relationships with patients and providing compassionate care. [6]

Ultimately, this synergistic approach between human and machine intelligence has the potential to revolutionize healthcare, making it faster, more precise, and ultimately, more effective. [1]

Along with it, we have also provided an about page which gives information about the symptoms & information about the diseases.

OBJECTIVES

The objectives of our project are as follows: -

1. **Early detection & treatment: -** Engineers and medical researchers are trying to develop machine learning algorithms and models that can identify chronic kidney disease at an early stage. The problem is that the data generated in the health industry is large and complex, making data analysis difficult. [7]

2. **Reducing the fatality of disease: -** Combination of estimated glomerular filtration rate (GFR), age, diet, existing medical conditions, and albuminuria can be used to assess the severity of kidney disease, but requires more accurate information about the risk to the kidney is required to make clinical decisions. [3]

3. **Chronic kidney disease (CKD)** is a life-threatening condition that can be difficult to diagnose early because there are no symptoms. The purpose of the proposed study is to develop and validate a predictive model for the prediction of chronic kidney disease.

   The main function of the kidney is to filter the blood in the body. Kidney disease is a silent killer because it can cause kidney failure without causing any symptoms or concern. [4]

4. **Chronic kidney disease** is defined as a decline in kidney function over a period of months or years. Kidney disease is often caused by diabetes and high blood pressure, and by using models such as these we would be able to predict, detect, prevent as well as properly cure such diseases.

LITERATURE REVIEW

Several studies have been performed on disease prediction using machine learning. The prediction of disease using different ml algorithm help us to find accuracy among ML models.

Bemando et al. investigated the relationship between blood-related diseases and their features utilizing classifier methods such as Gaussian NB, Bernoulli NB, and Random Forest.

These three algorithms anticipate and offer statistical findings in a variety of ways. In this experiment, we discovered that Nave Bayes estimated accuracy was higher than that of other algorithms. [6]

Kumar and Polepaka devised a technique for illness prediction in the medical field. They employed Random Forest and CNN as well as other machine learning methods.

For illness dataset classification, precision, recall, and F1-score, these algorithms deliver better results. In this experiment, Random Forest outperformed other algorithms in terms of accuracy and statistical performance.

On the kidney disease dataset, Nithya et al. developed a method for categorization and cluster-based analysis.

On diverse sets of photos, the authors utilized the K-Means clustering technique to collect the closest familiar images. They calculated 99.61 percent classification accuracy using Artificial Neural Networks for Kidney Disease Image Prediction. [3]

The purpose of the proposed model is to predict whether the patient will suffer or develop chronic kidney disease in the future if he continues their lifestyle.

This information can be used to determine whether the kidney disease is using eGFR (glomerular filtration rate), which helps the doctor plan the appropriate treatment.

Estimated glomerular filtration rate (eGFR) defines the degree of kidney disease and measures kidney function.

A combination of estimated glomerular filtration rate (GFR), age, diet, existing medical conditions, and albuminuria can be used to assess the severity of kidney disease, but requires more accurate information about the risk to the kidney is required to make clinical decisions about diagnosis, treatment, and prognosis. [8]

While the factors mentioned provide a valuable starting point, a more comprehensive evaluation might involve additional tests like kidney ultrasounds, biopsies, or genetic screenings. [10]

These can pinpoint the underlying cause of the kidney damage, assess the extent of scarring, and identify potential complications. [4]

This more detailed picture allows healthcare professionals to tailor treatment plans to address the specific issues at hand, potentially slowing disease progression and improving long-term outcomes for the patient.

METHODOLOGY/ PLANNING OF WORK

1. Importing Dependencies: -

In machine learning, achieving success hinges on the efficient utilization of various tools and algorithms. These tools, often packaged as libraries or frameworks, serve as the building blocks for data manipulation, model creation, and evaluation. Effectively managing these dependencies is crucial for project reproducibility, maintainability, and collaboration.

