



Hereditary Disease Prediction using Machine Learning

Dr.D.Suneetha¹ | Raavi Lakshmi² | Naviri Karuna² | Nallamudi Praveen Kumar² | Juluru Vinay Mouli²

¹Professor & HOD, Department of CSE, NRI Institute of Technology, Vijayawada, A.P., India.

²Department of CSE, NRI Institute of Technology, Vijayawada, A.P., India.

To Cite this Article

Dr.D.Suneetha, Raavi Lakshmi, Naviri Karuna, Nallamudi Praveen Kumar and Juluru Vinay Mouli. Hereditary Disease Prediction using Machine Learning. International Journal for Modern Trends in Science and Technology 2022, 8(03), pp. 105-108. <https://doi.org/10.46501/IJMTST0803020>

Article Info

Received: 08 February 2022; Accepted: 10 March 2022; Published: 16 March 2022.

ABSTRACT

Genetic diseases mean that defective genes are inherited from their parents. In this case, the genetic disorder is known as a hereditary disease. Heredity means people inherit one copy from their parents, which they inherit from their future generations. The hereditary disease can occur abruptly when healthful providers of a faulty recessive gene reproduce. However, this can also occur when the faulty gene is dominant. Most of the genetic disorders are quite rare and affect one person in every several thousand or million. So this process will predict the hereditary disease by using different machine learning techniques. Machine learning plays an efficient role in medical systems. Earlier identification of diseases is very crucial nowadays. The main objective of this system is to analyze the data and predict whether a person will inherit the disease from his or her family's roots. To predict the result, supervised learning algorithms like a decision tree and random forest are used to predict the result. The prediction of the result is based on accuracy, precision, recall, f1_measure which is implemented using the python programming language. This will show which algorithm gives more accuracy in predicting the result. Based on the result, the candidate can take some precautions so that people can lead healthy lives. This prediction is carried out on diseases like diabetes, cancer, heart attack, nerve weakness, and hair issues.

KEYWORDS: Hereditary, machine learning, decision tree, random forest, prediction

1. INTRODUCTION

Heredity is the sum of all biological processes by which particular characteristics are transmitted from parents to their children. The concept of heredity is explained in two ways, species from generation to generation and the variation among individuals within a species. Every member of a species has a fixed number of genes unique to that species. It is this set of genes that gives the fidelity of the species. Among the individuals within a species, however, variations can occur in the form each gene takes, providing the genetic basis for the truth that no individuals (besides the same twins) are precisely identical. The set of genes that children inherit from both parents, a combination of the genetic material

of each is called the organism's genotype. The genotype is contrasted to the phenotype, that is, the organism's outward appearance and the final development results of its genes. The phenotype consists of an organism's physical structures, physiological processes, and behaviors. Although the genotype determines the extensive limits of the capabilities an organism can develop, the capabilities that without a doubt develop, i.e., the phenotype, depend upon complicated interactions between genes and their environment. The genotype remains constant throughout an organism's lifetime. However, due to the fact the organism's inner and outside environments change continuously, so does its phenotype.

In carrying out genetic studies, it's far more critical to find out the degree to which the observable trait is due to the sample of genes inside the cells. Machine learning plays a major role in identifying hereditary diseases from genes and it provides efficient results. The pressure on man's health is increasing based on various factors like the environment, lifestyle, etc. Nowadays, hereditary diseases are very common, and predicting them before and changing our lifestyle will give better results. This ensures that our future generations will not suffer that particular disease. We consider the medical history of our past generations and predict diseases like heart attack, diabetes, nerve weakness, hemophilia, hair issues, and cancer. This system predicts whether the disease will pass on or not. Based on some conditions, some people might not get affected, but some people might get affected. The main objective is to efficiently predict hereditary diseases from the dataset. Use numerous ml algorithms to construct prediction models, examine the accuracy and overall performance of those models. To increase the accuracy of the classification results.

2. LITERATURE REVIEW

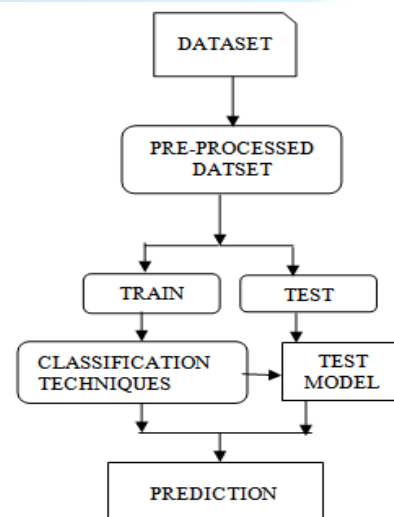
Hereditary diseases are exceeded from one generation to some other through defective genes. These hereditary diseases are transmitted within the same family. The chromosomes in humans are responsible for passing the traits from the parent to their children. In this type of approach, we can't do anything just to accept the truth. In the existing system, they have done the process based on various datasets, like a hereditary heart attack or hereditary diabetic dataset, and so on. In that, they have analyzed individually and predicted the dataset by using some machine learning techniques. In that analysis, there are some disadvantages, like low accuracy, and the performance is considerably very low. It can predict only the individual dataset like heredity heart attack, heredity cancer, or hereditary diabetics. Despite the significance of machine learning methods of strategic utility in conventional medicinal drugs, there may be no systematic literature review and classification for this field. This is the primary complete literature evaluation of the application of data mining strategies in conventional medicine. We reviewed five databases from 2000 to 2017 primarily based totally on the kitchenham systematic evaluation methodology.502

articles were identified and reviewed for their relevance to the application of machine learning methods in traditional medicine. There are 42 selected papers that are classified and categorized into four dimensions[1]. They are the application domain of data mining techniques in traditional medicine, the data mining methods most frequently used in traditional medicine, the main strength and limitation of data mining techniques in traditional medicine, and the performance evaluation methods in data mining methods in traditional medicine.

3. PROPOSE SYSTEM

The proposed model is introduced to overcome the disadvantages that are present in the existing system. This system will analyze the genetic diseases which occur from their family members to the next generations by using machine learning algorithms like a decision tree and random forest. Before, the prediction of diseases was meant for only a particular disease, but we have developed our project that this system can predict the chances of inheriting the diseases from their family members to their next generations. With the assistance of this, we will save you heredity sicknesses earlier than they're inherited with the aid of using the subsequent generations. It enhances the performance of the overall classification report. Predicting the heredity from human gene data and finding the accuracy is more reliable as it takes less time to predict and handle the dataset properly and generates decent results. Based on the accuracy generated, it will state which algorithm is more efficient.

4. SYSTEM ARCHITECTURE



5. IMPLEMENTATION

The modules included in this project are data selection and loading, data preprocessing, feature selection, classification, prediction.

A. Data selection and Loading

Data selection is the process of selecting the data predicting heredity. In this project, the dataset is loaded using the pandas read_csv() function.

B. Data preprocessing

Data pre-processing is the process of removing unwanted data from the dataset. Two formats are missing data removal, encoding categorical data. Missing data removal is the process in which null values consist of missing values that have been eliminated with the usage of the imputer library. Encoding categorical data means maximum machine learning algorithms require numerical input and output variables. For that reason, one-hot encoding is used to transform categorical data into integer data.

C. Splitting dataset into train and test data

Data splitting is the act of partitioning to be had statistics into two portions, normally for cross-validatory purposes. One part of the data is used to increase a predictive model. And the opposite to assess the model's performance. Separating data into training and testing sets is an essential part of evaluating data mining models. Typically, while you separate a data set right into a training set and testing set, a maximum of the data is used for training, and a smaller part of the data is used for testing.

D. Classification

Decision Tree

A decision tree is a supervised learning approach that may be used for each classification and regression problem, however frequently it's favored for fixing classification problems. It is a tree-based classifier, in which inner nodes constitute the features of a dataset, branches constitute the decision rules and every leaf node represents the outcome. In a decision tree, there are nodes which can be the decision node and leaf node. The decision or test is carried out on the idea of the capabilities of the given dataset. It is referred to as a decision tree because, just like a tree it begins with the root node, which expands on similar branches and constructs a tree-like structure. A decision tree in reality asks a question, and primarily based totally on the

answer yes or no, it similarly split up the tree into subtrees.