Importance of Dependency Management:

1. **Consistency:** Specifying dependencies ensures everyone involved in the project uses the same library versions. This eliminates compatibility issues and unexpected behavior during code execution.
2. **Reproducibility:** A well-defined dependency list allows for replicating the project's results on different machines. This is vital for scientific validation and sharing research findings.
3. **Isolation:** Virtual environments, which isolate project dependencies from the system-wide Python environment, prevent conflicts arising from different project requirements.

Dependency management fosters a streamlined development workflow, facilitates collaboration, and enables the creation of reliable and reproducible machine learning solutions.

2. Analyzing Features*:* -

Importance of Feature Analysis:

- **Better Predictions:** By selecting features that have a strong correlation with the target variable (what we are trying to predict), the model can learn more accurate patterns. Imagine training a model to predict house prices. Features like

square footage and location are likely more important than the colour of the doorknobs.

- **Reduced Complexity:** A high number of features can make a model cumbersome and prone to overfitting, where it memorizes the training data but performs poorly on new data. Feature analysis helps identify and remove irrelevant or redundant features, streamlining the model and improving generalization.
- **Interpretability:** With fewer, well-chosen features, it's easier to understand the inner workings of the model. This is crucial for tasks like fraud detection where you need to know why a certain transaction was flagged as suspicious.

3. Data Cleaning and Preprocessing*:* -

Data preprocessing is the essential first step in machine learning, cleaning raw data for use in models. It improves data quality and model understanding. Common techniques include handling missing values, encoding categorical data, scaling features, and splitting data into training and testing sets. The specific techniques depend on the data and machine learning task.

Steps In Data Preprocessing:

1. Gathering the data*:*

Data is raw information; it is the representation of both human and machine observation of the world. Dataset entirely depends on what type of problem we want to solve. Each problem in machine learning has its own unique approach.
Kaggle: Kaggle is one of the best ones to get the dataset.

2. Import the dataset & Libraries*:*

First step is usually importing the libraries that will be needed in the program. A library is essentially a collection of modules that can be called and used.

3. Dealing with Missing Value*s:*

Sometimes we may find some data are missing in the dataset. if we found then we will remove those rows or we can calculate either **mean, mode or median** of the feature and replace it with missing values. This is an approximation which can add variance to the dataset.

4. Check for null values*:*

We can check the null values in our dataset with pandas library With the help of *info()* we can found total number of entries as well as count of non-null values with datatype of all features.

We work on large dataset so it will be a good thing to get the count of all null values corresponding to each features and it will be done by using *sum()* we work on large dataset so it will be a good thing to get the count of all null values corresponding to each features and it will be done by using *sum().*

5. Replacing Null values with Strategy*:*

For replacing null values, we use the strategy that can be applied on a feature which has numeric data. We can calculate the *Mean, Median or Mode* of the feature and replace it with the missing values.

6. Divide the dataset into Dependent & Independent variable*:*

After importing the dataset, the next step would be to identify the independent variable (X) and the dependent variable (Y).

4. Feature Selection*:* -

Feature selection acts like a discerning chef carefully choosing ingredients. It is the process of selecting the most relevant features from your data to train a model. Just like

a well-chosen recipe yields a tastier dish, using the right features can significantly improve our machine learning model.

Feature selection offers several key benefits:

- **Enhanced Model Performance:** By eliminating irrelevant or redundant features, you focus the model on the data that truly matters. This can lead to better accuracy, reduced overfitting, and potentially faster training times.
- **Improved Interpretability:** With fewer features to consider, it becomes easier to understand the inner workings of your model. This allows you to diagnose issues more easily and gain valuable insights from the relationships between the chosen features and the target variable.
- **Reduced Computational Cost:** Training models with fewer features requires less computational power. This is particularly advantageous for datasets with a high number of dimensions or when working with limited resources.

5. Fitting into Model: -

Model fitting is the heart of machine learning. It is the process of training a model on data to learn the underlying patterns and relationships. The goal is to create a model that can accurately predict outcomes for new, unseen data.

Importance:

- **Generalizability:** A well-fit model can generalize its learnings to similar data beyond the training set. This allows for reliable predictions on real-world applications.
- **Accuracy:** The fitting process optimizes the model's parameters to minimize errors between predictions and actual values.