Random Forest

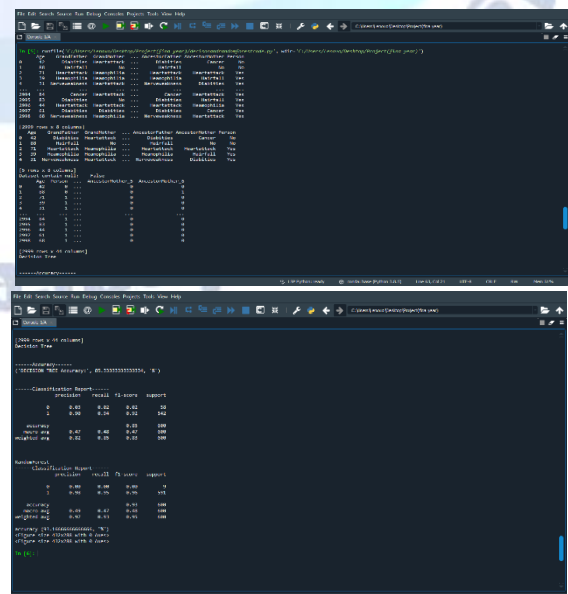
Random forest is a famous machine learning algorithm that belongs to the supervised learning technique. It may be used for classification and regression issues. It is primarily based totally on the idea of ensemble learning, that's a procedure of mixing more than one classifier to remedy complicated trouble and to enhance the overall performance of the model. As the call suggests, random forest is a classifier that includes several decision trees on various subsets of the given dataset and takes the common to enhance the predictive accuracy of that dataset. Instead of counting on one decision tree, the random forest takes the prediction from every tree and is primarily based totally on the bulk votes of predictions, and it predicts the very last output. The more wide variety of trees within the random forest ends in better accuracy and prevents the trouble of overfitting.

E. Prediction

Process of predicting hereditary genetic diseases from the dataset. This project will effectively predict whether a particular person inherits disease or not from the dataset by enhancing the performance of the overall prediction results. The performance of this proposed approach is evaluated using some measures like accuracy, precision, recall, F1-measure.

I. Sample screens

Sample screens are the output screens that show accuracy, classification report, the dataset, and the encoded data.



The image shows two screenshots of a Jupyter Notebook. The top screenshot displays a dataset with columns for 'Age', 'Sex', 'Height', 'Weight', 'Blood Pressure', 'Cholesterol', 'Glucose', and 'Diabetes'. The bottom screenshot shows the output of a classification model, including a confusion matrix and a classification report. The classification report shows a score of 0.99, indicating high accuracy.

6. CONCLUSION

In this study, the machine learning classifiers are used to predict the hereditary. The hereditary data is taken as input data and applied to the pre-processing method. In the pre-processing method, the process will be like cleaning the dataset and applying the label encoding. Then it is processed into a feature selection method, in this method the dataset is split into a training dataset and testing dataset. Finally, the classification method machine learning algorithm is used to predict the hereditary in human gene data and find the result based on accuracy.

7. FUTURE SCOPE FOR FURTHER DEVELOPMENT

In the future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms to achieve further increased performance. It is also possible to use other machine learning algorithms and can implement a real-world application for the user experience in a simple manner so that they can take precautions for hereditary diseases. The future enhancement for this project will be adding many more hereditary diseases. In the future, refine the stacking ensemble framework and utilize our approach to other real-world applications.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

- [1] A systematic literature review and classification of knowledge discovery in traditional medicine GoliArji, Reza Safdari, Hossein Rezaeizadeh, AlirezaAbbassian, MehrshadMokhtaran, Mohammad Hossein Ayati.
- [2] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, "Decision making in advanced heart failure: A scientific statement from the American heart association," *Circulation*, vol. 125, no. 15, pp. 1928–1952, 2012.
- [3] Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. RezaAlbarrán, and K. L. Ramaiya, "Diabetes in developing countries," *Journal of Diabetes*, vol. 11, no. 7, pp. 522-539, Mar. 2019.
- [4] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm-based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in *Proc. International Conference on Computing Networking and Informatics*, Oct. 2017, pp. 1-5.

- [5] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," *Nature Rev. Cardiol.*, vol. 8, no. 1, p. 30, 2011.