CONCLUSION

This training has introduced us to Machine Learning. Now, we know that **Machine Learning is a technique of training machines to perform the activities a human brain can do, a bit faster and better than an average human-being**. Today we have seen that the machines can beat human champions in games such as Chess, Mahjong, which are considered very complex.

We have seen that machines can be trained to perform human activities in several areas and can aid humans in living better lives. Machine learning is quickly growing field in computer science. It has applications in nearly every other field of study and is already being implemented commercially because **machine learning can solve problems too difficult or time consuming for humans to solve.**

To describe machine learning in general terms, a variety models are used to learn patterns in data and make accurate predictions based on the patterns it observes. Machine Learning can be a Supervised or Unsupervised. If we have a lesser amount of data and clearly labelled data for training, we opt for Supervised Learning.

Unsupervised Learning would generally give better performance and results for large data sets. If we have a huge data set easily available, we go for deep learning techniques. We also have learned Reinforcement Learning and Deep Reinforcement Learning. We now know what Neural Networks are, their applications and limitations.

Specifically, we have developed a thought process for approaching problems that machine learning works so well at solving. We have learnt how machine learning is different than descriptive statistics.

REFERENCES

1. C. Bemando, E. Miranda, M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," in 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), (IEEE, 2021), pp. 232–237

2. . R.P. Ram Kumar, Sanjeeva Polepaka, Performance comparison of random forest classifier and convolution neural network in predicting heart diseases, in Proceedings of the Third International Conference on Computational Intelligence and Informatics. ed. by K. Srujan Raju, A. Govardhan, B. Padmaja Rani, R. Sridevi, M. Ramakrishna Murty (Springer, Singapore, 2020)

3. A. Nithya, A. Appathurai, N. Venkatadri, D.R. Ramji, C.A. Palagan, Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images. Measurement (2020). https://doi.org/10.1016/j.measurement.2019.106952

4. F. Aqlan, R. Markle, A. Shamsan, "Data mining for chronic kidney disease prediction." in IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers, (IISE 2017), pp. 1789–1794

5. A. Nishanth, T. Thiruvaran, Identifying important attributes for early detection of chronic kidney disease. IEEE Rev. Biomed. Eng. 11, 208–216 (2018)

6. R.S. Walse, G.D. Kurundkar, S.D. Khamitkar, A.A. Muley, P.U. Bhalchandra, S.N. Lokhande, Effective use of naïve bayes, decision tree, and random forest techniques for analysis of chronic kidney disease, in International Conference on Information and Communication Technology for Intelligent Systems. ed. by T. Senjyu, P.N. Mahalle, T. Perumal, A. Joshi (Springer, Singapore, 20).

7. Application of Machine Learning in Chronic Kidney Disease: Current Status and Future Prospects [MDPI] (application of machine learning in chronic kidney disease current status and future prospects ON MDPI mdpi.com)

8. Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease [Hindawi] (machine learning hybrid model for the prediction of chronic kidney disease ON Hindawi hindawi.com)

9. Chronic Kidney Disease Prediction Using Machine Learning Methods [ResearchGate] (chronic kidney disease prediction using machine learning methods ON ResearchGate researchgate.net)

10. Chronic kidney disease prediction using machine learning techniques [Journal of Big Data] (chronic kidney disease prediction using machine learning techniques ON Springer journalofbigdata.springeropen.com)

11. Investigation on explainable machine learning models to predict chronic kidney diseases [Nature] (investigation on explainable machine learning models to predict chronic kidney diseases ON Nature Journal nature.com)

12. Machine learning to predict end-stage kidney disease in chronic kidney disease [Nature Research] (machine learning to predict end stage kidney disease in chronic kidney disease ON Nature Journal nature.com)

13. Novel Machine Learning Approach for Predicting Chronic Kidney Diseases (Shodhganga) (thesis chronic kidney disease prediction ON Shodhganga shodhganga.inflibnet.ac.in)

14. Role of machine learning in early diagnosis of kidney diseases (University of Louisville) (machine learning in early diagnosis of kidney diseases ON University of Louisville ir.library.louisville.edu)